Actes de JEP-TALN-RECITAL 2024. 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position

Mathieu Balaguer, Nihed Bendahman, Lydia-Mai Ho-Dac, Julie Mauclair, José G. Moreno, Julien Pinquier

35èmes Journées d'Études sur la Parole (JEP 2024)
31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024)
26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2024)

https://jep-taln2024.sciencesconf.org

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position

Mathieu BALAGUER, Nihed BENDAHMAN, Lydia-Mai HO-DAC, Julie MAUCLAIR, Jose G MORENO, Julien PINQUIER (Éds.)

Toulouse, France, 8 au 12 juillet 2024

Avec le soutien de

Actes JEP-TALN-RECITAL Toulouse 2024 – ISBN : 978-2-917490-37-2
Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles – ISBN : 978-2-917490-39-6

Préface

Organisée conjointement par les équipes de recherche IRIS, MELODI et SAMoVA de l'Institut de Recherche en Informatique de Toulouse (IRIT UMR 5505), l'équipe PLC du laboratoire Cognition, Langues, Langage, Ergonomie (CLLE UMR 5263) et l'axe neurocognition langagière, linguistique et phonétique cliniques du laboratoire de NeuroPsychoLinguistique (LNPL URI EA 4156), sous l'égide de l'Association Francophone de la Communication Parlée (AFCP) et l'Association pour le Traitement Automatique des Langues (ATALA), la conférence JEP-TALN-RECITAL 2024 regroupe :

— les 35ème Journées d'Études sur la Parole (JEP),
— la 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN),
— la 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL).

Les conférences TALN et JEP sont un rendez-vous qui offre le plus important forum d'échange francophone aux acteurs universitaires et industriels des technologies de la langue et la parole. Pour cette édition, nous avons plus de 200 inscrits dont une grande partie des étudiants qui construisent le futur de la recherche francophone et assurent le relais de son développement.

En tant que conférenciers invités, nous aurons Véronique HOSTE de l'Université de Ghent, Laurent BESACIER de Naver Labs Europe et Catia CUCCHIARINI de l'Université de Radboud. Ces trois conférenciers qui représentent un large spectre de thématiques entre le texte et la parole vont aborder les dernières avancées de leurs domaines d'expertise.

Cette édition permet aussi de célébrer les 30 ans de TALN. À cette occasion, nous avons dédié une session spéciale dans le programme. La session a comme objectif de rappeler l'historique de la conférence avec l'intervention des participants qui ont participé à sa pérennité afin de mieux transmettre les enjeux de ce rassemblement à la communauté scientifique du traitement automatique des langues naturelles.

En termes des soumissions, pour TALN, 66 articles pour la conférence principale ont été soumis, dont respectivement 18 ont été acceptés pour une présentation orale et 30 pour une présentation sous forme de posters. Également, nous avons reçu 13 résumés des articles publiés lors de conférences internationales qui ont été acceptés pour une présentation en format poster. En ce qui concerne RECITAL, 11 articles ont été soumis dont 7 ont été acceptés. L'ensemble des soumissions acceptées seront présentées sous forme de posters et 3 d'entre elles donneront lieu à une présentation orale. Pour les JEP, 64 articles ont été soumis et 62 ont été acceptés (17 sous forme de présentation orale et 45 sous format poster). L'alternance de sessions communes entre TALN, JEP et RECITAL et de sessions plus spécifiques devraient permettre de susciter des échanges fructueux. En complément de la conférence principale, se tiennent les ateliers "Parole Spontanée", "Défi Fouille de Texte" (DEFT), "Jurisprudence Prédictive" (JP'24), "Evaluation des modèles génératifs" (EvalLLM) et l'activité HackaTAL 2024. Ces événements illustrent à la fois des tendances nouvelles présentes dans la communauté et des activités récurrentes.

Il convient d'exprimer une profonde reconnaissance envers toutes les personnes qui ont participé à faire vivre la conférence, d'un côté les auteurs de toutes les soumissions et de l'autre les membres de différents comités scientifiques de la conférence. Un remerciement très chaleureux aux relecteurs qui ont accepté une charge importante et qui ont fait des relectures d'urgence afin de faciliter le bon déroulement de la conférence. La bienveillance et l'expertise des comités de programme ont permis la constitution d'un programme riche en thématiques et d'un niveau scientifique correspondant aux attentes de la communauté. Il est également essentiel d'exprimer notre gratitude envers les sponsors et les organisations qui ont subventionné la conférence. Leur soutien financier a permis à cet événement scientifique de se réaliser dans les meilleures conditions, rappelant l'importance des aspects financiers dans la réussite de telles

initiatives. Finalement, un grand merci aux différentes équipes présentes pour le bon fonctionnement, notamment des équipes de l'ATALA, l'AFCP et le CPRS qui nous ont accompagnés dans les différentes étapes de l'organisation.

Jose G Moreno	Lydia-Mai Ho-Dac Nihed Bendahman	Julie Mauclair
Président de TALN	Présidentes de RECITAL	Présidente de JEP

Comités

Comité de Programme

— Rachel Bawden, Inria
— Leonor Becerra-Bonache, Laboratoire d'Informatique et Systèmes
— Delphine Bernhard, LiLPa, Université de Strasbourg
— Nathalie Camelin, LIUM — Université du Maine
— Marie Candito, Universtité Paris 7 / INRIA
— Vincent Claveau, Irisa
— Géraldine Damnati, Orange Labs
— Iris Eshkol-Taravella, University of Orléans
— Benoit Favre, Aix-Marseille Université
— Natalia Grabar, STL CNRS Université Lille 3
— Thierry Hamon, France
— Lydia-Mai Ho-Dac, CLLE
— Philippe Langlais, Canada
— Jose G Moreno, IRIT – Université Paul Sabatier
— Emmanuel Morin, Université de Nantes, LS2N
— Vincent Segonne, Université Bretagne Sud, UMR CNRS 6074, IRISA, F-56000 Vannes, France
— Christophe Servan, Qwant Research
— Anne Vilnat, LIMSI-CNRS

Comité de Relecture

— Maxime Amblard, Université de Lorraine
— Jean-Yves Antoine, Université Franois Rabelais de Tours
— Lauriane Aufrant, Inria
— Frederic Bechet, Aix Marseille Université - LIF
— Patrice Bellot, Aix-Marseille Université - CNRS (LIS)
— Asma Ben Abacha, Microsoft Health AI
— Timothée Bernard, Université Paris Cité
— Romaric Besançon, CEA LIST
— Philippe Blache, LPL, AMU
— Chloé Braud, IRIT - CNRS
— Remi Cardon, CENTAL, IL&C, Université Catholique de Louvain
— Maximin Coavoux, CNRS, Université Grenoble Alpes
— Matthieu Constant, Université de Lorraine, ATILF, CNRS
— Caio Corro, Université Paris-Saclay
— Benoît Crabbé, Paris 7 et INRIA
— Béatrice Daille, Laboratoire d'Informatique Nantes Atlantique (LINA)
— Gaël de Chalendar, CEA LIST
— Gaël Dias, Normandie University
— Taoufiq Dkaki, IRIT, Intitut de Recherche en Informatique de Toulouse
— Benamara Farah, Univ. Paul Sabatier, Toulouse and IPAL, Singapore
— Olivier Ferret, CEA List
— Karën Fort, Sorbonne Université
— Amel Fraisse, Université de Lille
— Thomas Francois, Université catholique de Louvain
— Sahar Ghannay, LISN lab
— Cyril Grouin, LISN

— Gaël Guibon, Université de Lorraine - LORIA
— Nabil Hathout, CNRS
— Nicolas Hernandez, Nantes Université - LS2N CNRS UMR 6004
— Gilles Hubert, IRIT
— Luce Lefeuvre, DTIPG, SNCF
— Fabio Martínez Carrillo, Bivl2ab- Biomedical Imaging, vision and learning laboratory. Universidad Industrial de Santander
— Véronique Moriceau, IRIT Université Toulouse 3
— Philippe Muller, IRIT, Toulouse University
— Alexis Nasr, LIS
— Aurélie Névéol, Université Paris-Saclay, CNRS, LISN
— Jian-Yun Nie, University de Montreal
— Damien Nouvel, INaLCO
— Yannick Parmentier, LORIA - Université de Lorraine
— Patrick Paroubek, Universite Paris Sacle - CNRS
— Benjamin Piwowarski, CNRS / ISIR, Sorbonne Université
— Thierry Poibeau, LaTTiCe-CNRS
— Solen Quiniou, LS2N - Nantes Université
— Benoît Sagot, INRIA
— Djamé Seddah, Alpage/Université Paris la Sorbonne
— Nasredine Semmar, CEA
— Ludovic Tanguy, CLLE-ERSS
— Xavier Tannier, Sorbonne Université, INSERM, LIMICS
— Julien Tourille, CEA, LIST
— Guillaume Wisniewski, LLF - Université de Paris
— François Yvon, CNRS
— Pierre Zweigenbaum, Université Paris-Saclay, CNRS, LISN

Table des matières

Première partie

Articles présentés oralement

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

À propos des difficultés de traduire automatiquement de longs documents

Ziqian Peng[1, 2] Rachel Bawden[2] François Yvon[1]
(1) Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR, F-75005 Paris, France
(2) Inria, Paris, France
prénom.nom@isir.upmc.fr, prénom.nom.@inria.fr

RÉSUMÉ

Les nouvelles architectures de traduction automatique sont capables de traiter des segments longs et de surpasser la traduction de phrases isolées, laissant entrevoir la possibilité de traduire des documents complets. Pour y parvenir, il est nécessaire de surmonter un certain nombre de difficultés liées à la longueur des documents à traduire. Dans cette étude, nous discutons de la traduction des documents sous l'angle de l'évaluation, en essayant de répondre à une question simple : comment mesurer s'il existe une dégradation des performances de traduction avec la longueur des documents ? Nos analyses, qui évaluent des systèmes encodeur-décodeur et un grand modèle de langue à l'aune de plusieurs métriques sur une tâche de traduction de documents scientifiques, suggèrent que traduire d'un bloc des documents longs reste un problème difficile.

ABSTRACT

Document Level Machine Translation: does length matter?

Today's machine translation architectures can process long segments and go beyond the translation of isolated sentences, opening up the possibility of translating full documents. To achieve this goal, it is necessary to overcome several difficulties related to the length of source documents. In this work, we discuss document-level machine translation from an evaluation perspective, trying to answer a simple question: how can we measure whether translation performance degrades with document length? Our analysis, which compares encoder-decoder systems and a large language model using multiple metrics on a scientific document translation task, suggests that translating long documents holistically remains a challenging problem.

MOTS-CLÉS : Traduction Automatique, Évaluation de la traduction, Traitement de documents.

KEYWORDS: Machine Translation, Evaluation of Machine translation, Document-level processing.

1 Introduction

Les évolutions récentes des modèles *Transformer* (Vaswani *et al.*, 2017) permettent de traiter (c'est-à-dire d'encoder et de décoder) de très longs contextes contenant des centaines, voire des milliers d'unités : voir (Tay *et al.*, 2023) pour quelques-unes des méthodes qui rendent ce calcul possible. Cette capacité à encoder un long contexte qui va conditionner le processus de génération ouvre de nouvelles perspectives pour de nombreuses applications du traitement automatique des langues (TAL). Nous nous intéressons ici à la traduction automatique (TA), en nous interrogeant sur les difficultés de

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 2–21.

traduire des documents [1] de manière holistique.

La *traduction holistique* d'un document consiste à l'encoder intégralement, puis à générer d'un trait toute la traduction, à l'instar, par exemple, de (Zhang *et al.*, 2018; Junczys-Dowmunt, 2019; Liu *et al.*, 2020a). Cette stratégie revient à traiter les documents comme sont traitées les phrases dans les systèmes de traduction standard. Elle se distingue des méthodes qui exploitent un contexte étendu aux phrases précédentes pour traduire la phrase courante (Post & Junczys-Dowmunt, 2023), comme de celles qui traduisent d'un bloc des segments étendus (correspondant à des fragments de taille fixe, ou bien à des paragraphes) (Tiedemann & Scherrer, 2017; Bawden *et al.*, 2018; Lopes *et al.*, 2020; Ma *et al.*, 2021) [2]. Avec le déploiement de grands modèles de langue multilingues dotés de capacité de traduction et capables d'interpréter des contextes très longs, cette stratégie devient de plus en plus commune (Hendy *et al.*, 2023; Zhang *et al.*, 2023), ce qui implique d'en analyser les principes et d'en diagnostiquer le fonctionnement.

Comme discuté §2, traduire holistiquement induit un changement notable par rapport à la traduction de phrases, aussi bien du point de vue des calculs réalisés que du point de vue des métriques, qui ne peuvent plus s'appuyer sur une comparaison phrase-à-phrase des sorties des systèmes et des références.

Dans la suite de cette contribution, nous nous intéressons plus particulièrement à une question essentielle pour le succès de ces approches, à savoir la capacité à traiter des documents de longueur variable. Plus précisément, notre contribution principale est de nature méthodologique et met à l'épreuve les méthodes expérimentales utilisées pour répondre à cette question. Nous pointons, dans un premier temps, les problèmes des comparaisons automatiques réalisées dans les articles de l'état de l'art, qui reposent sur des *métriques globales* calculées au niveau du corpus (le score BLEU (Papineni *et al.*, 2002) ou sa variante d-BLEU, voir §2.3). Nous présentons ensuite deux manières alternatives d'aborder les questions de longueur, en nous appuyant sur des *métriques locales*, qui individualisent les scores au niveau des documents. Nos expérimentations sont menées sur des documents de taille modeste (de quelques phrases à quelques dizaines de phrases), avec 7 systèmes de traduction automatique. Elles nous permettent néanmoins de conclure que la longueur des documents reste un problème, les scores de traduction ayant une tendance à se dégrader avec le nombre d'unités, un effet qui se manifeste clairement en fin de document.

2 La traduction automatique holistique de documents

2.1 Vers les systèmes traduisant des documents

Comme discuté, par exemple par Sun *et al.* (2022), l'approche holistique (aussi nommée « Doc2Doc ») se distingue de la plupart des travaux sur la traduction de documents qui le plus souvent consistent à augmenter le contexte (surtout côté source) des quelques phrases précédentes - avec potentiellement des encodeurs distincts pour la phrase courante et l'historique (Libovický *et al.*, 2018), tout en continuant de traduire phrase par phrase (« Sent2Sent »). Ces variantes sont appelées « Doc2Sent »

1. C'est-à-dire dont la longueur varie entre quelques phrases et quelques centaines de phrases, soit un résumé comme pour la campagne d'évaluation sur la TA biomédicale à WMT (Neves *et al.*, 2023), soit un exposé, comme pour la campagne d'évaluation de IWSLT 2023 (Salesky *et al.*, 2023), voire un article complet.

2. La terminologie « Traduction Automatique pour les Documents » (*Document-Level Machine Translation* ») ne distingue pas ces approches, dont certaines ne traitent pas des documents complets, mais plutôt des contextes élargis.

par Sun *et al.* (2022) et seraient plus correctement décrites comme des méthodes de traduction avec contexte étendu. Une étude de l'état de l'art, qui ne distingue pas ces deux méthodes, est dans (Maruf *et al.*, 2021). Les résultats expérimentaux qui implémentent ces techniques sont contrastés, ce qui a conduit certains à remettre en cause le bénéfice de contextes étendus (Kim *et al.*, 2019).

L'approche Doc2Doc, qui est notre principal sujet d'étude, est conceptuellement simple, mais elle introduit de multiples changements par rapport à la situation de référence où chaque phrase est encodée et décodée séparément des autres. Nous examinons ci-dessous les principaux changements pour des architectures encodeur-décodeur, sachant que les mêmes observations valent pour les approches à base de grands modèles de langue, lorsqu'on les emploie à des fins de traduction (Wang *et al.*, 2023; Karpinska & Iyyer, 2023). Traduire des documents de manière holistique signifie en particulier que :

— l'encodeur traite l'entièreté du document D source, composé de L phrases $D = (s_1 \dots s_L)$, comme une longue séquence, avec ou sans identification préalable des frontières de phrases ;
— pour générer la $l^{\text{ème}}$ phrase cible, le décodeur a accès à l'intégralité de D, ainsi qu'à toutes les phrases cibles déjà produites $t_{<l} = t_1 \dots t_{l-1}$.

Ces changements ont de nombreuses conséquences, certaines positives (4), d'autres négatives (1)-(3) :

1. les séquences à traiter sont plus longues, entraînant un surcoût computationnel car l'attention dans l'encodeur et le décodeur sont quadratiques en la longueur de l'entrée. Des implémentations approximatives efficaces de ce calcul permettent de conserver des temps de traitement raisonnables (Tay *et al.*, 2023) pour des longues séquences (jusqu'à quelques milliers d'unités).
2. lors du décodage des mots correspondant à la phrase source s_t, le décodeur ne peut plus s'appuyer sur un alignement explicite entre phrases et repose donc entièrement sur l'attention croisée, qui doit calculer une forme d'alignement de mots sur l'intégralité de D [3]. Cet effet est en particulier analysé par Bao *et al.* (2021).
3. décoder des séquences plus longues augmente les effets liés à l'accumulation des erreurs et au biais d'exposition (*exposure bias*) - dû au fait que l'apprentissage du modèle ne considère que des contextes cibles $t_{<l}$ corrects, alors qu'à l'inférence ils peuvent être erronés (Ranzato *et al.*, 2016; Mihaylova & Martins, 2019). Décoder des séquences plus longues réduit également la diversité des hypothèses représentées dans le faisceau de recherche.
4. l'allongement des contextes sources et cibles permet d'intégrer des dépendances plus longues, qui aident à désambiguïser des ambiguïtés lexicales ou des références pronominales.
5. les segments générés ne sont plus nécessairement en correspondance un-pour-un avec les segments sources, ce qui complique, voire obère, le calcul des métriques usuelles (voir §2.3).

Les impacts de ces changements ont été le plus souvent ignorés par les approches « mono-encodeur » qui traduisent simplement des segments longs comme s'il s'agissait de phrases isolées — en s'assurant toutefois que la longueur des segments d'apprentissage est cohérente avec celle qui sera vue au test [4].

2.2 Architectures pour la traduction de documents

Plus récemment, toutefois, divers travaux ont proposé des modifications significatives de l'architecture encodeur-décodeur de base portant notamment sur :

— la stratégie d'apprentissage, qui doit inclure des segments de longueur variable allant de la phrase isolée à des groupes de phrases (appelé « apprentissage multi-résolution » par Sun *et al.*

3. Alors que l'alignement des phrases est en général simple et monotone, contrairement à l'alignement de mots.
4. Et que la configuration du système est appropriée, c-à-d. qu'elle permet effectivement d'encoder des longs segments, que l'encodage des positions est correctement réalisé, etc.

(2022)). Le besoin de prendre en compte des documents d'apprentissage suffisamment longs est également pointé par (Zhuocheng *et al.*, 2023; Wu *et al.*, 2024) ;
— l'architecture du réseau, qui doit être plus profonde (augmentation de capacité) pour modéliser plus de phrases (Junczys-Dowmunt, 2019; Post & Junczys-Dowmunt, 2023), et également mieux régularisée (Kim *et al.*, 2019; Sun *et al.*, 2022) ;
— l'encodage des positions au sein du document source (Li *et al.*, 2022; Lupo *et al.*, 2023) ;
— la structure de l'auto-attention et de l'attention croisée avec des contraintes qui aident à localiser les phrases parallèles au sein d'un document (Bao *et al.*, 2021; Zhuocheng *et al.*, 2023; Herold & Ney, 2023) ;
— les méta-paramètres utilisés par le décodeur durant la génération (pénalité de longueur, largeur du faisceau, etc.) ;

L'enjeu principal de ces études est de parvenir à montrer que (a) les solutions techniques proposées parviennent effectivement à résoudre les problèmes de longueur (« *length bias* ») causés par la longueur des documents, et qu'une fois ces problèmes résolus (b) les méthodes holistiques surpassent les méthodes à traduisant phrase-à-phrase. Avant d'analyser de manière critique les arguments avancés pour prouver que la longueur n'est plus un problème (§3.1), nous nous intéressons dans un premier temps aux méthodes et métriques utilisées pour prouver (b).

2.3 La question de l'évaluation, nuances de BLEU

Répondre à cette question requiert des métriques permettant de comparer des traductions holistiques avec des traductions de phrases : comme le nombre de segments produits par les premières peut différer du nombre de segments sources, ces métriques doivent pouvoir comparer des documents ayant des longueurs différentes. La plupart des travaux en traduction de documents utilisent le score BLEU (Papineni *et al.*, 2002), ou plutôt une variante baptisée *d-BLEU* par Liu *et al.* (2020a) [5], et ceci en dépit des limites de cette métrique (Callison-Burch *et al.*, 2006; Reiter, 2018; Mathur *et al.*, 2020).

Le calcul de BLEU repose sur le décompte, phrase par phrase, du nombre de n-grammes (pour $1 \leq n \leq 4$) partagés par l'hypothèse de traduction et la référence humaine ; ces scores sont agrégés et normalisés, enfin moyennés (géométriquement) au niveau du corpus ; une pénalité de longueur, est enfin appliquée pour dégrader le score lorsque la longueur (agrégée) des hypothèses est plus courte que celle des références. BLEU est donc un score global, qui repose sur des alignements de phrases.

d-BLEU est également un score global, qui s'affranchit toutefois des appariements phrase-à-phase, et effectue les décomptes des n-grammes partagés au niveau des documents. Une conséquence est que d-BLEU, qui repose sur des correspondances élargies, tend à être plus élevé que BLEU, puisque les opportunités de trouver des n-grammes sont plus grandes dans une fenêtre plus large. Cet effet est connu et on peut le visualiser par exemple dans (Koehn & Knowles, 2017, Fig. 1), où l'on observe que BLEU augmente quand on considère des groupes de phrases de longueur croissante (au moins pour une certaine plage de longueur), là où on s'attendrait à une baisse (la longueur est souvent liée à la complexité syntaxique et donc à la difficulté de traduction). Il est facile de reproduire cette observation en calculant d-BLEU pour des systèmes qui traduisent phrase-à-phrase (voir le tableau 2).

Une alternative à d-BLEU consiste à réaligner traduction automatique et référence, par exemple avec

5. Hendy *et al.* (2023) considèrent également une variante de COMET (Rei *et al.*, 2022) ; Zhuocheng *et al.* (2023) introduisent, par analogie avec d-BLEU, la métrique d-chrF, qui est une variante de chrF (Popović, 2015).

l'algorithme de Wicks & Post (2022) [6], tel qu'il soit plus comparable avec les scores BLEU calculé au niveau de phrases reportés dans les publications. Ce problème se pose à l'identique en traduction de parole, l'algorithme de réalignement le plus utilisé dans ce contexte étant dû à Matusov *et al.* (2005). On se ramène ainsi au cas où référence et traduction ont même longueur et où BLEU peut être calculé. Avec cette approche, les scores obtenus dépendront des heuristiques utilisées pour l'alignement.

3 Évaluer les effets de longueur

Une question récurrente dans les études sur la traduction au niveau des documents est la question de la longueur. Un système entraîné uniquement avec des phrases parallèles aura tendance à sous-traduire des documents, stoppant typiquement le processus de génération après la première phrase. Comme expliqué §2.2, d'autres sources de problèmes pour les documents longs sont relatives aux encodages des positions (pour des positions non observées à l'apprentissage), ou encore à des méta-paramètres du système (longueurs maximales acceptées par l'encodeur et le décodeur, etc).

Ces problèmes affectent négativement la pénalité de longueur (BP) du score d-BLEU. Dans nos expériences (§5), le système FT SCIPAR, entraîné sur des phrases isolées, obtient un d-BLEU de 45,0 (BP=1) sur le corpus **THE** traduit phrase par phrase ; lorsqu'on utilise ce système pour traduire des documents complets, d-BLEU tombe à 0,9 (BP$= 0,02$). Ces problèmes sont assez faciles à corriger : dans nos expériences, en affinant le modèle avec des documents, d-BLEU remonte à 45,6 (BP$= 1$).

3.1 Les courbes BLEU / longueur

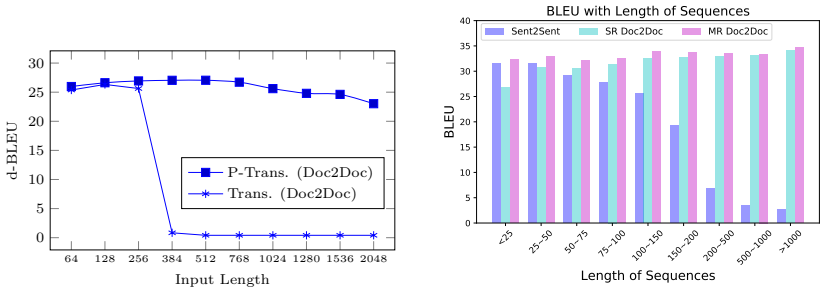


FIGURE 1 – Graphes BLEU / longueur extraits de (Li *et al.*, 2022, fig. 7) (à gauche) et de (Sun *et al.*, 2022, fig. 1) (à droite).

Pour mettre en évidence ces problèmes de longueur, ou pour prouver qu'ils ont été résolus, une première approche consiste à segmenter les documents en fragments de taille fixe, qui sont traduits séparément puis réassemblés pour calculer d-BLEU (Bao *et al.*, 2021, fig. 7), (Li *et al.*, 2022, fig. 7) et (Zhuocheng *et al.*, 2023, fig. 1 et 5). Le graphe 1 (gauche), extrait de Li *et al.* (2022) illustre la difficulté des modèles Sent2Sent à traduire des longs fragments et l'amélioration réalisée par des modèles Doc2Doc correctement entraînés. Cette comparaison occulte un biais qui joue en défaveur de la segmentation en fragments plus courts, qui (a) bénéficie de moins de contexte ; (b) traduit des fragments moins cohérents (démarrant ou s'achevant en milieu de phrases).

6. L'approche de Junczys-Dowmunt (2019) inclut un ensemble de tags qui contraignent les entrées et sortie à avoir le même nombre de phrases, voir aussi (Li *et al.*, 2022).

Une alternative consiste à regrouper les documents en fonction de leur taille et à calculer le score BLEU pour chacun des groupes de longueur – comme sur la figure 1 (droite), tirée de (Sun *et al.*, 2022). Cette manière de procéder évite l'écueil précédent, mais induit une légère confusion, puisqu'elle conduit à aligner sur un même graphique des scores incomparables, puisqu'ils sont calculés sur des corpus de test différents [7]. En se basant de telles analyses, il semble difficile de répondre avec certitude aux questions posées §2.2, ce qui motive les propositions développées ci-dessous.

3.2 Nuances de BLEU : nos analyses

Un premier changement méthodologique consiste à établir un lien direct entre longueur du document et métrique automatique, en calculant un score pour chaque document, plutôt qu'un score unique pour un groupe de documents. Nous utilisons pour cette analyse une variante du score s-BLEU. s-BLEU (*sentence-level BLEU*) est attribué à Lin & Och (2004) et consiste essentiellement à appliquer la métrique BLEU à chaque phrase, puis (éventuellement) à moyenner les scores au niveau d'un groupe de phrases ou de tout le corpus. Le calcul de s-BLEU impose toutefois de lisser les précisions n-grammes afin d'éviter les valeurs nulles. Il existe de multiples manières de réaliser ce lissage (Chen & Cherry, 2014), sans qu'aucune ne parvienne à faire de ce score une bonne évaluation de la qualité au niveau des phrases (Reiter, 2018). Notons que ce lissage est d'autant plus nécessaire que les séquences sont courtes et que la TA est de mauvaise qualité [8].

Calculer s-BLEU pour des documents (et des systèmes de bonne qualité) limite ces problèmes et permet d'obtenir un score par document, appelé dans la suite *ds-BLEU*. Dans nos analyses, nous traitons ces scores comme des réalisations d'une variable aléatoire, dont nous pouvons alors étudier la distribution et les relations avec d'autres variables, comme la longueur des documents sources.

4 Protocole expérimental

4.1 Données pour l'apprentissage et le test

Nos expérimentations considèrent des résumés et des transcriptions d'exposés dans le domaine du traitement automatique des langues, et se focalisent sur la traduction de l'anglais vers le français.

SciPar (Roussis *et al.*, 2022) est un corpus multilingue de phrases parallèles extraites de documents scientifiques collectés sur le Web, dont nous conservons uniquement la partie en–fr. Elle comprend 1,1M phrases, desquelles nous extrayons aléatoirement 3000 phrases pour la validation et le test.

TAL est constitué de résumés d'articles et de thèses dans le domaine du TAL, comprenant d'une part 1701 résumés de thèses récupérés de theses.fr et 1357 résumés d'articles extraits de ISTEX. Ces documents ont été segmentés avec Trankit (Nguyen *et al.*, 2021) et alignés phrase-à-phrase avec hunalign [9] (Varga *et al.*, 2005). Pour la traduction holistique, les phrases parallèles au sein de chaque document sont concaténées et traduites d'un bloc.

7. On constate d'ailleurs que d-BLEU semble augmenter (pour les systèmes Doc2Doc) avec la longueur des documents.
8. s-BLEU est récemment utilisé dans un contexte de TA de documents, par exemple dans Bao *et al.* (2021, tab. 2) : comme ce score est calculé phrase par phrase, il nécessite, comme pour BLEU, de réaligner références et traductions automatiques.
9. https://github.com/danielvarga/hunalign

Jeux de test : **THE** (dev et test) contient deux échantillons aléatoires de 101 et 100 résumés dans le domaine du TAL extraits également de theses.fr sans recouvrement avec **TAL**. **rTAL** contient enfin 246 résumés parallèles d'articles publiés dans la revue TAL. Ces articles sont alignés au niveau des phrases avec la même méthode que pour **TAL** ; ils ont également fait l'objet d'un filtrage avec TransQuest (Ranasinghe *et al.*, 2020) et d'une révision manuelle des alignements. Les statistiques du tableau 1 décrivent ces différents corpus. Pour analyser plus précisément les effets de longueur, nous avons enfin constitué un ensemble de 53 pseudo-documents, désigné par **IWSLT**, en segmentant 10 présentations orales transcrites, puis traduites, préparées pour la campagne IWSLT 2023 (Salesky *et al.*, 2023). La méthode utilisée pour construire ces pseudo-documents est décrite dans l'annexe B.

	SciPar			TAL				
	appr.	valid.	test	appr.	valid.	**THE**	**rTAL**	**IWSLT**
Nb. phrases	1116325	3000	3000	2858	101	100	246	53
Longueur moyenne des documents	37	38	37	265	317	327	129	402
Longueur moyenne des phrases dans un document	-	-	-	34	35	33	32	24

TABLE 1 – Statistiques des données d'apprentissage (appr.), de validation (valid.) et de test. La longueur est donnée en nombre de tokens calculés par le modèle BPE de MBART50(1-M).

4.2 Systèmes comparés

MBART50 (1-M) est un modèle encodeur-décodeur « classique » dérivé du modèle multilingue BART en poursuivant l'apprentissage avec des données parallèles associant anglais en source avec 49 langues en cible (Liu *et al.*, 2020b; Tang *et al.*, 2021). Ce modèle est dans un temps premier affiné avec les phrases parallèles du corpus **SciPar** (FT SCIPAR), puis adapté au TAL en présentant les exemples de **TAL** soit phrase-par-phrase (FT TAL-S), soit document-par-document (FT TAL-D). Un second système Doc2Doc (FT TAL-MR) est obtenu en augmentant **TAL-D** avec des pseudo-documents contenant des sous-parties des documents originaux (reproduisant l'apprentissage *multi-résolution* de Sun *et al.* (2022)). Les détails concernant l'affinage de MBART50 sont dans l'annexe C.

TOWERBASE (Alves *et al.*, 2024) est un grand modèle de langue dérivé de LLAMA2 (Touvron *et al.*, 2023) en continuant le préapprentissage avec des données multilingues en 10 langues, contenant une large proportion de données parallèles. Nous utilisons la version 7B, et l'amorce suivante : « English : SRC \n French : » en mode « zéro-exemple » [10]. Nous avons également analysé les traductions de TOWERINSTRUCT, les détails sont dans l'annexe E. À des fins de comparaison, nous utilisons la version professionnelle de DEEPL [11], sans adaptation, pour traduire les jeux de test **THE** et **rTAL**. Ce système est supposé fournir une indication de l'état-de-l'art. Dans nos expériences, nous fournissons au système des documents complets ; la sortie contient toujours autant de segments que l'entrée.

4.3 Méthodes

4.3.1 Analyse de l'impact de la longueur

Pour mesurer l'impact de la longueur sur le score de traduction, nous calculons pour chaque système la corrélation statistique entre ds-BLEU et la longueur du document : une corrélation négative indique

10. Nous avons également testé 3 exemples et 5 exemples sur le jeu de validation du **TAL**, sans amélioration de BLEU.
11. https://deepl.com

l'existence certaine d'un problème de longueur ; l'absence de corrélation indiquant qu'on ne peut pas conclure sur cette question. Un contraste intéressant pour cette analyse est de considérer des systèmes qui traduisent phrase par phrase et dont on peut considérer qu'ils n'ont aucun problème intrinsèque de longueur. Leur co-variation avec ds-BLEU donne donc une indication de l'impact des différents effets listés ci-dessus. Pour compléter cette analyse, nous reportons en annexe F une analyse des corrélations du score ds-BLEU avec la longueur moyenne des phrases dans un document.

4.3.2 Analyse des effets de position

La dégradation des scores BLEU avec la longueur peut être uniforme au sein d'un document, ou bien affecter surtout les phrases qui sont en fin de document. Analyser plus spécifiquement cet effet requiert de pouvoir comparer les scores de traduction pour des phrases au sein d'un même document. Pour ce faire, nous procédons comme suit : pour chaque document d, nous comparons trois situations impliquant une traduction holistique : (a) d est traduit isolément ; (b) d est traduit dans un pseudo-document, précédé de d' ; (c) d est traduit dans un pseudo-document, suivi de d', où d' un document sélectionné aléatoirement [12]. Les situations (b) et (c) sont éventuellement répétées pour plusieurs sélections de d'. Comparer ds-BLEU des situations (a) et (b) permet de mettre en évidence l'impact de la position. Dans la mesure où (b) introduit aussi un bruit aléatoire (lié à l'introduction du contexte d'), nous comparons également avec (c), dans lequel ce bruit est présent, mais sans changement de position par rapport à (a). Cette méthode demande simplement d'identifier dans la sortie de la traduction de dd' la frontière entre les deux documents, que nous obtenons par réalignement.

Une autre méthode consiste à recopier la première phrase de chaque document en position finale, puis de traduire de manière globale pour évaluer à quel point le changement de position (début vs. fin) modifie le texte généré. Cette méthode demande de réaligner pour identifier les frontières de phrases.

	Scores	MBART50(1-M)	FT SciPar	FT TAL-S	FT TAL-D	FT TAL-MR	TowerBase	DeepLPro
THE	BLEU*	29,0 (0,82)	1,1 (0,03)	1,7 (0,05)	43,3 (1,00)	43,0 (1,00)	39,9 (1,00)	44,5 (1,00)
	d-BLEU	32,0 (0,81)	0,9 (0,02)	1,7 (0,03)	45,6 (1,00)	45,3 (1,00)	42,2 (1,00)	46,8 (1,00)
	dS-BLEU	29,2 (0,77)	3,4 (0,07)	3,9 (0,08)	43,3 (0,98)	43,4 (0,98)	39,9 (0,97)	45,0 (0,98)
rTAL	BLEU*	21,1 (0,75)	3,9 (0,12)	4,7 (0,14)	34,9 (0,99)	35,0 (1,00)	31,9 (0,99)	36,0 (1,00)
	d-BLEU	23,2 (0,74)	4,1 (0,11)	5,0 (0,13)	36,5 (0,99)	36,7 (0,99)	33,5 (0,99)	37,7 (1,00)
	dS-BLEU	21,5 (0,73)	6,7 (0,19)	7,4 (0,21)	33,9 (0,95)	34,0 (0,96)	30,9 (0,95)	34,9 (0,96)
IWSLT	BLEU*	31,8 (0,79)	nan	nan	48,1 (0,97)	49,4 (0,98)	48,3 (0,98)	52,9 (1,00)
	d-BLEU	33,7 (0,78)	0,8 (0,02)	1,1 (0,02)	50,2 (0,96)	51,3 (0,98)	50,1 (0,98)	54,2 (1,00)
	dS-BLEU	32,8 (0,71)	3,4 (0,07)	4,0 (0,08)	49,9 (0,96)	50,8 (0,97)	50,6 (0,94)	52,6 (0,99)

TABLE 2 – Variantes du score BLEU (et pénalité de longueur), calculés sur les résumés de **THE**, de **rTAL** et les transcriptions de **IWSLT**. * indique qu'un réalignement est nécessaire pour l'évaluation.

5 Résultats et analyses

Le tableau 2 donne les scores BLEU [13] pour les trois corpus de test, traduits de manière holistique [14]. On constate en premier lieu, comme attendu, que d-BLEU est toujours supérieur au score BLEU

12. Modulo les contraintes sur la longueur cumulée de d et d' qui doit rester compatible avec les limites de l'encodeur.
13. Tous les scores BLEU et variantes sont calculés avec SacreBLEU (Post, 2018) version 2.4.0.
14. Pour **IWSLT**, certains scores sont absents, à cause de l'impossibilité de réaligner une sortie trop courte avec la référence.

(obtenu par réalignement), avec un écart de deux à trois points. Les systèmes entraînés à traduire des phrases ont des scores médiocres, en particulier causés par la très faible pénalité de longueur, démontrant leur inadéquation pour cette tâche [15]. Les systèmes entrainés pour les documents, y compris TOWERBASE, obtiennent des performances bien meilleures, proches de DEEPLPRO.

5.1 Impact de la longueur des documents

	DEEPLPRO	MBART50(1-M)	FT SCIPAR	FT TAL-S	FT TAL-D	FT TAL-MR	TOWERBASE
THE	0,100 (0,323)	0,139 (0,169)	**-0,416** (0,000)	**-0,309** (0,002)	0,100 (0,322)	0,078 (0,442)	0,080 (0,431)
rTAL	**0,214** (0,001)	**0,136** (0,032)	**-0,469** (0,000)	**-0,417** (0,000)	**0,237** (0,000)	**0,220** (0,001)	**0,249** (0,000)
IWSLT	0,099 (0,480)	0,010 (0,943)	**-0,532** (0,000)	**-0,500** (0,000)	0,002 (0,987)	-0,055 (0,694)	-0,151 (0,279)
THE diff	-	0,080 (0,429)	**-0,234** (0,019)	**-0,233** (0,020)	-0,094 (0,353)	-0,123 (0,221)	-0,162 (0,107)
rTAL diff	-	**-0,136** (0,033)	**-0,471** (0,000)	**-0,467** (0,000)	-0,009 (0,885)	-0,031 (0,625)	-0,021 (0,740)
IWSLT diff	-	-0,059 (0,673)	-0,164 (0,241)	-0,199 (0,154)	-0,093 (0,509)	-0,105 (0,453)	-0,252 (0,069)

TABLE 3 – Corrélation de Spearman (et *p-values*) entre ds-BLEU et la longueur des sources (L_s) (haut) ; corrélation de Spearman entre L_s et la différence des scores ds-BLEU entre chaque système et DEEPLPRO (bas).

Le tableau 3 présente les corrélations de Spearman relevées entre ds-BLEU et la longueur des documents. Pour DEEPLPRO, notre référence, on observe une légère corrélation positive (significative pour **rTAL**), induite par le biais de ds-BLEU en faveur des documents plus longs (§2.3). Des résultats similaires sont obtenus pour les systèmes holistiques (FT TAL-D, FT TAL-MR et TOWERBASE), laissant penser qu'ils n'ont pas plus de problèmes de longueur que DEEPLPRO. Cette tendance s'inverse pour **IWSLT**, dont la distribution de longueur est plus équilibrée. Les déficiences des systèmes entraînés avec des phrases isolées se traduisent par de fortes corrélations négatives. La partie inférieure du tableau vise à neutraliser les effets liés à la difficulté intrinsèque de chaque document, en soustrayant de chaque valeur de ds-BLEU le score obtenu par DEEPLPRO. Nous observons alors que toutes les corrélations pour les systèmes holistiques deviennent négatives (de manière non significative) : plus les documents sont longs, plus les scores ds-BLEU s'écartent de ceux de DEEPLPRO, ce qui suggère que la longueur reste un facteur de difficulté pour la traduction des documents holistiques.

5.2 Analyse des changements de position

Les graphes de la figure 2 mettent en évidence un effet de position très clair pour les deux systèmes dérivés de MBART : décaler un document de quelques centaines de tokens dans l'encodeur cause une perte moyenne d'environ 1,5 points pour la métrique ds-BLEU, alors qu'ajouter un second document à la suite introduit un léger bruit moyen et une dégradation faible de ds-BLEU. Pour TOWERBASE, les deux transformations sont également problématiques, ce qui s'explique par une architecture différente (décodeur pur) qui est plus impactée par l'adjonction du document d'.

Ces résultats sont confirmés par le tableau 5 où nous comparons les traductions des phrases en position initiale avec celle de leur copie en position finale : les premières sont légèrement meilleures que les secondes, surtout pour le corpus **THE**, dont les documents sont plus longs que **rTAL**.

15. L'annexe D montre que lorsque l'on traduit phrase à phrase, ces systèmes obtiennent les meilleurs scores BLEU.

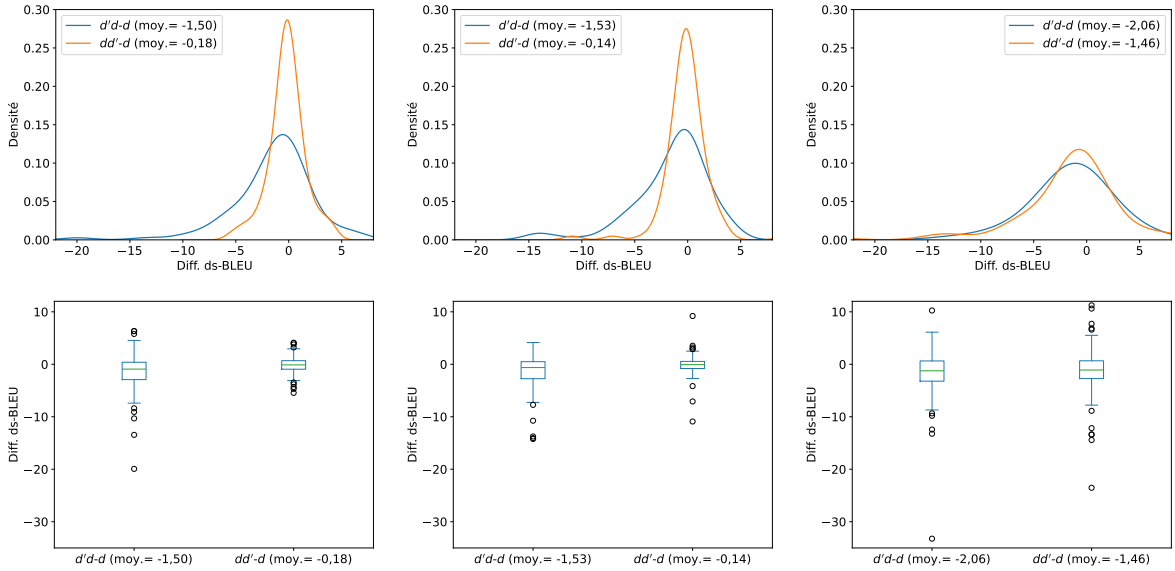


FIGURE 2 – Densité (haut) et boîtes à moustaches (bas) de la différence entre ds-BLEU des documents traduits par les méthodes (b) ou (c) et par la méthode (a) (voir §4.3) : FT TAL-D (gauche), FT TAL-MR (milieu) et TOWEBASE (droite).

		BLEU	s-BLEU	d-BLEU	ds-BLEU
FT TAL-D	$d'd$	40,1 (0,98)	37,6 (0,92)	42,9 (0,98)	41,8 (0,97)
	dd'	43,0 (1,00)	40,7 (0,94)	45,2 (1,00)	43,1 (0,98)
FT TAL-MR	$d'd$	40,6 (1,00)	38,1 (0,93)	43,3 (1,00)	41,9 (0,97)
	dd'	42,8 (1,00)	40,6 (0,95)	45,1 (1,00)	43,3 (0,98)
TOWERBASE	$d'd$	38,4 (0,98)	35,9 (0,91)	40,6 (0,98)	37,9 (0,94)
	dd'	37,8 (0,98)	35,0 (0,89)	40,2 (0,98)	38,5 (0,94)

TABLE 4 – Variantes de BLEU, calculées sur les documents d en position initiale (i.e. dd'), ou finale (i.e. $d'd$). Les scores sont moyennés sur 6 répétitions en variant le choix de d'.

		MBART50(1-M)	FT TAL-D	FT TAL-MR
THE	début	36,7 (0,94)	44,5 (0,99)	45,0 (0,99)
	fin	14,7 (0,45)	43,1 (0,99)	43,7 (0,98)
	début :fin	33.1 (0,50)	90,9 (0,99)	91,5 (0,99)
rTAL	début	25,7 (0,92)	34,7 (0,98)	35,6 (0,98)
	fin	4,8 (0,30)	34,8 (0,98)	35,6 (0,98)
	début :fin	17,2 (0,36)	95,0 (1,00)	95,6 (1,00)

TABLE 5 – Score BLEU calculé sur l'ensemble des premières phrases, traduites respectivement au début et à la fin. "début :fin" évalue (avec BLEU) la distance entre les deux traductions.

6 Conclusion

La traduction au niveau du document semble à notre portée, mais il reste des défis à relever, notamment en ce qui concerne les longs documents. Dans cette étude, nous avons examiné certains des avantages et inconvénients théoriques de la traduction automatique holistique et exploré les différentes manières dont le score BLEU peut être utilisé pour évaluer la traduction au niveau du document. Notre étude expérimentale indique qu'il existe toujours un effet négatif visible de la longueur du document sur la qualité de la traduction, comme le montre le score BLEU, et cet effet négatif semble croître lorsque la longueur du document augmente. Parmi les pistes de travail, mentionnons l'étude des effets du choix de l'encodage positionnel pour les documents longs et l'impact de la longueur des documents vus à l'apprentissage sur la capacité du modèle à traduire des documents de test de longueur variable.

Remerciements

Nous adressons nos remerciements à Mathilde Huguin pour l'extraction des données brutes de ISTEX et à Maxime Bouthors pour les données brutes de theses.fr. Nous remercions également Jean-François Nominé pour les traductions avec DeepLPro. Nous remercions enfin Paul Lerner pour ses retours sur une version préliminaire de cet article. Ces travaux sont financés par l'Agence Nationale de la Recherche (ANR) dans le cadre du projet MaTOS. La contribution de R. Bawden a été partiellement financée par sa chaire à l'institut PRAIRIE financé par l'agence nationale française ANR dans le cadre du programme "Investissements d'avenir" sous la référence ANR-19- P3IA-0001.

Références

ALVES D. M., POMBAL J., GUERREIRO N. M., MARTINS P. H., ALVES J., FARAJIAN A., PETERS B., REI R., FERNANDES P., AGRAWAL S., COLOMBO P., DE SOUZA J. G. C. & MARTINS A. F. T. (2024). Tower : An open multilingual large language model for translation-related tasks.

BAO G., ZHANG Y., TENG Z., CHEN B. & LUO W. (2021). G-transformer for document-level machine translation. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 3442–3455, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.267.

BAWDEN R., SENNRICH R., BIRCH A. & HADDOW B. (2018). Evaluating discourse phenomena in neural machine translation. In M. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 1304–1313, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1118.

CALLISON-BURCH C., OSBORNE M. & KOEHN P. (2006). Re-evaluating the role of Bleu in machine translation research. In D. MCCARTHY & S. WINTNER, Éds., *11th Conference of the European Chapter of the Association for Computational Linguistics*, p. 249–256, Trento, Italy : Association for Computational Linguistics.

CHEN B. & CHERRY C. (2014). A systematic comparison of smoothing techniques for sentence-level BLEU. In O. BOJAR, C. BUCK, C. FEDERMANN, B. HADDOW, P. KOEHN, C. MONZ, M. POST & L. SPECIA, Éds., *Proceedings of the Ninth Workshop on Statistical Machine Translation*, p. 362–367, Baltimore, Maryland, USA : Association for Computational Linguistics. DOI : 10.3115/v1/W14-3346.

HENDY A., ABDELREHIM M., SHARAF A., RAUNAK V., GABR M., MATSUSHITA H., KIM Y. J., AFIFY M. & AWADALLA H. H. (2023). How good are GPT models at machine translation ? a comprehensive evaluation.

HEROLD C. & NEY H. (2023). Improving long context document-level machine translation. In M. STRUBE, C. BRAUD, C. HARDMEIER, J. J. LI, S. LOAICIGA & A. ZELDES, Éds., *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, p. 112–125, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.codi-1.15.

JUNCZYS-DOWMUNT M. (2019). Microsoft translator at WMT 2019 : Towards large-scale document-level neural machine translation. In O. BOJAR, R. CHATTERJEE, C. FEDERMANN, M. FISHEL, Y. GRAHAM, B. HADDOW, M. HUCK, A. J. YEPES, P. KOEHN, A. MARTINS, C. MONZ,

M. NEGRI, A. NÉVÉOL, M. NEVES, M. POST, M. TURCHI & K. VERSPOOR, Éds., *Proceedings of the Fourth Conference on Machine Translation (Volume 2 : Shared Task Papers, Day 1)*, p. 225–233, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-5321.

KARPINSKA M. & IYYER M. (2023). Large language models effectively leverage document-level context for literary translation, but critical errors persist. In P. KOEHN, B. HADDOW, T. KOCMI & C. MONZ, Éds., *Proceedings of the Eighth Conference on Machine Translation*, p. 419–451, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.wmt-1.41.

KIM Y., TRAN D. T. & NEY H. (2019). When and why is document-level context useful in neural machine translation ? In A. POPESCU-BELIS, S. LOÁICIGA, C. HARDMEIER & D. XIONG, Éds., *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, p. 24–34, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-6503.

KOEHN P. & KNOWLES R. (2017). Six challenges for neural machine translation. In T. LUONG, A. BIRCH, G. NEUBIG & A. FINCH, Éds., *Proceedings of the First Workshop on Neural Machine Translation*, p. 28–39, Vancouver : Association for Computational Linguistics. DOI : 10.18653/v1/W17-3204.

LI Y., LI J., JIANG J., TAO S., YANG H. & ZHANG M. (2022). P-Transformer : Towards Better Document-to-Document Neural Machine Translation. arXiv :2212.05830 [cs].

LIBOVICKÝ J., HELCL J. & MAREČEK D. (2018). Input combination strategies for multi-source transformer decoder. In O. BOJAR, R. CHATTERJEE, C. FEDERMANN, M. FISHEL, Y. GRAHAM, B. HADDOW, M. HUCK, A. J. YEPES, P. KOEHN, C. MONZ, M. NEGRI, A. NÉVÉOL, M. NEVES, M. POST, L. SPECIA, M. TURCHI & K. VERSPOOR, Éds., *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 253–260, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6326.

LIN C.-Y. & OCH F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, p. 605–612, Barcelona, Spain. DOI : 10.3115/1218955.1219032.

LIU C., ZHANG Q., ZHANG X., SINGH K., SARAF Y. & ZWEIG G. (2020a). Multilingual graphemic hybrid ASR with massive data augmentation. In D. BEERMANN, L. BESACIER, S. SAKTI & C. SORIA, Éds., *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, p. 46–52, Marseille, France : European Language Resources association.

LIU Y., GU J., GOYAL N., LI X., EDUNOV S., GHAZVININEJAD M., LEWIS M. & ZETTLEMOYER L. (2020b). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, **8**, 726–742. DOI : 10.1162/tacl_a_00343.

LOPES A., FARAJIAN M. A., BAWDEN R., ZHANG M. & MARTINS A. F. T. (2020). Document-level neural MT : A systematic comparison. In A. MARTINS, H. MONIZ, S. FUMEGA, B. MARTINS, F. BATISTA, L. COHEUR, C. PARRA, I. TRANCOSO, M. TURCHI, A. BISAZZA, J. MOORKENS, A. GUERBEROF, M. NURMINEN, L. MARG & M. L. FORCADA, Éds., *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, p. 225–234, Lisboa, Portugal : European Association for Machine Translation.

LUPO L., DINARELLI M. & BESACIER L. (2023). Encoding sentence position in context-aware neural machine translation with concatenation. In S. TAFRESHI, A. AKULA, J. SEDOC, A. DROZD, A. ROGERS & A. RUMSHISKY, Éds., *The Fourth Workshop on Insights from Negative Results in NLP*, p. 33–44, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.insights-1.4.

MA Z., EDUNOV S. & AULI M. (2021). A comparison of approaches to document-level machine translation.

MARUF S., SALEH F. & HAFFARI G. (2021). A Survey on Document-Level Neural Machine Translation : Methods and Evaluation. *ACM Comput. Surv.*, **54**(2). Place : New York, NY, USA Publisher : Association for Computing Machinery, DOI : 10.1145/3441691.

MATHUR N., BALDWIN T. & COHN T. (2020). Tangled up in BLEU : Reevaluating the evaluation of automatic machine translation evaluation metrics. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 4984–4997, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.448.

MATUSOV E., LEUSCH G., BENDER O. & NEY H. (2005). Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

MIHAYLOVA T. & MARTINS A. F. T. (2019). Scheduled sampling for transformers. In F. ALVA-MANCHEGO, E. CHOI & D. KHASHABI, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop*, p. 351–356, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-2049.

NEVES M., JIMENO YEPES A., NÉVÉOL A., BAWDEN R., DI NUNZIO G. M., ROLLER R., THOMAS P., VEZZANI F., VICENTE NAVARRO M., YEGANOVA L., WIEMANN D. & GROZEA C. (2023). Findings of the WMT 2023 biomedical translation shared task : Evaluation of ChatGPT 3.5 as a comparison system. In P. KOEHN, B. HADDOW, T. KOCMI & C. MONZ, Éds., *Proceedings of the Eighth Conference on Machine Translation*, p. 43–54, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.wmt-1.2.

NGUYEN M. V., LAI V. D., POURAN BEN VEYSEH A. & NGUYEN T. H. (2021). Trankit : A light-weight transformer-based toolkit for multilingual natural language processing. In D. GKATZIA & D. SEDDAH, Éds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : System Demonstrations*, p. 80–90, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-demos.10.

OTT M., EDUNOV S., BAEVSKI A., FAN A., GROSS S., NG N., GRANGIER D. & AULI M. (2019). fairseq : A fast, extensible toolkit for sequence modeling. In W. AMMAR, A. LOUIS & N. MOSTAFAZADEH, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, p. 48–53, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-4009.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In P. ISABELLE, E. CHARNIAK & D. LIN, Éds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

POPOVIĆ M. (2015). chrF : character n-gram F-score for automatic MT evaluation. In O. BOJAR, R. CHATTERJEE, C. FEDERMANN, B. HADDOW, C. HOKAMP, M. HUCK, V. LOGACHEVA & P. PECINA, Éds., *Proceedings of the Tenth Workshop on Statistical Machine Translation*, p. 392–395, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/W15-3049.

POST M. (2018). A call for clarity in reporting BLEU scores. In O. BOJAR, R. CHATTERJEE, C. FEDERMANN, M. FISHEL, Y. GRAHAM, B. HADDOW, M. HUCK, A. J. YEPES, P. KOEHN, C. MONZ, M. NEGRI, A. NÉVÉOL, M. NEVES, M. POST, L. SPECIA, M. TURCHI & K. VERSPOOR, Éds., *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 186–191, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6319.

POST M. & JUNCZYS-DOWMUNT M. (2023). Escaping the sentence-level paradigm in machine translation. arXiv :2304.12959 [cs].

RANASINGHE T., ORASAN C. & MITKOV R. (2020). TransQuest : Translation quality estimation with cross-lingual transformers. In D. SCOTT, N. BEL & C. ZONG, Éds., *Proceedings of the 28th International Conference on Computational Linguistics*, p. 5070–5081, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.445.

RANZATO M., CHOPRA S., AULI M. & ZAREMBA W. (2016). Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016*, San Juan, Puerto Rico.

REI R., C. DE SOUZA J. G., ALVES D., ZERVA C., FARINHA A. C., GLUSHKOVA T., LAVIE A., COHEUR L. & MARTINS A. F. T. (2022). COMET-22 : Unbabel-IST 2022 submission for the metrics shared task. In P. KOEHN, L. BARRAULT, O. BOJAR, F. BOUGARES, R. CHATTERJEE, M. R. COSTA-JUSSÀ, C. FEDERMANN, M. FISHEL, A. FRASER, M. FREITAG, Y. GRAHAM, R. GRUNDKIEWICZ, P. GUZMAN, B. HADDOW, M. HUCK, A. JIMENO YEPES, T. KOCMI, A. MARTINS, M. MORISHITA, C. MONZ, M. NAGATA, T. NAKAZAWA, M. NEGRI, A. NÉVÉOL, M. NEVES, M. POPEL, M. TURCHI & M. ZAMPIERI, Éds., *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 578–585, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics.

REITER E. (2018). A structured review of the validity of BLEU. *Computational Linguistics*, **44**(3), 393–401. DOI : 10.1162/coli_a_00322.

ROUSSIS D., PAPAVASSILIOU V., PROKOPIDIS P., PIPERIDIS S. & KATSOUROS V. (2022). SciPar : A collection of parallel corpora from scientific abstracts. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 2652–2657, Marseille, France : European Language Resources Association.

SALESKY E., DARWISH K., AL-BADRASHINY M., DIAB M. & NIEHUES J. (2023). Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In E. SALESKY, M. FEDERICO & M. CARPUAT, Éds., *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, p. 62–78, Toronto, Canada (in-person and online) : Association for Computational Linguistics. DOI : 10.18653/v1/2023.iwslt-1.2.

SUN Z., WANG M., ZHOU H., ZHAO C., HUANG S., CHEN J. & LI L. (2022). Rethinking document-level neural machine translation. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Findings of the Association for Computational Linguistics : ACL 2022*, p. 3537–3548, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-acl.279.

TANG Y., TRAN C., LI X., CHEN P.-J., GOYAL N., CHAUDHARY V., GU J. & FAN A. (2021). Multilingual translation from denoising pre-training. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 3450–3466, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.304.

TAY Y., DEHGHANI M., BAHRI D. & METZLER D. (2023). Efficient Transformers : A Survey. *ACM Computing Surveys*, **55**(6), 1–28. DOI : 10.1145/3530811.

TIEDEMANN J. & SCHERRER Y. (2017). Neural machine translation with extended context. In B. WEBBER, A. POPESCU-BELIS & J. TIEDEMANN, Éds., *Proceedings of the Third Workshop on Discourse in Machine Translation*, p. 82–92, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/W17-4811.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S., BIKEL D., BLECHER L., FERRER C. C., CHEN M., CUCURULL G., ESIOBU D., FERNANDES J., FU J., FU W., FULLER B., GAO C., GOSWAMI V., GOYAL N., HARTSHORN A., HOSSEINI S., HOU R., INAN H., KARDAS M., KERKEZ V., KHABSA M., KLOUMANN I., KORENEV A., KOURA P. S., LACHAUX M.-A., LAVRIL T., LEE J., LISKOVICH D., LU Y., MAO Y., MARTINET X., MIHAYLOV T., MISHRA P., MOLYBOG I., NIE Y., POULTON A., REIZENSTEIN J., RUNGTA R., SALADI K., SCHELTEN A., SILVA R., SMITH E. M., SUBRAMANIAN R., TAN X. E., TANG B., TAYLOR R., WILLIAMS A., KUAN J. X., XU P., YAN Z., ZAROV I., ZHANG Y., FAN A., KAMBADUR M., NARANG S., RODRIGUEZ A., STOJNIC R., EDUNOV S. & SCIALOM T. (2023). Llama 2 : Open foundation and fine-tuned chat models.

VARGA D., HALAÁCSY P., KORNAI A., NAGY V., NÉMETH L. & TRÓN V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005 Conference*, p. 590–596.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is All you Need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 30*, p. 5998–6008 : Curran Associates, Inc.

WANG L., LYU C., JI T., ZHANG Z., YU D., SHI S. & TU Z. (2023). Document-level machine translation with large language models. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 16646–16661, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.1036.

WICKS R. & POST M. (2022). Does sentence segmentation matter for machine translation ? In P. KOEHN, L. BARRAULT, O. BOJAR, F. BOUGARES, R. CHATTERJEE, M. R. COSTA-JUSSÀ, C. FEDERMANN, M. FISHEL, A. FRASER, M. FREITAG, Y. GRAHAM, R. GRUNDKIEWICZ, P. GUZMAN, B. HADDOW, M. HUCK, A. JIMENO YEPES, T. KOCMI, A. MARTINS, M. MORISHITA, C. MONZ, M. NAGATA, T. NAKAZAWA, M. NEGRI, A. NÉVÉOL, M. NEVES, M. POPEL, M. TURCHI & M. ZAMPIERI, Éds., *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 843–854, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics.

WU M., WANG Y., FOSTER G. F., QU L. & HAFFARI G. (2024). Importance-aware data augmentation for document-level neural machine translation. *CoRR*, **abs/2401.15360**. DOI : 10.48550/ARXIV.2401.15360.

ZHANG B., HADDOW B. & BIRCH A. (2023). Prompting large language model for machine translation : A case study. In A. KRAUSE, E. BRUNSKILL, K. CHO, B. ENGELHARDT, S. SABATO & J. SCARLETT, Éds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 de *Proceedings of Machine Learning Research*, p. 41092–41110 : PMLR.

ZHANG J., LUAN H., SUN M., ZHAI F., XU J., ZHANG M. & LIU Y. (2018). Improving the transformer translation model with document-level context. In E. RILOFF, D. CHIANG, J. HOCKENMAIER & J. TSUJII, Éds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 533–542, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1049.

ZHUOCHENG Z., GU S., ZHANG M. & FENG Y. (2023). Addressing the length bias challenge in document-level neural machine translation. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 11545–11556, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.773.

A Algorithme de réalignement

Le calcul des scores BLEU et s-BLEU pour les systèmes de traduction holistiques demande un réalignement des phrases traduites automatiquement avec les phrases de la référence. Notre stratégie de réalignement est présentée dans l'algorithme 1. Elle consiste essentiellement à aligner les traductions automatiques avec les références, puis à segmenter les sorties automatiques au niveau des frontières de phrase de la référence. L'alignement est réalisé au niveau des caractères avec la bibliothèque Python `edlib`.

Pour évaluer l'exactitude de l'algorithme 1, nous avons traduit les phrases de **THE-S** séparément et calculé un premier score BLEU. Nous avons ensuite concaténé les sorties pour former des documents complets, que nous avons réalignés sur les phrases sources pour obtenir **THE-S-D-S**. Le tableau 6 donne la différence des scores BLEU évalués sur ces deux traductions, qui est toujours très faible (inférieure à 0,2), montrant que la segmentation obtenue par réalignement est très proche de la segmentation initiale. Les mêmes observations valent pour **rTAL**.

	THE-S	**THE-S-D-S**	**rTAL-S**	**rTAL-S-D-S**
MBART50(1-M)	36,2 (1,00)	36,0 (1,00)	29,4 (1,00)	29,2 (1,00)
FT SciPar	42,8 (1,00)	42,7 (1,00)	34,1 (0,99)	33,9 (0,99)
FT TAL-S	44,1 (1,00)	43,9 (1,00)	34,7 (0,98)	34,5 (0,98)
FT TAL-D	44,2 (1,00)	44,1 (1,00)	34,9 (0,99)	34,7 (0,99)
FT TAL-MR	43,9 (1,00)	43,8 (1,00)	35,2 (0,99)	35,0 (0,99)

TABLE 6 – Scores BLEU calculés sur **THE-S** et **THE-S-D-S**, où nous concaténons les phrases de **THE-S** pour former des documents, puis réalignons avec l'algorithme 1, De même pour **rTAL-S**.

Algorithm 1: Calcul d'un réalignement au niveau des phrases

Data: SYS : l'ensemble des traductions produites automatiquement.
Data: [sep] : un délimiteur qui sépare les phrases au sein des documents.
Data: REF : les traductions de référence, dans lesquelles SEP est inséré aux frontières de phrases
Result: SYS_{sent}, REF_{sent} : alignement de phrases entre SYS et REF
begin
  $N \longleftarrow$ nombre de documents dans REF
  $\text{SYS}_{sent} \longleftarrow$ liste vide
  $\text{REF}_{sent} \longleftarrow$ liste vide
  for $I \in \{1, \ldots, N\}$ **do**
    $D_R \longleftarrow$ le I$^{\text{ème}}$ document de REF
    $D_S \longleftarrow$ le I$^{\text{ème}}$ document de SYS
    $I_{SEP} \longleftarrow$ liste des positions de [sep] dans D_R
    $D_R \longleftarrow$ supprimer [sep] de D_R
    dériver l'alignement optimal de D_R et D_S du calcul de la distance de Levenshtein
    $I_{split} \longleftarrow$ indices des caractères de D_R alignés avec les positions de I_{SEP}
    $\text{SYS}_{sent} \longleftarrow$ segmenter D_S en phrases aux positions de I_{split}
    $\text{REF}_{sent} \longleftarrow$ segmenter D_R en phrases aux positions de I_{SEP}
  end
end

B Construction du corpus IWSLT

À partir des 10 documents du corpus **IWSLT**, nous construisons un ensemble de 53 pseudo-documents équilibrés en longueur de la manière suivante. Soit L l'ensemble des longueurs $L = \{32, 24, 16, 8, 4\}$, nous construisons pour chaque document D une permutation aléatoire de L notée $L(D)$, puis segmentons d selon $L(D)$ en affectant les $L(D)[1]$ premières phrases au pseudo-document D_1, puis les $L(D)[2]$ phrases suivantes au pseudo-document D_2, etc. Les phrases en excédent (au-delà de 84) sont affectées au pseudo-document D_5.

Nb. phrases	4	6	7	8	12	16	21	23	24	32	33
Effectif	10	2	1	8	1	9	1	1	10	9	1
Longueur moyenne des documents	99	172	118	198	490	349	573	601	584	777	657
Longueur moyenne des phrases dans un document	25	29	17	25	41	22	27	26	24	24	20

TABLE 7 – Statistiques de **IWSLT** : pour chaque longueur de documents (en nombre de phrases) nous donnons la longueur moyenne (en nombre d'unités) des documents et la longueur moyenne (en nombre d'unités) des phrases dans chaque document.

C Affinage de MBART

L'affinage de MBART50(1-M) est implémenté avec *fairseq* (Ott *et al.*, 2019). Tous les modèles sont constitués de 12 couches pour l'encodeur et le décodeur, de dimension 1024 avec 16 têtes d'attention. Les systèmes sont entraînés avec un GPU de type `NVIDIA RTX A6000 48G` et 12 CPU avec chacun 8G de mémoire. La taille des lots est fixée à 4096 avec une mise à jour tous les 4 lots pour FT SCIPAR, et à 2048 avec une mise à jour tous les deux lots pour les autres systèmes. Afin d'éviter le sur-apprentissage, nous utilisons une procédure d'arrêt précoce avec une patience de 5 époques, en fonction des scores BLEU évalués sur le jeu de validation.

La première étape consiste à entraîner FT SCIPAR en affinant MBART50(1-M) avec le corpus **SciPar** pour réaliser une adaptation au domaine scientifique. À partir de FT SCIPAR, nous avons continué à affiner le système avec les documents du domaine du TAL, présentés soit document par document (FT TAL-D) soit phrase par phrase (FT TAL-S).

Pour implanter l'approche multi-résolution (MR) de (Sun *et al.*, 2022), un jeu d'apprentissage **TAL-MR** est construit avec les documents du corpus **TAL**, avec lequel nous dérivons FT TAL-MR par affinage de FT SCIPAR. Cette approche MR consiste à constituer un jeu d'apprentissage des sous-documents de longueur équilibrée en coupant chaque document (de **TAL-D** dans notre cas) en K sous-documents plusieurs fois, avec $K \in \{1, 2, 4, 8, \dots\}$. C'est-à-dire, un document de longueur 8 est réparti en 15 sous-parties, avec un document de 8 phrases, 2 sous-documents de 4 phrases, 4 sous-documents de 2 phrases et 8 sous-documents d'une phrase.

Pour le décodage avec `fairseq-interactive`, nous utilisons les valeurs de paramètres `max-len-a`=1,5 au lieu de la valeur par défaut (1,2), et `max-len-b`=10. Ces deux paramètres servent à contrôler la longueur des phrases générées par traduction.

D Analyse des systèmes MBART

Le tableau 8 présente les résultats de traduction pour l'ensemble des systèmes dérivés de MBART50. Pour tous les systèmes, l'adaptation au registre scientifique, puis au domaine du TAL a de larges effets positifs. Traduire phrase par phrase donne des scores d-BLEU assez comparables pour les trois systèmes adaptés. Les traductions automatiques sont d'une longueur adéquate. Il en va tout autrement pour les deux systèmes entraînés sur des phrases isolées, qui ont des pénalités de longueur proches de zéro et des scores d-BLEU insignifiants. L'affinage par document suffit à corriger ce problème et conduit à des performances légèrement inférieures (pour **THE**) ou comparables (pour **rTAL**) à la traduction phrase à phrase.

	Sent2Sent		Doc2Doc	
	THE	**rTAL**	**THE**	**rTAL**
MBART50 (1-M)	38,3 (1,0)	31,0 (1,0)	32,0 (0,8)	23,3 (0,7)
FT SciPar	45,0 (1,0)	35,7 (1,0)	0,9 (0,0)	4,1 (0,1)
FT TAL-S	46,2 (1,0)	36,2 (1,0)	1,7 (0,0)	5,1 (0,1)
FT TAL-D	46,4 (1,0)	36,5 (1,0)	45,6 (1,0)	36,6 (1,0)
FT TAL-MR	46,0 (1,0)	36,8 (1,0)	45,3 (1,0)	36,8 (1,0)

TABLE 8 – d-BLEU et pénalité de longueur pour MBART50 (1-M) et l'ensemble des modèles affinés pour les données de test traduites par phrases et par documents.

E Résultats avec TowerInstruct

TowerInstruct est développé pour les tâches concernant la TA, en affinant TowerBase avec instruction sur les jeux de données **TOWERBLOCKS**. **TOWERBLOCKS** contient une grande partie des données pour la traduction au niveau des phrases, la détection des erreurs, conversation et code. Il inclut également une petite portion des corpus parallèles pour la traduction en contexte (Alves *et al.*, 2024, fig.3). Nous utilisons la version 7B de TowerInstruct, et l'amorce suivante [16] : « Translate the following text from French into English.\n English : SRC \n French : » en mode « zéro-exemple ».

Nous reportons pour tous les systèmes les scores BLEU dans le tableau 9, les corrélations de Spearman entre ds-BLEU et la longueur de source dans le tableau 10. Pour **THE** et **rTAL**, les traductions de TowerInstruct sont meilleures que TowerBase en terme des scores BLEU. La corrélation entre ds-BLEU et la longueur de la source montre la même tendance précédemment, même si elle est moins forte que pour les autres systèmes holistiques. Pour **IWSLT**, le score ds-BLEU de TowerInstruct est inférieur à celui de TowerBase. La corrélation positive (non significative) devient moins faible lorsque nous soustrayons le ds-BLEU de TowerInstruct avec celui de DeepLPro.

Le tableau 11 rassemble les évaluations des méthodes de traduction (a), (b) et (c) présentées dans §4.3 pour tous les systèmes, y compris TowerInstruct. Nous constatons que la différence entre les ds-BLEU des traductions par les méthodes (b) et (c) a diminué par rapport aux résultats de TowerBase. Ce phénomène est aussi illustré dans la figure 3. Il montre que la performance de TowerInstruct est plus stable pour traduire des textes situant aux différentes positions dans notre scénario, avec des séquences d'entrée $d'd$ ou dd' moins longues que 1024.

16. https://huggingface.co/Unbabel/TowerInstruct-7B-v0.1#prompt-format

	Scores	MBART50(1-M)	FT SciPar	FT TAL-S	FT TAL-D	FT TAL-MR	TowerBase	TowerInstruct	DeepLPro
THE	BLEU*	29,0 (0,82)	1,1 (0,03)	1,7 (0,05)	43,3 (1,00)	43,0 (1,00)	39,9 (1,00)	41,0 (1,00)	44,5 (1,00)
	d-BLEU	32,0 (0,81)	0,9 (0,02)	1,7 (0,03)	45,6 (1,00)	45,3 (1,00)	42,2 (1,00)	43,3 (1,00)	46,8 (1,00)
	ds-BLEU	29,2 (0,77)	3,4 (0,07)	3,9 (0,08)	43,3 (0,98)	43,4 (0,98)	39,9 (0,97)	41,6 (0,98)	45,0 (0,98)
rTAL	BLEU*	21,1 (0,75)	3,9 (0,12)	4,7 (0,14)	34,9 (0,99)	35,0 (1,00)	31,9 (0,99)	33,3 (1,00)	36,0 (1,00)
	d-BLEU	23,2 (0,74)	4,1 (0,11)	5,0 (0,13)	36,5 (0,99)	36,7 (0,99)	33,5 (0,99)	35,0 (1,00)	37,7 (1,00)
	ds-BLEU	21,5 (0,73)	6,7 (0,19)	7,4 (0,21)	33,9 (0,95)	34,0 (0,96)	30,9 (0,95)	32,3 (0,96)	34,9 (0,96)
IWSLT	BLEU*	31,8 (0,79)	nan (nan)	nan (nan)	48,1 (0,97)	49,4 (0,98)	48,3 (0,98)	48,5 (0,98)	52,9 (1,00)
	d-BLEU	33,7 (0,78)	0,8 (0,02)	1,1 (0,02)	50,2 (0,96)	51,3 (0,98)	50,1 (0,98)	50,1 (0,98)	54,2 (1,00)
	ds-BLEU	32,8 (0,71)	3,4 (0,07)	4,0 (0,08)	49,9 (0,96)	50,8 (0,97)	50,6 (0,94)	48,5 (0,97)	52,6 (0,99)

TABLE 9 – Variantes du score BLEU (et pénalité de longueur), calculés sur les résumés de **THE**, de **rTAL** et les transcriptions de **IWSLT**. * indique qu'un réalignement est nécessaire pour l'évaluation.

	DeepLPro	MBART50(1-M)	FT SciPar	FT TAL-S	FT TAL-D	FT TAL-MR	TowerBase	TowerInstruct
THE	0,100 (0,323)	0,139 (0,169)	**-0,416** (0,000)	**-0,309** (0,002)	0,100 (0,322)	0,078 (0,442)	0,080 (0,431)	0,066 (0,512)
rTAL	**0,214** (0,001)	**0,136** (0,032)	**-0,469** (0,000)	**-0,417** (0,000)	**0,237** (0,000)	**0,220** (0,001)	**0,249** (0,000)	**0,231** (0,000)
IWSLT	0,099 (0,480)	0,010 (0,943)	**-0,532** (0,000)	**-0,500** (0,000)	0,002 (0,987)	-0,055 (0,694)	-0,151 (0,279)	0,234 (0,092)
THE diff	-	0,080 (0,429)	**-0,234** (0,019)	**-0,233** (0,020)	-0,094 (0,353)	-0,123 (0,221)	-0,162 (0,107)	-0,069 (0,494)
rTAL diff	-	**-0,136** (0,033)	**-0,471** (0,000)	**-0,467** (0,000)	-0,009 (0,885)	-0,031 (0,625)	-0,021 (0,740)	-0,025 (0,697)
IWSLT diff	-	-0,059 (0,673)	-0,164 (0,241)	-0,199 (0,154)	-0,093 (0,509)	-0,105 (0,453)	-0,252 (0,069)	0,023 (0,869)

TABLE 10 – Corrélation de Spearman (et *p-values*) entre ds-BLEU et la longueur des sources (L_s) (haut) ; corrélation de Spearman entre L_s et la différence des scores ds-BLEU entre chaque système et DeepLPro (bas).

		BLEU	s-BLEU	d-BLEU	ds-BLEU
FT TAL-D	d	43,3 (1,00)	40,9 (0,94)	45,6 (1,00)	43,3 (0,98)
	$d'd$	40,1 (0,98)	37,6 (0,92)	42,9 (0,98)	41,8 (0,97)
	dd'	43,0 (1,00)	40,7 (0,94)	45,2 (1,00)	43,1 (0,98)
FT TAL-MR	d	43,0 (1,00)	41,0 (0,95)	45,3 (1,00)	43,4 (0,98)
	$d'd$	40,6 (1,00)	38,1 (0,93)	43,3 (1,00)	41,9 (0,97)
	dd'	42,8 (1,00)	40,6 (0,95)	45,1 (1,00)	43,3 (0,98)
TowerBase	d	39,9 (1,00)	37,2 (0,93)	42,2 (1,00)	39,9 (0,97)
	$d'd$	38,4 (0,98)	35,9 (0,91)	40,6 (0,98)	37,9 (0,94)
	dd'	37,8 (0,98)	35,0 (0,89)	40,2 (0,98)	38,5 (0,94)
TowerInstruct	d	41,0 (1,00)	39,0 (0,95)	43,3 (1,00)	41,6 (0,98)
	$d'd$	39,8 (1,00)	37,5 (0,94)	42,2 (1,00)	40,1 (0,98)
	dd'	39,7 (0,99)	37,3 (0,92)	42,1 (0,99)	40,3 (0,97)

TABLE 11 – Variantes de BLEU calculées sur les documents d en position initiale (i.e. dd'), ou finale (i.e. $d'd$), ou tout seul (i.e. d) et moyennés sur 6 répétitions avec des documents d' différents.

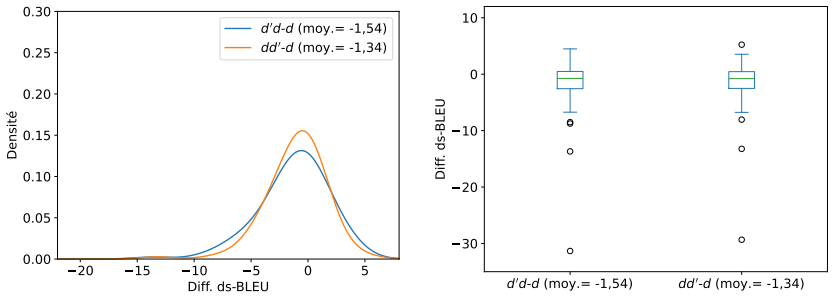


FIGURE 3 – Densité (gauche) et boîtes à moustaches (droite) de la différence entre ds-BLEU des documents traduits par les méthodes (b) ou (c) et par la méthode (a) (voir §4.3) avec TOWEINSTRUCT.

F Analyse des performances selon la longueur moyenne des documents

Le tableau 12 présente des analyses de corrélation de *la longueur moyenne* des phrases au sein du document avec le score ds-BLEU et visent à confirmer l'intérêt de l'analyse de ces corrélations. On observe des corrélations négatives (pour **THE** et **IWSLT**) ou faiblement positives (pour **TAL**), les documents ayant des phrases moyennes plus longues recevant des scores ds-BLEU en moyenne plus faibles, que l'on peut interpréter comme étant causés par la complexité plus grande des phrases à traduire.

	DEEPLPRO	MBART50(1-M)	FT SCIPAR	FT TAL-S	FT TAL-D	FT TAL-MR	TOWERBASE	TOWERINSTRUCT
THE	-0,149 (0,139)	**-0,211** (0,035)	0,021 (0,834)	0,010 (0,920)	-0,165 (0,101)	-0,177 (0,078)	-0,159 (0,115)	-0,179 (0,074)
rTAL	0,044 (0,491)	0,043 (0,501)	**0,245** (0,000)	**0,228** (0,000)	0,016 (0,800)	0,019 (0,763)	0,001 (0,989)	0,014 (0,826)
IWSLT	0,051 (0,718)	-0,047 (0,737)	0,161 (0,250)	0,095 (0,497)	-0,006 (0,969)	-0,074 (0,601)	-0,169 (0,226)	-0,042 (0,767)
THE diff	-	0,006 (0,953)	**0,217** (0,03)	**0,201** (0,045)	-0,080 (0,432)	-0,079 (0,435)	-0,002 (0,981)	-0,036 (0,724)
rTAL diff	-	0,001 (0,986)	0,091 (0,157)	0,098 (0,124)	-0,062 (0,335)	-0,087 (0,174)	-0,102 (0,109)	-0,079 (0,216)
IWSLT diff	-	-0,051 (0,719)	0,162 (0,246)	0,129 (0,356)	-0,073 (0,605)	-0,120 (0,393)	-0,254 (0,066)	-0,134 (0,337)

TABLE 12 – Corrélation de Spearman (et *p-values*) entre ds-BLEU et la longueur moyenne des phrases sources (L_m) (haut) ; corrélation de Spearman entre L_m et la différence des scores ds-BLEU entre chaque système et DEEPLPRO (bas).

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Approches cascade et de bout-en-bout pour la traduction automatique de la parole en pictogrammes

Cécile Macaire[1] Chloé Dion[1] Didier Schwab[1] Benjamin Lecouteux[1]
Emmanuelle Esperança-Rodier[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France
prenom.nom@univ-grenoble-alpes.fr

RÉSUMÉ

La traduction automatique de la parole en pictogrammes (Parole-à-Pictos) est une nouvelle tâche du Traitement Automatique des Langues (TAL) ayant pour but de proposer une séquence de pictogrammes à partir d'un énoncé oral. Cet article explore deux approches distinctes : (1) en cascade, qui combine un système de reconnaissance vocale avec un système de traduction, et (2) de bout-en-bout, qui adapte un système de traduction automatique de la parole. Nous comparons différentes architectures état de l'art entraînées sur nos propres données alignées parole-pictogrammes. Nous présentons une première évaluation automatique des systèmes et réalisons une évaluation humaine pour analyser leur comportement et leur impact sur la traduction en pictogrammes. Les résultats obtenus mettent en évidence la capacité d'une approche en cascade à générer des traductions acceptables à partir de la parole lue et dans des contextes de la vie quotidienne.

ABSTRACT

Cascade and End-to-End Approaches for Automatic Speech-to-Pictograms Translation

The automatic translation of speech into pictograms (Speech-to-Pictograms) is a new Natural Language Processing (NLP) task whose purpose is to generate a sequence of pictograms based on a speech utterance. This article explores two distinct approaches : (1) the cascade approach, which combines a speech recognition system with a machine translation system, and (2) the end-to-end approach, which adapts a speech translation system. We compare different state-of-the-art architectures trained on our own aligned speech-to-pictogram data. We present a first automatic evaluation of the systems and conduct a human evaluation to analyze their behavior and their impact on pictogram translation. The results highlight the ability of the cascade approach to generate acceptable translations from spoken language in everyday life situations.

MOTS-CLÉS : Pictogrammes, Parole, Traduction Automatique.

KEYWORDS: Pictograms, Speech, Machine Translation.

1 Introduction

La Communication Alternative et Augmentée (CAA) regroupe un ensemble d'outils et de stratégies conçus pour faciliter la communication des individus confrontés à des troubles du langage (Beukelman & Mirenda, 2017). Ces troubles impactent un ensemble de capacités langagières, allant de la production et la compréhension de la parole à l'écoute, la lecture et l'écriture. Ils peuvent avoir différentes origines, telles que certaines maladies génétiques, des troubles du spectre autistique, un déficit

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 22–35.

intellectuel, pour en citer quelques-uns. Nous retrouvons, dans la CAA, l'utilisation de pictogrammes, pour transmettre des messages dans des situations de la vie quotidienne. Un pictogramme est une représentation graphique associée à un concept (objet, personne, action, etc.) (Pereira *et al.*, 2022b). Ceux-ci présentent plusieurs avantages, notamment, qu'ils permettent de visualiser la syntaxe, de manipuler des mots et de faciliter l'accès au langage (Cataix-Nègre, 2017). D'un point de vue social, une enquête menée par la Croix-Rouge (2021) a identifié une réduction du stress, une augmentation de l'autonomie et un impact positif du bien-être général des utilisateurs de CAA.

Pourtant, selon la même étude, la CAA est confrontée à différents freins environnementaux qui limitent son utilisation et sa diffusion. L'étude cite spécifiquement le manque de sensibilisation des accompagnants au potentiel de la CAA et la difficulté d'accéder aux outils (absence d'informations, de formation, de moyens financiers et de temps).

Nous pensons que la mise en place de systèmes de traduction de la parole en une séquence de pictogrammes, tâche que nous appellerons Parole-à-Pictos (PAP), pourrait permettre de relever ces défis. Un système PAP prédit une suite de termes, chacun associé à un pictogramme unique ARASAAC [1] à partir d'un segment audio (cf. Figure 1). Notre objectif est de construire le premier système PAP pour le français.

FIGURE 1 – Illustration de la tâche PAP avec la traduction d'un segment audio en pictogrammes ARASAAC.

Pour cela, nous proposons deux approches. La première s'appuie sur un système en cascade, qui imbrique un modèle de Reconnaissance Automatique de la Parole (RAP) avec un modèle de Traduction Automatique (TA). La deuxième approche adapte une architecture de bout-en-bout de Traduction de la Parole (TP) pour la tâche Parole-à-Pictos. Nous présentons une première évaluation automatique et humaine sur nos propres jeux de données. Nous résumons nos contributions ci-dessous :

— La présentation de deux approches pour traduire automatiquement la parole en une séquence de pictogrammes.
— La construction de trois corpus de données alignées parole-texte-pictogrammes pour cette tâche [2].
— L'implémentation et la publication de plusieurs modèles de Reconnaissance de la Parole, de Traduction Automatique et de Traduction de la parole affinés sur ces jeux de données. Le code et les modèles sont disponibles en ligne [3] et les expériences sont entièrement reproductibles.
— La présentation d'une première évaluation automatique et humaine des modèles pour la tâche Parole-à-Pictos.

1. https://arasaac.org/
2. https://www.ortolang.fr/market/corpora/propicto
3. https://github.com/macairececile/speech-to-pictograms

2 État de l'art

Les précédents travaux se sont majoritairement focalisés sur la traduction du texte vers les pictogrammes, plutôt qu'à partir de la parole. Sevens *et al.* (2015) ont notamment proposé Text2Picto, un système de traduction texte-pictogrammes pour le néerlandais, ensuite étendu à l'anglais, à l'espagnol (Sevens, 2018) et au français (Norré *et al.*, 2021). De récents travaux (Pereira *et al.*, 2022a, 2023) se sont intéressés à la prédiction d'un pictogramme selon le contexte, en utilisant des modèles de type BERT (Devlin *et al.*, 2019). Le but est de proposer un modèle prédictif pour compléter une phrase en construction dans les systèmes de CAA.

La traduction automatique de la parole en une séquence de pictogrammes du Français est étudiée pour la première fois dans Vaschalde *et al.* (2018). Leur méthodologie s'appuie sur l'adaptation du système Text2Picto (Vandeghinste *et al.*, 2017) à la parole. Le modèle proposé imbrique quatre modules : un système de Reconnaissance Automatique de la Parole (RAP), un système de simplification, un modèle de désambiguïsation lexicale et un dernier module qui affiche la séquence de pictogrammes. L'évaluation est réalisée sur deux jeux de données. Le premier regroupe 15 histoires pour enfants manuellement traduites en pictogrammes. Le second est un ensemble de 20 phrases extraites du corpus ESLO (Baude & Dugua, 2017). Aucune évaluation automatique ou humaine ne sont rapportées. Récemment, les travaux de Macaire *et al.* (2022, 2023) ont exploré la traduction Parole-à-Pictos en proposant Voice2Picto. Bien que le système propose une approche novatrice, il ne compare pas différents systèmes de RAP. De plus, la traduction utilise un vocabulaire pictographique non adapté à la parole et aucune donnée spécifique à cette tâche n'est employée. Aucune évaluation, qu'elle soit automatique ou humaine, n'a été réalisée.

Pour la traduction de la parole en pictogrammes, nous nous appuyons sur les travaux précédents qui associent un système de Reconnaissance Automatique de la Parole et un système de Traduction Automatique.

Reconnaissance Automatique de la Parole `Wav2Vec2.0` (Baevski *et al.*, 2020) est un modèle basé sur l'apprentissage auto-supervisé. Celui-ci apprend des représentations robustes de la parole sur une collection importante de données non étiquetées pendant la phase dite de pré-entraînement. L'architecture est ensuite affinée sur un jeu de données étiqueté pour une tâche en aval. Plus récemment, deux modèles multimodaux et multilingues montrent des résultats compétitifs sans nécessiter une phase d'affinage. `Whisper` (Radford *et al.*, 2023) utilise l'architecture encodeur-décodeur Transformer (Vaswani *et al.*, 2017). Le modèle est appris sur 680 000 heures de données étiquetées multilingues (plus de 100 langues). `SeamlessM4T` (Barrault *et al.*, 2023) est un modèle massif de traduction automatique multimodale (traduction parole-parole, parole-texte, texte-parole, texte-texte et transcription) et multilingue sur une centaine de langues. Contrairement à `Whisper`, `SeamlessM4T` préserve les éléments de la prosodie et du style vocal dans toutes les langues couvertes.

Traduction Automatique Ott *et al.* (2018) présente un modèle neuronal Transformer (Vaswani *et al.*, 2017) séquence-à-séquence. L'architecture utilise un vocabulaire commun à chaque paire de langues. Les données sont tokenisées en sous-mots avec l'algorithme *Byte-Pair Encoding*. L'architecture est entraînée à partir de zéro. Les modèles suivants sont pré-entraînés sur des données multilingues. Liu *et al.* (2020) présente `mBART`, un modèle auto-encodeur séquence-à-séquence pré-entraîné sur une quantité importante de données monolingues dans plusieurs langues. L'architecture applique un objectif BART (Lewis *et al.*, 2020), modèle de type Transformer. L'article souligne l'avantage de `mBART` sur des langues ne figurant pas dans les données de pré-entraînement. Raffel *et al.* (2020) propose `T5`, une approche basée sur l'apprentissage par transfert. Chaque donnée textuelle

est considérée, en entrée, comme un problème texte-à-texte, permettant ainsi de réaliser différentes tâches (résumé de documents, analyse de sentiments, traduction automatique, etc.) via un modèle unique. Le modèle utilise 20TB de données textuelles de langues anglaise, française, roumaine et allemande. Costa-jussà *et al.* (2022) propose `NLLB`, un modèle de type Transformer massivement multilingue capable de traduire automatiquement dans 200 langues. Cette couverture linguistique peut être bénéfique entre deux langues apparentées via un transfert interlinguistique (Conneau *et al.*, 2020; Fan *et al.*, 2021). Plusieurs travaux présentent des approches basées sur les représentations de phrases, notamment LASER(Artetxe & Schwenk, 2019), LabSE (Feng *et al.*, 2022) et SONAR (Duquenne *et al.*, 2023). Ce dernier obtient des résultats compétitifs en TA par rapport au modèle NLLB 1B.

Traduction Automatique de la Parole La traduction automatique de parole de bout-en-bout est explorée dans plusieurs travaux. Nous pouvons citer `Fairseq S2T` (Wang *et al.*, 2020), qui combine un modèle RNN et Transformer. Ye *et al.* (2021) présentent `XSTNET`, un modèle transversal parole-texte avec `Wav2Vec2.0` comme encodeur vocal, suivi d'un entraînement progressif multitâche (modèle de TA pré-entraîné et affinage multitâche). Plus récemment, les travaux de Ye *et al.* (2022) proposent `ConST`, un modèle fondé sur une approche d'apprentissage contrastive. Celui-ci cherche à encoder les représentations audio et textuelles similaires dans un espace proche. Composé de quatre modules, `ConST` intègre un encodeur vocal utilisant les représentations `Wav2vec2.0`, une couche de plongement de mots et un encodeur-decodeur Transformer. Les scores BLEU rapportés sur MUST-C (Di Gangi *et al.*, 2019) démontrent des performances état de l'art, notamment pour des paires de langues peu dotées.

3 Méthode proposée

Nous appliquons deux approches pour la tâche Parole-à-Pictos. La première est une approche cascade constituée d'un système de RAP et d'un système de TA. Ici, la transcription fournie par le système de RAP est le point d'entrée du système de TA, dont le but est de traduire la langue source (ici le français) dans la langue cible. Pour la tâche Parole-à-Pictos, la langue cible est ce que nous nommons "langage pictographique" qui correspond à la séquence de termes (mot unique, expression polylexicale, ou phrase entière), chacun associé à un pictogramme ARASAAC. Pour les systèmes de RAP, nous comparons `Wav2Vec2.0`, `Whisper` et `SeamlessM4T`. L'objectif est de confronter les performances d'un modèle ajusté à nos données à celles de modèles massivement multilingues, multitâches et du domaine général, mais non reproductibles. Nous considérons, pour la Traduction Automatique, les modèles état de l'art présentés Section 2. Notre objectif est de comparer une architecture à entraîner à partir de zéro (*from scratch*) avec des architectures pré-entraînées sur des données multilingues et à affiner sur ces propres données. La seconde approche adapte les systèmes de bout-en-bout de Traduction Automatique de la parole à notre tâche. Dans cet article, nous testons le modèle `ConST`.

4 Données

Nous construisons deux ensembles de données issus de deux corpus de parole préexistants pour entraîner nos systèmes. Ces deux corpus se différencient par le type de parole qu'ils contiennent

(parole lue et parole spontanée). L'objectif est d'évaluer la robustesse des modèles face à diverses situations acoustiques. Un jeu de données supplémentaire est également utilisé pour l'évaluation, qui se rapproche des interactions du public cible.

Les pictogrammes utilisés proviennent d'ARASAAC, une ressource riche de plus de 25 000 pictogrammes uniques, continuellement mise à jour. Distribués sous la licence Creative Commons CC-BY-NC-SA et téléchargeables gratuitement, ces pictogrammes sont très largement utilisés dans la communauté CAA.

Propicto-orféo Nous récupérons les données alignées parole/texte issues du Corpus d'Étude pour le Français Contemporain (CEFC) (Benzitoun *et al.*, 2016), comprenant un ensemble de 12 corpus sources. Nous retrouvons des situations de parole diverses (dialogues, réunions, etc.) et dans des domaines variés. Propicto-orféo contient 290 036 segments audio pour un total de 233 h. Chaque segment audio est accompagné d'une transcription. À partir de celles-ci, nous appliquons la méthode présentée par Macaire *et al.* (2024) pour générer une traduction en pictogrammes, qui suit des règles et un lexique précis. Les données ont la forme suivante, avec *tokens* se référant à la liste des termes associés à chaque pictogramme présent dans *pictos* :

```
{
  "id": "cefc-tcof-Hen_sai_vin_reunion_08-190",
  "text": "ça fera trop rapproché",
  "pictos": [9829, 6906, 6190, 25708, 6879],
  "tokens": "prochain celle-là faire trop approcher"
}
```

Propicto-commonvoice Nous récupérons la partie française du corpus de parole lue CommonVoice version 15 (Ardila *et al.*, 2020). Cette version comprend 967 heures d'enregistrements issues de 17 911 locuteurs uniques. Sur le même principe décrit précédemment, nous appliquons la méthode de Macaire *et al.* (2024) pour générer la traduction de chaque segment audio en pictogrammes.

Propicto-eval Nous utilisons un jeu de données test pour évaluer les différentes approches sur un domaine et un type de parole restreints. Propicto-eval est un corpus de parole lue multilocuteurs (62 au total). Les données textuelles proviennent d'histoires pour enfants, de situations de la vie quotidienne et de phrases du domaine médical. Ces contextes sont particulièrement pertinents, car ils reflètent les types d'interactions de notre public cible.

5 Expériences

Données et pré-traitement Nous répartissons les données Propicto-orféo et Propicto-commonvoice en trois ensembles entraînement, validation et test selon une répartition 90/5/5. Nous supprimons la ponctuation et convertissons les transcriptions en minuscules. Chaque segment audio représente une phrase de moins de 30 secondes, la taille maximale pouvant être encodée par les systèmes de RAP. Les segments ne comprenant pas de traductions en pictogrammes ne sont pas conservés dans notre ensemble. Ces segments contiennent des disfluences ou des termes non traduits en pictogramme, conséquence de la limite de la méthode de Macaire *et al.* (2024) et des limites d'ARASAAC (certains domaines sont sous-représentés en pictogrammes). Les données sont détaillées Table 1.

	Propicto-commonvoice		Propicto-orféo	
	# phrases	# heures	# phrases	# heures
entraînement	527 554	756	231 374	147
validation	16 132	25	28 796	18
test	16 132	26	29 009	14

TABLE 1 – Répartition des données en trois ensembles (entraînement, validation, test) avec, pour chaque corpus, le nombre de phrases et le nombre d'heures.

Détails des entraînements Nous utilisons la boîte à outils SpeechBrain (Ravanelli *et al.*, 2021) et la recette fournie [4] pour affiner le modèle `Wav2Vec2.0` de RAP, avec, comme modèle pré-entrainé, *LeBenchmark/wav2vec2-FR-7K-large* (Evain *et al.*, 2021). Les segments audio de moins de 3 secondes et de plus de 10 secondes ont été écartés de l'entraînement pour éviter les segments audio trop courts ou vides.

Pour les différents systèmes de Traduction Automatique (TA), nous exploitons deux boîtes à outils : Fairseq (Ott *et al.*, 2019) et HuggingFace (Wolf *et al.*, 2020). Nous adaptons la recette proposée par Fairseq [5] du modèle de traduction *from scratch* `NMT`. Une phase de tokenisation (BPE) segmente le texte en unités de sous-mots. Un vocabulaire de 10 000 jetons est généré. Fairseq est également utilisée pour affiner le modèle `mBART`, ici, *mbart-large-cc25* appris sur 25 langues [6]. La même méthode de tokenisation décrite précédemment est appliquée. L'affinage des modèles `T5-large` et `NLLB-200` (*facebook/nllb-200-1.3B*) est réalisée en adaptant la recette proposée par HuggingFace [7]. Les principaux paramètres des modèles sont décrits Table 2 [8].

Modèle ↓	# paramètres	taux d'apprentissage	taille du lot	# epoch
Whisper large-v3 (Radford *et al.*, 2023)	1550M	-	-	-
SeamlessM4T-Large v2 (Barrault *et al.*, 2023)	2.3B	-	-	-
Wav2Vec2 (Baevski *et al.*, 2020) + CTC greedy search	318,7M	1e-4	8	30
NMT (Ott *et al.*, 2018)	51M	5e-4	8	40
mBART25 (Liu *et al.*, 2020)	610M	3e-5	8	40
T5-large (Raffel *et al.*, 2020)	220M	2e-5	32	40
NLLB-200 (Costa-jussà *et al.*, 2022)	600M	2e-5	32	40
ConST (Ye *et al.*, 2022)	150M	1e-4	8	40

TABLE 2 – Paramètres des modèles de Reconnaissance Automatique de la Parole, de Traduction Automatique et de Traduction Automatique de la Parole.

Enfin, nous suivons le pipeline basé sur Fairseq [9] pour entraîner le modèle `ConST`. Le modèle pré-entrainé *LeBenchmark/wav2vec2-FR-7K-base* est employé.

4. https://github.com/speechbrain/speechbrain/tree/develop/recipes/LibriSpeech/ASR/CTC
5. https://github.com/facebookresearch/fairseq/blob/main/examples/translation/
6. https://github.com/facebookresearch/fairseq/tree/main/examples/mbart
7. https://huggingface.co/docs/transformers/tasks/translation#train
8. Veuillez vous référer aux recettes citées pour les informations complètes sur les paramètres, ceux-ci n'ont pas été modifiés.
9. https://github.com/ReneeYe/ConST/tree/main

Résultats Nous présentons les résultats des trois approches de Reconnaissance Automatique de la Parole dans la Table 3. Notre évaluation rapporte le taux d'erreur au niveau des mots (Woodard & Nelson, 1982; Morris *et al.*, 2004). Pour les deux corpus, l'approche `Wav2Vec2.0` obtient les meilleures performances. Sur Propicto-commonvoice, la différence entre les trois modèles est minime (3 points d'écart), contrairement à Propicto-orféo avec 18,9 points séparant `Wav2vec2.0` et `Whisper`. Une hypothèse plausible pour expliquer ce phénomène pourrait résider dans le type de parole que renferment les corpus. Propicto-orféo est un corpus de parole spontanée contenant des chevauchements entre locuteurs et des disfluences (hésitations, répétitions, faux-départ). Nous supposons que `Whisper` et `SeamlessM4T`, ayant été entraînés sur une majorité de parole lue, se généralisent donc mal à des corpus plus complexes.

Modèle ↓	validation		test	
	Propicto-commonvoice	Propicto-orféo	Propicto-commonvoice	Propicto-orféo
Whisper large-v3 (Radford *et al.*, 2023)	-	-	14.34	37.69
SeamlessM4T-Large v2 (Barrault *et al.*, 2023)	-	-	12.45	46.50
Wav2Vec2 + CTC greedy search (Baevski *et al.*, 2020)	9.14	23.24	**11.21**	**27.56**

TABLE 3 – Taux d'erreur au niveau des mots (%) rapportés sur Propicto-commonvoice et Propicto-orféo entre les trois modèles de RAP.

Les résultats des modèles de traduction texte-à-pictogrammes sont présentés Table 4. Nous rapportons le score BLEU [10] par modèle et par corpus. Celui-ci est calculé en comparant la séquence de termes prédits par rapport à la séquence de termes "gold" (*tgt* : "prochain celle-là faire trop approcher", *hyp* : "celle-là faire non trop approcher"). Nous constatons des scores similaires entre les deux corpus. `mBART` présente un écart significatif avec les autres modèles (plus de 12 points en moins). De plus, le modèle `NMT`, approche entraînée à partir de zéro surpasse `mBART`. Les résultats ne démontrent pas un apport important des modèles pré-entraînés multilingues pour cette tâche de traduction. Nous n'excluons pas que la méthode de segmentation utilisée peut influencer les performances. D'autres techniques sont à explorer.

Modèle ↓	validation		test	
	Propicto-commonvoice	Propicto-orféo	Propicto-commonvoice	Propicto-orféo
Neural Machine Translation (NMT) (Ott *et al.*, 2018)	86.06	**87.28**	82.60	**87.43**
mBART25 (Liu *et al.*, 2020)	72.39	75.26	72.31	75.62
T5-large (Raffel *et al.*, 2020)	86.36	85.21	86.58	85.88
NLLB-200 (Costa-jussà *et al.*, 2022)	**87.41**	86.32	**87.66**	86.92

TABLE 4 – Scores BLEU des quatre modèles de Traduction Automatique par corpus. Les résultats sont présentés sur les données de validation et de test.

C'est lors de l'association des systèmes de Reconnaissance Automatique de la Parole et les systèmes de Traduction Automatique (notre approche cascade) que certains modèles se démarquent par rapport aux autres. La Table 5 présente les scores BLEU sur les données test en combinant chaque modèle. Pour Propicto-commonvoice, l'association de `SeamlessM4T` et `NLLB-200` obtient le score BLEU le plus élevé. Quant aux deux autres modèles de RAP avec `NLLB-200`, leurs scores sont très étroitement alignés, avec une différence de seulement 0,75. Nous expliquons cette similarité dans les

10. Les modèles sont évalués avec sacreBLEU (Post, 2018).

scores par le fait qu'il y a un écart non significatif entre les performances des systèmes de RAP. Les performances de Propicto-orféo subissent une baisse significative lorsque le système de traduction utilise les transcriptions prédites par le système de RAP en entrée. Précisément, nous observons une diminution de plus de 24 points du score BLEU, pour atteindre un score de 62,48 avec l'association de `Wav2Vec2.0` et `NLLB-200`. Le système de RAP impacte fortement les performances de la traduction en pictogrammes. De plus, bien que le modèle `NMT` soit le plus performant en TA, c'est `NLLB-200` et `T5-large` avec `Wav2Vec2.0` qui obtiennent les meilleurs scores BLEU dans notre approche cascade. Nous supposons que les modèles pré-entraînés massivement multilingues sont plus robustes lorsqu'ils sont confrontés à des termes déformés par le système de RAP.

Modèle RAP ↓	Modèle TA ↓	test	
		Propicto-commonvoice	Propicto-orféo
Whisper large-v3	NMT	73.72	58.07
	mBART25	67.02	52.05
	T5-large	78.67	57.80
	NLLB-200	**79.49**	58.82
SeamlessM4T-Large v2	NMT	73.72	52.38
	mBART25	67.61	48.71
	T5-large	79.45	53.96
	NLLB-200	**80.15**	54.86
Wav2Vec2 + CTC greedy search	NMT	73.45	61.37
	mBART25	67.02	55.49
	T5-large	78.55	**61.66**
	NLLB-200	**79.49**	**62.48**

TABLE 5 – Scores BLEU sur les données test obtenus en combinant chaque modèle de RAP avec les modèles de TA.

Nous concluons nos expériences en présentant le score BLEU Table 6 du modèle de traduction de la parole de bout-en-bout `ConST`. Les résultats indiquent des performances inférieures à celles d'une approche en cascade. Cependant, nous observons des résultats compétitifs, voire meilleurs par rapport à certaines associations de modèles RAP et TA. D'autres architectures restent à explorer dans des travaux futurs.

Modèle ↓	validation		test	
	Propicto-commonvoice	Propicto-orféo	Propicto-commonvoice	Propicto-orféo
ConST (Ye *et al.*, 2022)	73.13	62.21	71.65	60.21

TABLE 6 – Scores BLEU obtenus sur les données de validation et test par corpus du modèle de traduction de la parole de bout-en-bout `ConST`.

Le score BLEU n'offre pas d'informations précises sur les comportements spécifiques de chaque approche dans le contexte d'une traduction en pictogrammes. À cette fin, nous conduisons une évaluation humaine.

6 Évaluation humaine

Nous menons une évaluation humaine sur les deux modèles ayant obtenu le score BLEU le plus élevé pour chaque corpus. Nous adaptons un cadre analytique de conseils et de procédures pour mesurer la qualité d'une traduction, défini par Burchardt (2013), MQM [11]. Cette évaluation permet de déterminer si la traduction proposée répond aux spécifications convenues par les parties prenantes. L'évaluation analytique associe les erreurs à des mots et à des phrases spécifiques du texte (source et/ou cible). Le rôle de l'évaluateur est d'examiner le texte traduit par rapport au texte source et aux spécifications, puis d'annoter les erreurs conformément à la métrique (c'est-à-dire identifier, marquer et attribuer un type d'erreur et un niveau de gravité).

En analysant de façon globale les traductions proposées par les différents systèmes, nous sélectionnons 12 erreurs réparties en 4 catégories :

— *Précision* — ajout, omission, erreur de traduction, sur-traduction, sous-traduction,
— *Fluidité* — inintelligible, ambigu, cohésion, ordre des mots, offensif,
— *Vérité* — exhaustivité (le texte source est-il en adéquation avec le public cible ?),
— *Design* — longueur (écart important entre la longueur du texte source et celle du texte cible).

À chaque erreur est associé un niveau de gravité :

— *neutre*,
— *mineur* (aucune incidence sur la facilité d'utilisation ou la compréhensibilité du contenu),
— *majeur* (incidence sur la facilité d'utilisation sans pour autant rendre la traduction inutilisable),
— *critique* (traduction inutilisable, ce qui entraîne des dommages aux personnes, au matériel ou à la réputation d'une organisation si celles-ci ne sont pas corrigées).

Deux annotateurs experts du projet ont annoté 100 phrases sélectionnées aléatoirement parmi les données tests de Propicto-orféo et Propicto-commonvoice. Nous présentons Table 7 le Score de Qualité Globale par modèle. Celui-ci est calculé en multipliant la note de pénalité (qui résulte de l'association du nombre d'erreurs par catégorie et le niveau de gravité auquel est associé un poids - neutre : 0, mineur : 1, majeur : 5, critique : 25) par la valeur maximale (généralement 100). Le système de traduction ne peut être validé si le Score de Qualité Globale (SQG) est inférieur à la valeur de seuil. Le SQG est calculé en divisant le nombre total de pénalités (nombre d'erreurs multiplié par le poids du niveau de gravité associé à chacune) avec le nombre total de termes évalués. Par les différentes observations, les parties prenantes ont défini la limite de compréhension et d'utilisation d'une traduction à deux erreurs majeures et une erreur mineure, ce qui équivaut à une valeur de seuil de 89 ($100 - ((2 * 5) + (1 * 1))$).

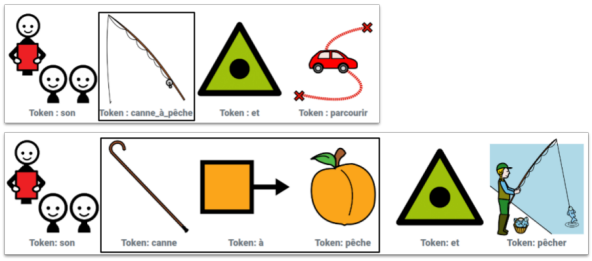
Modèle ↓		Score de Qualité Globale
Propicto-orféo	Wav2Vec2 + CTC / T5-large	45.78
	Wav2Vec2 + CTC / NLLB-200	44.56
Propicto-commonvoice	Wav2Vec2 + CTC / NLLB-200	80.65
	SeamlessM4T-Large v2 / NLLB-200	81.96
Propicto-eval	Whisper large-v3 / T5-large	87.78
	Whisper large-v3 / NLLB-200	**90.01**

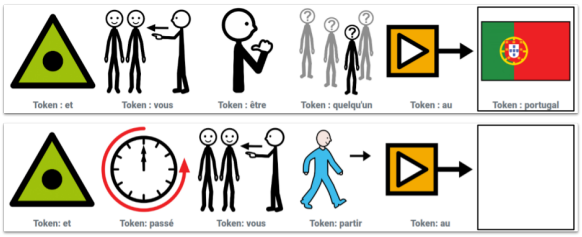
TABLE 7 – Score de qualité globale calculé sur 100 phrases annotées pour les deux meilleurs modèles par corpus.

11. https://themqm.org/

Les systèmes de traduction sur Propicto-orféo et Propicto-commonvoice n'ont pas atteint un score supérieur au seuil, rejetant ainsi leur utilisation auprès d'un public cible. Ce score s'explique par certains comportements observés par les annotateurs :

— la traduction des entités nommées est inexacte, par exemple une ville sera traduite par un pictogramme représentant une personne et inversement,
— les termes polylexicaux sont découpés en n pictogrammes au lieu d'un, par exemple "canne à pêche",
— certains homonymes sont incorrects ("avocat" pour le fruit au lieu du métier), ce problème découle des données utilisées,
— certains termes ne sont pas traduits, majoritairement dû aux erreurs générées par les systèmes de RAP,
— les chiffres sont découpés en plusieurs pictogrammes (800 traduit par 8 et 100).





Nous réalisons une dernière évaluation sur l'ensemble test Propicto-eval. En appliquant les différentes approches, c'est l'association de `Whisper` avec `T5-large` (score BLEU de 77,23) et `Whisper` avec `NLLB-200` (score BLEU de 74,97) qui obtiennent les meilleures performances. L'évaluation humaine conduite valide l'utilisabilité de notre approche avec `Whisper` et `NLLB-200` sur un corpus de parole lue et représentant des situations de la vie quotidienne, car le Score de qualité globale est supérieur à 89. Notre approche reste donc restreinte à ce cadre particulier. Elle n'est pas robuste à des situations acoustiques dites difficiles et à des domaines spécifiques (ce qu'on retrouve dans Propicto-orféo et Propicto-commonvoice). Des pistes de recherche pourraient porter sur l'amélioration de la gestion des termes non traduits et mal traduits. Nous pourrions également tester de nouvelles approches de bout-en-bout et des méthodes pour faire émerger des pictogrammes générés par des systèmes génératifs. Enfin, comparer les sorties entre modèles permettrait d'identifier leurs avantages et inconvénients respectifs.

7 Conclusion

Dans cet article, nous introduisons deux approches pour traduire automatiquement la parole en pictogrammes. Nous présentons des données spécifiquement créées pour cette tâche, couvrant diverses situations acoustiques et divers domaines. Bien que l'approche cascade présente des résultats légèrement supérieurs à l'approche de bout-en-bout sur chaque ensemble de données étudié, nous notons des résultats compétitifs avec cette dernière. Nous n'excluons donc pas cette approche pour des travaux futurs. L'évaluation humaine révèle plusieurs limitations, notamment l'impact significatif des systèmes de reconnaissance vocale sur la traduction, ainsi que la difficulté à traduire certains phénomènes linguistiques tels que les unités polylexicales et les entités nommées. Cette nouvelle tâche Parole-à-Pictos est proposée dans le cadre de la campagne d'évaluation ImageCLEF (Ionescu *et al.*, 2024), intégrée à la conférence CLEF (Conference and Labs of the Evaluation Forum) 2024.

Remerciements

Ce travail a bénéficié d'un financement de l'Agence Nationale de la Recherche, via le projet PROPICTO (ANR-20-CE93-0005). Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2023-AD011013625R1 attribuée par GENCI. Ces travaux ont nécessité l'utilisation de 1 400 heures de GPUs V100, ce qui équivaut à 33 kg de CO_2. Les données utilisées sont libres de droit.

Références

ARDILA R., BRANSON M., DAVIS K., KOHLER M., MEYER J., HENRETTY M., MORAIS R., SAUNDERS L., TYERS F. & WEBER G. (2020). Common voice : A massively-multilingual speech corpus. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4218–4222, Marseille, France : European Language Resources Association.

ARTETXE M. & SCHWENK H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, **7**, 597–610. DOI : 10.1162/tacl_a_00288.

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, **33**, 12449–12460.

BARRAULT L., CHUNG Y.-A., MEGLIOLI M. C., DALE D., DONG N., DUPPENTHALER M., DUQUENNE P.-A., ELLIS B., ELSAHAR H., HAAHEIM J. *et al.* (2023). Seamless : Multilingual expressive and streaming speech translation. *arXiv preprint arXiv :2312.05187*.

BAUDE O. & DUGUA C. (2017). Les ESLO, du portrait sonore au paysage digital. *Corpus*. HAL : halshs-01679544.

BENZITOUN C., DEBAISIEUX J.-M. & DEULOFEU H.-J. (2016). Le projet orféo : un corpus d'étude pour le français contemporain. *Corpus*, (15).

BEUKELMAN D. R. & MIRENDA P. (2017). *Communication alternative et améliorée : Aider les enfants et les adultes avec des difficultés de communication*. De Boeck Superieur.

BURCHARDT A. (2013). Multidimensional quality metrics : a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK : Aslib.

CATAIX-NÈGRE E. (2017). *Communiquer autrement : Accompagner les personnes avec des troubles de la parole ou du langage*. De Boeck Superieur.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised cross-lingual representation learning at scale. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

COSTA-JUSSÀ M. R., CROSS J., ÇELEBI O., ELBAYAD M., HEAFIELD K., HEFFERNAN K., KALBASSI E., LAM J., LICHT D., MAILLARD J. *et al.* (2022). No language left behind : Scaling human-centered machine translation. *arXiv preprint arXiv :2207.04672*.

CROIX-ROUGE (2021). Communiquons autrement : Déploiement de la communication alternative améliorée dans les établissements handicap de la croix-rouge française.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DI GANGI M. A., CATTONI R., BENTIVOGLI L., NEGRI M. & TURCHI M. (2019). MuST-C : a Multilingual Speech Translation Corpus. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 2012–2017, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1202.

DUQUENNE P.-A., SCHWENK H. & SAGOT B. (2023). Sonar : sentence-level multimodal and language-agnostic representations. *arXiv e-prints*.

EVAIN S., NGUYEN H., LE H., ZANON BOITO M., MDHAFFAR S., ALISAMIR S., TONG Z., TOMASHENKO N., DINARELLI M., PARCOLLET T., ALLAUZEN A., ESTÈVE Y., LECOUTEUX B., PORTET F., ROSSATO S., RINGEVAL F., SCHWAB D. & BESACIER L. (2021). LeBenchmark : A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. In *INTERSPEECH 2021 : Conference of the International Speech Communication Association*, Brno, Czech Republic. HAL : hal-03317730.

FAN A., BHOSALE S., SCHWENK H., MA Z., EL-KISHKY A., GOYAL S., BAINES M., CELEBI O., WENZEK G., CHAUDHARY V. *et al.* (2021). Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, **22**(107), 1–48.

FENG F., YANG Y., CER D., ARIVAZHAGAN N. & WANG W. (2022). Language-agnostic BERT sentence embedding. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 878–891, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.62.

IONESCU B., MÜLLER H., DRĂGULINESCU A. M., IDRISSI-YAGHIR A., RADZHABOV A., HERRERA A. G. S. D., ANDREI A., STAN A., STORÅS A. M., ABACHA A. B., LECOUTEUX B., STEIN B., MACAIRE C., FRIEDRICH C. M., SCHMIDT C. S., SCHWAB D., ESPERANÇA-RODIER E., IOANNIDIS G., ADAMS G., SCHÄFER H., MANGUINHAS H., COMAN I., SCHÖLER J., KIESEL J., RÜCKERT J., BLOCH L., POTTHAST M., HEINRICH M., YETISGEN M., RIEGLER M. A., SNIDER N., HALVORSEN P., BRÜNGEL R., HICKS S. A., THAMBAWITA V., KOVALEV V., PROKOPCHUK Y. & YIM W.-W. (2024). Advancing multimedia retrieval in medical, social media and content recommendation applications with imageclef 2024. In N. GOHARIAN, N. TONELLOTTO, Y. HE, A. LIPANI, G. MCDONALD, C. MACDONALD & I. OUNIS, Éds., *Advances in Information Retrieval*, p. 44–52, Cham : Springer Nature Switzerland.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). BART : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7871–7880, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.703.

LIU Y., GU J., GOYAL N., LI X., EDUNOV S., GHAZVININEJAD M., LEWIS M. & ZETTLEMOYER L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, **8**, 726–742. DOI : 10.1162/tacl_a_00343.

MACAIRE C., DION C., ARRIGO J., LEMAIRE C., ESPERANÇA-RODIER E., LECOUTEUX B. & SCHWAB D. (2024). A multimodal french corpus of aligned speech, text, and pictogram sequences for speech-to-pictogram machine translation. In *LREC*.

MACAIRE C., ESPERANÇA-RODIER E., LECOUTEUX B. & SCHWAB D. (2023). Voice2Picto : un système de traduction automatique de la parole vers des pictogrammes. In C. SERVAN & A. VILNAT, Éds., *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 5 : démonstrations*, p. 10–13, Paris, France : ATALA.

MACAIRE C., ORMAECHEA-GRIJALBA L. & PUPIER A. (2022). Une chaîne de traitements pour la simplification automatique de la parole et sa traduction automatique vers des pictogrammes (simplification and automatic translation of speech into pictograms). In Y. ESTÈVE, T. JIMÉNEZ, T. PARCOLLET & M. ZANON BOITO, Éds., *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 : 24e Rencontres Etudiants Chercheurs en Informatique pour le TAL (RECITAL)*, p. 111–123, Avignon, France : ATALA.

MORRIS A., MAIER V. & GREEN P. (2004). From wer and ril to mer and wil : improved evaluation measures for connected speech recognition.

NORRÉ M., VANDEGHINSTE V., BOUILLON P. & FRANÇOIS T. (2021). Extending a text-to-pictograph system to French and to arasaac. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, p. 1050–1059, Held Online : INCOMA Ltd.

OTT M., EDUNOV S., BAEVSKI A., FAN A., GROSS S., NG N., GRANGIER D. & AULI M. (2019). fairseq : A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019 : Demonstrations*.

OTT M., EDUNOV S., GRANGIER D. & AULI M. (2018). Scaling neural machine translation. In O. BOJAR, R. CHATTERJEE, C. FEDERMANN, M. FISHEL, Y. GRAHAM, B. HADDOW, M. HUCK, A. J. YEPES, P. KOEHN, C. MONZ, M. NEGRI, A. NÉVÉOL, M. NEVES, M. POST, L. SPECIA, M. TURCHI & K. VERSPOOR, Éds., *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 1–9, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6301.

PEREIRA J., NOGUEIRA R., ZANCHETTIN C. & FIDALGO R. (2023). Predictive authoring for brazilian portuguese augmentative and alternative communication. *arXiv preprint arXiv :2308.09497*.

PEREIRA J. A., DE MEDEIROS S., ZANCHETTIN C. & FIDALGO R. D. N. (2022a). Pictogram prediction in alternative communication boards : a mapping study. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, p. 705–717 : SBC.

PEREIRA J. A., MACÊDO D., ZANCHETTIN C., DE OLIVEIRA A. L. I. & DO NASCIMENTO FIDALGO R. (2022b). Pictobert : Transformers for next pictogram prediction. *Expert Systems with Applications*, **202**, 117231.

POST M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 186–191, Belgium, Brussels : Association for Computational Linguistics.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, p. 28492–28518 : PMLR.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, **21**(140), 1–67.

RAVANELLI M., PARCOLLET T., PLANTINGA P., ROUHE A., CORNELL S., LUGOSCH L., SUBAKAN C., DAWALATABAD N., HEBA A., ZHONG J., CHOU J.-C., YEH S.-L., FU S.-W., LIAO C.-F., RASTORGUEVA E., GRONDIN F., ARIS W., NA H., GAO Y., MORI R. D. & BENGIO Y. (2021). SpeechBrain : A general-purpose speech toolkit. arXiv :2106.04624.

SEVENS L. (2018). *Words divide, pictographs unite : Pictograph communication technologies for people with an intellectual disability*. Netherlands Graduate School of Linguistics.

SEVENS L., VANDEGHINSTE V., SCHUURMAN I. & VAN EYNDE F. (2015). Extending a Dutch text-to-pictograph converter to English and Spanish. In J. ALEXANDERSSON, E. ALTINSOY, H. CHRISTENSEN, P. LJUNGLÖF, F. PORTET & F. RUDZICZ, Éds., *Proceedings of SLPAT 2015 : 6th Workshop on Speech and Language Processing for Assistive Technologies*, p. 110–117, Dresden, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/W15-5119.

VANDEGHINSTE V., SEVENS I. S. L. & VAN EYNDE F. (2017). Translating text into pictographs. *Natural Language Engineering*, **23**(2), 217–244.

VASCHALDE C., TRIAL P., ESPERANÇA-RODIER E., SCHWAB D. & LECOUTEUX B. (2018). Automatic pictogram generation from speech to help the implementation of a mediated communication. In *Conference on Barrier-free Communication*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.

WANG C., TANG Y., MA X., WU A., OKHONKO D. & PINO J. (2020). Fairseq S2T : Fast speech-to-text modeling with fairseq. In D. WONG & D. KIELA, Éds., *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing : System Demonstrations*, p. 33–39, Suzhou, China : Association for Computational Linguistics.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., LE SCAO T., GUGGER S., DRAME M., LHOEST Q. & RUSH A. (2020). Transformers : State-of-the-art natural language processing. In Q. LIU & D. SCHLANGEN, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-demos.6.

WOODARD J. & NELSON J. (1982). An information theoretic measure of speech recognition performance.

YE R., WANG M. & LI L. (2021). End-to-End Speech Translation via Cross-Modal Progressive Training. In *Proc. Interspeech 2021*, p. 2267–2271. DOI : 10.21437/Interspeech.2021-1065.

YE R., WANG M. & LI L. (2022). Cross-modal contrastive learning for speech translation. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 5099–5113, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.376.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Au-delà de la performance des modèles : la prédiction de liens peut-elle enrichir des graphes lexico-sémantiques du français ?

Hee-Soo Choi[1, 2] Priyansh Trivedi[4]
Mathieu Constant[1] Karën Fort[2, 3] Bruno Guillaume[2]
(1) ATILF, CNRS, Université de Lorraine, 54000 Nancy, France
(2) LORIA, Université de Lorraine, 54506 Vandoeuvre-lès-Nancy, France
(3) Sorbonne Université, 75006 Paris, France, (4) Chercheur indépendant
hee-soo.choi@loria.fr, mail@priyansh.page,
mathieu.constant@atilf.fr, karen.fort@loria.fr, bruno.guillaume@loria.fr

RÉSUMÉ

Cet article présente une étude sur l'utilisation de modèles de prédiction de liens pour l'enrichissement de graphes lexico-sémantiques du français. Celle-ci porte sur deux graphes, `RezoJDM16k` et `RL-fr` et sept modèles de prédiction de liens. Nous avons étudié les prédictions du modèle le plus performant, afin d'extraire de potentiels nouveaux triplets en utilisant un score de confiance que nous avons évalué avec des annotations manuelles. Nos résultats mettent en évidence des avantages différents pour le graphe dense `RezoJDM16k` par rapport à `RL-fr`, plus clairsemé. Si l'ajout de nouveaux triplets à `RezoJDM16k` offre des avantages limités, `RL-fr` peut bénéficier substantiellement de notre approche.

ABSTRACT

Beyond Model Performance : Can Link Prediction Enrich French Lexical Graphs ?

This paper presents a resource-centric study of link prediction approaches over French lexical-semantic graphs. Our study incorporates two graphs, `RezoJDM16k` and `RL-fr` and seven link prediction models. We studied the predictions of the best-performing model to extract potential new triples based on a confidence score that we evaluated with manual annotations. Our findings highlight different benefits for the dense graph `RezoJDM16k` compared to the sparser graph `RL-fr`. While the addition of new triples to `RezoJDM16k` offers limited advantages, `RL-fr` can benefit substantially from our approach.

MOTS-CLÉS : graphes lexicaux, ressources du français, prédiction de liens, complétion de graphes.

KEYWORDS: lexical graphs, French resources, link prediction, graph completion.

1 Introduction

Les ressources lexicales ont longtemps été essentielles pour le développement des systèmes de Traitement Automatique des Langues (TAL). Malgré les récentes avancées de l'apprentissage non supervisé et des modèles de langue, ces ressources et plus spécifiquement les graphes de connaissances ou lexico-sémantiques restent des sources de connaissances linguistiques permettant l'amélioration des systèmes dans la résolution de tâches de TAL (Hwang *et al.*, 2021; d'Amato *et al.*, 2023). En outre, ces objets décrivant la langue sont également importants dans des domaines tels que la linguistique

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 36–49.

ou l'enseignement des langues. Dans cet article, nous nous intéressons à l'utilisation de modèles de prédiction de liens pour enrichir des graphes lexico-sémantiques du français. Plus précisément, nous cherchons ici à améliorer la complétude du Réseau Lexical du français, `RL-fr` (Lux-Pogodalla & Polguère, 2011), un graphe lexico-sémantique relativement peu dense créé manuellement par des lexicographes. Pour mieux comparer l'efficacité de notre approche, nous avons également appliqué nos expériences sur `RezoJDM16k` (Mirzapour *et al.*, 2022), un graphe lexico-sémantique très dense. Les principales contributions de l'article sont les suivantes : i) nous avons évalué les performances de sept modèles de prédictions de liens sur deux graphes lexico-sémantiques du français, dont un modèle présentant des résultats état-de-l'art, ii) nous avons ajouté un score de confiance sur les prédictions générées par le modèle `CompGCN-ConvE` pour extraire de potentiels nouveaux triplets et iii) nous avons analysé qualitativement ces prédictions à partir d'annotations manuelles. Notre expérience produit des résultats encourageants et ouvre une voie nouvelle vers l'enrichissement semi-automatique des ressources lexico-sémantiques.

2 Tâche de prédiction de liens

Les graphes de connaissances sont généralement incomplets en raison de l'impossibilité de décrire le monde ou la langue de manière exhaustive. La tâche de prédiction de liens consiste alors à prédire des triplets manquants dans un graphe. Il existe deux variantes principales de cette tâche : la prédiction transductive et la prédiction inductive. Dans la prédiction transductive, l'entraînement et l'inférence se font sur le même graphe. Inversement, dans la prédiction inductive, l'inférence peut avoir lieu sur un graphe différent et l'échantillon de test peut inclure des nœuds inconnus (Galkin *et al.*, 2022). Dans cet article, nous nous concentrons uniquement sur la prédiction transductive. Les graphes de connaissances peuvent être décrits comme un ensemble de triplets, désignés par *(h, r, t)* pour *head* (tête), *relation* et *tail* (queue). Étant donné des triplets incomplets tels que *(h, r, ?)* ou *(?, r, t)*, le modèle doit prédire l'entité manquante. Pour ce faire, les modèles neuronaux sont entraînés à donner un meilleur score aux triplets positifs qu'aux triplets négatifs créés *negative sampling* (Bordes *et al.*, 2013), technique consistant à corrompre les triplets positifs en remplaçant l'entité de tête ou de queue par une autre entité choisie au hasard. Les scores sont calculés à l'aide d'une fonction score, qui dépend du type de modèle. Dans cette section, nous donnons un aperçu général de certains de ces types. Nous renvoyons les lecteurs intéressés à l'étude complète de Chen *et al.* (2020) pour une compréhension approfondie des approches neuronales de prédiction de liens.

Modèles translationnels Ces modèles utilisent la distance entre les plongements de nœuds et de relations comme fonction de score. `TransE` (Bordes *et al.*, 2013) utilise la distance euclidienne, les modèles suivants comme `TransH` (Wang *et al.*, 2014), `TransR` (Lin *et al.*, 2015) et `TransD` (Ji *et al.*, 2015) offrant diverses extensions. D'autres, comme `RotatE` (Sun *et al.*, 2019), utilisent des espaces vectoriels complexes pour représenter les entités et définissent les relations comme des rotations entre elles. Cela permet de modéliser des modèles de relations plus complexes comme la symétrie/asymétrie, les inversions et les compositions.

Modèles *semantic-matching* Ces modèles utilisent une fonction de score dérivée de la similarité sémantique pour découvrir les connexions sémantiques potentielles entre les entités et les relations. Parmi les exemples notables, nous pouvons citer `RESCAL` (Nickel *et al.*, 2011), qui capture les interactions par paire entre les entités, `DistMult` (Yang *et al.*, 2015) qui réduit la charge de calcul mais est limité aux relations symétriques et `ComplEx` (Trouillon *et al.*, 2016), qui introduit des plongements basés sur un espace vectoriel complexe pour une capacité de modélisation plus large.

Architectures neuronales profondes Les approches présentées ci-dessus se limitent à l'utilisation d'opérations mathématiques simples, telles que le produit scalaire ou les multiplications de matrices sur les plongements d'entités et de relations. Par conséquent, leur capacité de modélisation ne peut être augmentée qu'en modifiant les dimensions des plongements. L'application des réseaux neuronaux profonds à ces graphes peut être une alternative pour pallier ces limitations. Non triviale, elle a été sous-explorée jusqu'aux deux développements suivants. Premièrement, Dettmers *et al.* (2018) ont proposé `ConvE` qui applique des couches de convolution sur l'espace d'intégration latent pour modéliser les interactions entité-relation et utilise une couche dense pour calculer le score. Parallèlement, les réseaux de convolution de graphes (GCN), proposés dans Kipf & Welling (2017), ont permis de propager les informations de différents nœuds à travers les chemins du graphe, ce qui a conduit à des représentations d'entités et de relations tenant compte du voisinage. Des modèles comme `R-GCN` (Schlichtkrull *et al.*, 2018) et `CompGCN` (Vashishth *et al.*, 2020) proposent d'autres modifications pour gérer les graphes multi-relationnels. Toutefois, les GCN eux-mêmes ne résolvent pas la tâche de prédiction des liens, mais fournissent des moyens plus riches d'encoder le graphe.

Métriques Inspirées de la recherche d'information (RI), les métriques traditionnelles pour la tâche de prédiction de liens sont basées sur le classement des scores des prédictions correctes parmi toutes les prédictions générées :

— **Mean Rank (MR) :** étant donné un ensemble de triplets classés par leur score de prédiction, le MR calcule le rang moyen des triplets corrects. Un MR faible signifie une meilleure performance.
— **Mean Reciprocal Rank (MRR) :** MRR est la moyenne des inverses des rangs des triplets corrects. Un MRR élevé indique une meilleure performance.
— **Hits@k :** cette mesure calcule la proportion de triplets corrects apparaissant dans le top k de la liste classée des triplets prédits. Une valeur Hits@k plus élevée indique une précision de prédiction supérieure pour les triplets les plus importants.

3 État de l'art

Si la tâche de complétion des graphes lexico-sémantiques reste sous-explorée, il existe une vaste littérature sur l'enrichissement des graphes lexicaux. De manière générale, ces approches visent à accroître la couverture de la ressource, c'est-à-dire à ajouter de nouveaux nœuds au graphe, en s'appuyant sur des ressources externes. L'une des premières approches se base sur les co-occurrences statistiques pour amorcer les graphes lexico-sémantiques existants (Biemann *et al.*, 2004). Ces approches, ainsi que d'autres antérieures (Riloff & Shepherd, 1997), nécessitaient toutefois une intervention manuelle importante en raison de la faible qualité des prédictions. L'une des avancées majeures dans ce domaine a été la création de la ressource multilingue `BabelNet` (Navigli & Ponzetto, 2012), qui a étendu `WordNet` (Miller, 1995) en croisant les sens des mots avec les articles de `Wikipedia`. Des modèles de traductions automatiques ainsi que d'autres approches (Oliver & Climent, 2012; Lam *et al.*, 2014) ont également été utilisées à des degrés divers. Les progrès de la levée d'ambiguïté lexicale et l'existence de corpus ou de dictionnaires parallèles ont également permis d'enrichir les Wordnets d'autres langues à partir de l'anglais (Taghizadeh & Faili, 2016; Arcan *et al.*, 2016). En outre, les approches visant à optimiser la complétude de ces graphes adoptent rarement une perspective centrée sur les ressources. En effet, bien qu'il existe de nombreux travaux utilisant des modèles transductifs de prédiction de liens sur ces graphes, très peu d'entre eux ont conduit à des ajouts concrets à ladite ressource, à l'exception des travaux de Fellbaum (1998) sur l'ajout de liens entre les nœuds de `WordNet` par extraction de motifs (*pattern mining*) à partir de corpus.

Il est également important de souligner que ce domaine a un biais important en faveur des ressources de l'anglais, comme le montre l'utilisation de jeux de données de référence tels que WN18RR (Dettmers *et al.*, 2018) et FB15K-237 (Toutanova & Chen, 2015). Si la plupart des techniques sont en théorie transférables à n'importe quel graphe, très peu d'efforts empiriques ont été faits. Pour les ressources du français, Mirzapour *et al.* (2022) ont étudié l'efficacité des modèles de prédiction de liens sur un graphe lexico-sémantique, mais se concentrent davantage sur la création du jeu de données et l'évaluation des modèles plutôt que sur l'enrichissement de la ressource.

4 Graphes lexico-sémantiques du français

RezoJDM et RezoJDM16k RezoJDM (Lafourcade & Joubert, 2008; Lafourcade & Le Brun, 2020) est un réseau lexico-sémantique du français développé *via* des jeux ayant un but, des approches contributives et des mécanismes d'inférence. La plateforme JeuxDeMots [1] propose des jeux permettant de développer le réseau en ajoutant de nouvelles entrées et de vérifier des informations du réseau. Le jeu principal invite les joueurs à saisir des termes dans un délai imparti selon un terme et un type de relation donnés. L'utilisation des jeux ayant un but a ainsi permis de créer un graphe dirigé très dense, puisqu'il comprend actuellement plus de 537 millions de relations et six millions de nœuds [2]. Dans RezoJDM, les nœuds représentent principalement des termes (type n_term) mais contiennent aussi des informations telles que des étiquettes de partie du discours (type n_pos) ou des formes fléchies (n_form). Les relations sont divisées en trois catégories : lexicales (synonymie, antonymie...), ontologiques (hyperonymie, méronymie...) et prédicatives (agent, conséquences...). Les nœuds et les relations ont la particularité de présenter des poids en fonction de la dynamique du jeu. Pour les relations, un poids positif code une relation vraie et un poids négatif code une relation fausse, attribut relativement rare dans les graphes de connaissances. La polysémie d'un terme est exprimée en distinguant un nœud générique de ses nœuds de raffinement. Cependant, comme le jeu demande au joueur d'entrer autant de termes que possible dans un temps limité, les joueurs ont tendance à ne pas affiner leurs réponses, ce qui conduit à une faible densité des nœuds de raffinement. Par exemple, le nœud générique accord a un degré de 10 549, tandis que les degrés des nœuds de raffinement accord>pacte et accord>acceptation sont respectivement de 194 et 123.

Mirzapour *et al.* (2022) ont créé le sous-graphe RezoJDM16k en appliquant divers filtres aux nœuds et relations de RezoJDM. Seuls les nœuds de type n_term et de poids supérieur à 50 ont été retenus. Le même filtre de poids a été appliqué aux relations et certains types ont également été supprimés. En outre, les types de relations apparaissant moins de 100 fois et les nœuds ayant un degré inférieur à 45 ont été exclus, afin d'améliorer l'efficacité des modèles. Le graphe final est composé de 15 746 nœuds et de 832 093 relations.

RL-fr Le Réseau Lexical du Français RL-fr est un réseau lexico-sémantique du français créé par des lexicographes, où les nœuds correspondent à des unités lexicales et les arêtes à des relations lexico-sémantiques ou combinatoires (Lux-Pogodalla & Polguère, 2011). Dans cette section, nous présentons des informations concernant la version 2.1 de RL-fr, utilisée pour nos expériences [3].

Les unités lexicales dans RL-fr sont les entités fondamentales pour la description lexicographique et peuvent être soit un lexème, soit un idiome. Les lexèmes sont des unités lexicales monolexémiques

1. https://www.jeuxdemots.org
2. En octobre 2023.
3. https://www.ortolang.fr/market/lexicons/lexical-system-fr/v2.1

et correspondent aux sens des mots. Par conséquent, un mot polysémique, appelé vocable, est représenté comme une collection d'unités lexicales interconnectées par une relation de copolysémie. La version 2.1 du RL-fr contient 29 220 unités lexicales et 18 625 vocables. Voici les lexèmes du vocable jambe, à titre d'exemple :

— Jambe I.1 : Marc attend patiemment, les **jambes** croisées.
— Jambe I.2a : Le cheval s'est blessé à la **jambe**.
— Jambe I.2b : Il y a de la **jambe** de porc au menu.
— Jambe II : La **jambe** droite du pantalon est déchirée.
— Jambe III : Une des **jambes** de suspension doit être changée.

Contrairement aux dictionnaires classiques, où les sens d'un mot sont généralement seulement listés, le RL-fr représente les types de relation entre les sens avec la notion de copolysémie, décrite comme la relation entre les différents sens d'un mot, opposée à la polysémie qui est la propriété des mots d'exprimer plusieurs sens (Polguère, 2018). Il existe donc plusieurs relations de copolysémie, telle que la métonymie ou la métaphore (cf. Annexe A).

La ressource est basée sur la théorie Sens-Texte qui permet d'encoder les relations paradigmatiques et syntagmatiques avec les fonctions lexicales (Mel'čuk, 1996). Quelques exemples de fonctions lexicales paradigmatiques sont donnés ci-dessous :

— Synonymie (Syn) : vélo → bicyclette
— Antonymie (Anti) : accord → désaccord
— Hyperonymie (Gener) : amour → sentiment

Les relations syntagmatiques comprennent, entre autres, les collocations et les verbes supports :
— Intensifier (Magn) : boire → comme un trou
— Verbe support (Oper) : danger → courir

Les fonctions lexicales peuvent être utilisées pour représenter des relations sémantiques simples ou complexes. Dans la version 2.1 du RL-fr, il existe 686 fonctions lexicales différentes. Pour nos expériences, nous avons utilisé les familles de relations pour réduire le nombre de types de relations, ce qui correspond à 95 types de fonctions lexicales et 11 types de relations de copolysémie. Au total, il existe 62 641 relations encodées par des fonctions lexicales et 9 413 relations de copolysémie.

Deux graphes différents mais complémentaires Malgré un format de réseau commun, RezoJDM16k et RL-fr présentent des différences significatives en termes de création, de portée et de représentation de la polysémie. Le tableau 1 illustre le contraste topologique entre ces deux graphes : RezoJDM16k est très dense, avec un degré moyen de nœuds de 105,7 et un nombre d'arêtes 10 fois supérieur à celui de RL-fr. Par ailleurs, RL-fr présente davantage de nœuds, en raison de la représentation des sens des mots dans des nœuds distincts. La figure 1 montre un aperçu du réseau RL-fr autour du vocable accord, représenté par deux nœuds accord 1 et accord 2[4].

La méthode de création des ressources a un impact sur la structure du graphe, l'enrichissement manuel étant un processus chronophage. Cependant, elle garantit un contenu contrôlé et de qualité qui est vérifié par des experts du domaine. Concernant RezoJDM16k, la qualité de la ressource n'est pas nécessairement inférieure, car elle est fournie par des bénévoles motivés par leur intérêt pour la langue (Lafourcade & Le Brun, 2020), mais elle présente un certain bruit malgré les vérifications semi-automatiques.

4. https://spiderlex.atilf.fr/fr/q/*accord***

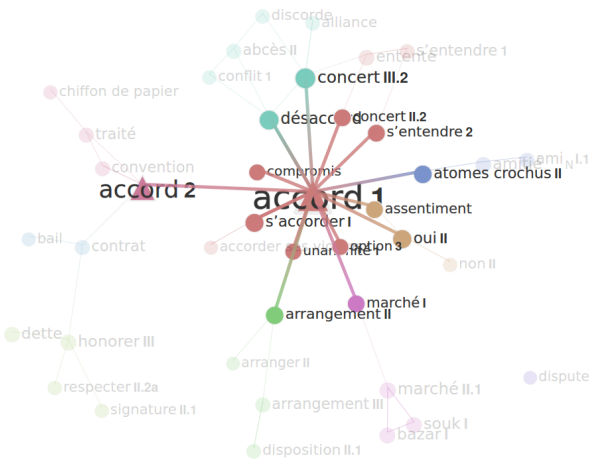


FIGURE 1 – Vocable `accord` représenté dans le `RL-fr`.

	`RezoJDM16k`	`RL-fr`
# nœuds	15 746	29 220
# arêtes	832 093	72 054
# types d'arête	150	106
Degré moyen d'un nœud	105,7	5,2
Degré minimum	2	1
Degré maximum	10 403	187

TABLE 1 – Statistiques sur `RezoJDM16k` et `RL-fr`.

5 Expériences

Pré-traitement des jeux de données La prédiction transductive des liens implique que tous les nœuds des échantillons de validation et de test soient présents dans l'échantillon d'entraînement. Nous avons décidé de diviser aléatoirement les triplets entre les échantillons d'entraînement, de validation et de test (80 %, 10 %, 10 %) et de vérifier si tous les nœuds de la validation et du test sont présents dans l'entraînement. Si ce n'est pas le cas, nous supprimons le triplet de l'ensemble. Pour un graphe très dense comme `RezoJDM16k`, aucun triplet n'a été perdu même après dix divisions aléatoires différentes. En revanche, un graphe peu dense tel que `RL-fr` implique des pertes d'arêtes et de nœuds (- 2 152 nœuds, - 1 037 arêtes). Nous obtenons un graphe final de 27 068 nœuds et 71 017 arêtes, que nous appelons `RLF27k`. Le tableau 2 présente `RezoJDM16k` et `RLF27k` et leurs échantillons d'entraînement, de validation et de test respectifs.

	`RezoJDM16k`	`RLF27k`
# nœuds	15 746	27 068
# arêtes	832 093	71 017
# triplets entraînement	665 674	57 643
# triplets validation	83 209	6 674
# triplets test	83 210	6 700

TABLE 2 – Statistiques sur les échantillons d'entraînement, validation et test de `RezoJDM16k` et `RLF27k` après une division en 80 %, 10 %, 10 %.

Performances des modèles Nous avons évalué les performances de modèles de prédiction de liens sur les deux ressources en utilisant les six mêmes modèles que ceux décrits dans Mirzapour *et al.* (2022). Nous avons reproduit leurs expériences sur `RezoJDM16k` et nous les avons menées sur `RLF27k`. De plus, nous avons utilisé un modèle `ConvE` avec l'encodeur `CompGCN` (Vashishth *et al.*, 2020) (appelé `CompGCN-ConvE`), afin d'explorer l'efficacité d'un modèle GNN (*Graph Neural Networks*) sur des données en français. Les résultats pour `RezoJDM16k` sont présentés dans le tableau 3. Nous constatons que `CompGCN-ConvE` surpasse tous les autres modèles dans presque

Modèle	MRR ↑	MR ↓	Hits@10 ↑	Hits@3 ↑	Hits@1 ↑
TransE (Bordes *et al.*, 2013)	0,180	200,78	0,437	0,242	0,040
TransH (Wang *et al.*, 2014)	0,217	173,28	0,503	0,293	0,064
TransD (Ji *et al.*, 2015)	0,216	**168,18**	0,500	0,290	0,065
DistMult (Yang *et al.*, 2015)	0,219	194,16	0,446	0,252	0,109
ComplEx (Trouillon *et al.*, 2016)	0,256	190,79	0,539	0,309	0,119
RotatE (Sun *et al.*, 2019)	0,312	177,04	0,587	0,409	0,155
CompGCN-ConvE (Vashishth *et al.*, 2020)	**0,461**	171,26	**0,659**	**0,514**	**0,357**

TABLE 3 – Résultats des modèles de prédiction de liens sur RezoJDM16k.

Modèle	MRR ↑	MR ↓	Hits@10 ↑	Hits@3 ↑	Hits@1 ↑
TransE (Bordes *et al.*, 2013)	0,278	**2594,24**	0,624	0,497	0,033
TransH (Wang *et al.*, 2014)	0,250	2957,59	0,581	0,465	0,011
TransD (Ji *et al.*, 2015)	0,255	2752,03	0,587	0,472	0,016
DistMult (Yang *et al.*, 2015)	0,373	2748,25	0,613	0,502	0,216
ComplEx (Trouillon *et al.*, 2016)	0,413	3447,98	0,593	0,524	0,284
RotatE (Sun *et al.*, 2019)	0,399	3650,92	0,490	0,454	0,336
CompGCN-ConvE (Vashishth *et al.*, 2020)	**0,515**	2808,68	**0,627**	**0,559**	**0,450**

TABLE 4 – Résultats des modèles de prédiction de liens sur RLF27k.

toutes les métriques. Il atteint notamment un Hits@1 de 0,357, qui est plus de deux fois supérieur à celui de RotatE. Les scores Hits@3, Hits@10 et MRR sont également les plus élevés et le MR reste proche du meilleur score obtenu par TransD. Le tableau 4 présente les résultats pour RLF27k. Une fois de plus, CompGCN-ConvE se distingue avec un MRR de 0,515 et les meilleurs scores Hits@k. Cependant, il convient de noter que le MR élevé indique une disparité dans le classement des triplets corrects : 60 % d'entre eux se classent parmi les dix premiers, tandis que les 40 % restants sont nettement moins bien classés.

Score de confiance Nous avons effectué une analyse approfondie des prédictions du meilleur modèle, CompGCN-ConvE. Au-delà d'évaluer la performance du modèle dans la prédiction des triplets de test, nous avons observé toutes les prédictions du modèle pour une entité et une relation données en classant les prédictions selon leur score. La figure 2 montre les 20 meilleures prédictions pour la tête bonnet I et la relation synonymie de RLF27k. Le modèle prédit logiquement un triplet présent dans l'entraînement avec un score élevé (0,893), puis prédit deux triplets de test avec des scores d'environ 0,08. Nous supposons que les triplets restants, qui n'existent pas dans le graphe original, pourraient être des nouveaux potentiels triplets. Néanmoins, la fonction de score ne permet pas d'évaluer significativement la pertinence de ces triplets, dans la mesure où tous les scores sont très faibles (environ 0,01).

Nous avons donc décidé d'utiliser un meilleur algorithme d'inférence, à savoir la technique de Monte-Carlo (MC) Dropout (Gal & Ghahramani, 2016). Le *dropout* fonctionne en désactivant de manière aléatoire des neurones d'un réseau de neurones. Généralement, lors d'une inférence standard, nous désactivons le *dropout* pour obtenir des prédictions déterministes et moins bruitées. Pour l'inférence basée sur MC Dropout, nous générons plusieurs prédictions pour la même entrée, en échantillonnant un masque *dropout* différent à chaque fois. On obtient ainsi une distribution prédictive pour un modèle et des entrées donnés, qui produit un ensemble plus riche d'informations sur les prédictions du modèle, comme la possibilité de calculer des scores de confiance pour n'importe quelle prédiction (cf.

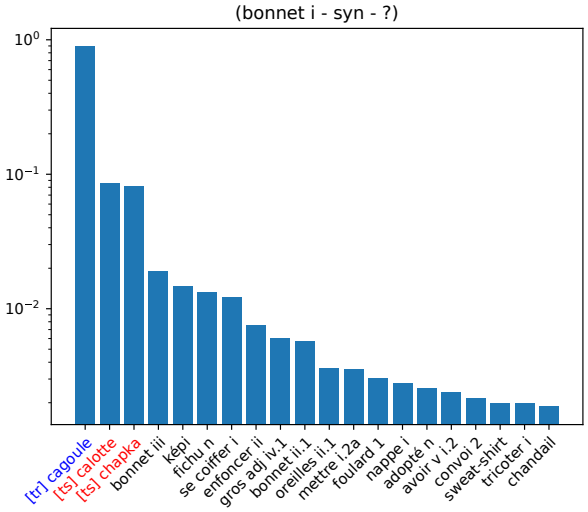


FIGURE 2 – Prédictions pour la tête `bonnet I` et la relation `synonymie`. Les triplets qui existent déjà dans `RLF27k` sont en rouge ([tr] pour *training* (entraînement) et [ts] pour test). L'axe des ordonnées est mis à l'échelle logarithme pour une meilleure lisibilité.

Annexe B). Considérons par exemple le triplet incomplet `(bonnet I, syn, ?)`. Étant donnée la distribution prédictive originale du modèle, nous appliquons le MC Droput pour générer $n = 100$ nouvelles distributions pour le triplet. Ensuite, nous définissons un critère d'inclusion pour transformer les prédictions à valeur réelle en décisions binaires, ici nous vérifions si l'entité prédite figure parmi les 10 premières prédictions notées. Nous calculons enfin le score de confiance comme le rapport entre le nombre de fois où l'entité prédite apparaît dans le top-10 et le nombre de distributions prédictives n. Cela nous permet d'établir des scores de confiance pour les prédictions du modèle et de faire des affirmations telles que « Selon ce modèle, l'entité queue `kepi` est dans le top-10 des prédictions pour l'entité tête `bonnet I` et la relation `synonymie` avec 75 % de confiance ».

6 Analyse qualitative

Nous cherchons à déterminer si le score de confiance permet d'identifier des triplets pertinents susceptibles d'être ajoutés aux graphes. Pour cela, nous avons généré toutes les combinaisons possibles de nœuds et de relations pour chaque jeu de données et nous avons supprimé les triplets déjà présents dans les graphes. Au total, 533 551 triplets ont été générés pour `RLF27k` et 1 720 454 pour `RezoJDM16k`. Nous nous sommes concentrés sur les triplets dont les entités ne sont pas reliées par un chemin orienté. Pour `RLF27k`, nous avons obtenu 95 766 triplets finaux. Pour `RezoJDM16k`, en raison de la forte densité du graphe, le plus court chemin entre deux entités est au maximum de longueur 4. Pour appliquer une méthodologie similaire à celle utilisée pour `RLF27k`, nous avons conservé les triplets avec des chemins de longueurs 3 et 4, ce qui donne un total de 154 168 triplets.

Pour évaluer le score de confiance, 240 triplets par jeu de données ont été annotés par quatre annotateurs, chacun ayant annoté 120 triplets ce qui permet d'obtenir deux annotations pour chaque triplet et calculer un accord inter-annotateurs (AIA). Les scores de confiance sont représentés de manière homogène dans les échantillons de chaque annotateur, regroupés par intervalle de 0,1. La tâche d'annotation consiste à déterminer si un lien sémantique ou syntaxique existe entre deux entités.

Les annotateurs ne disposent d'aucune information en dehors du triplet et trois étiquettes d'annotation sont possibles : (1) il existe un lien entre les entités, (-1) il n'y a pas de lien, (0) le lien est ambigu ou discutable. L'AIA est calculé avec un kappa de Cohen et s'avère bien plus élevé pour le RLF27k avec l'accord le plus fort à 0,84 contre 0,61 pour RezoJDM16k (cf. Annexe C). L'accord minimum pour RezoJDM16k est particulièrement faible, à 0,1, contrairement à 0,49 pour le RLF27k. La différence de l'AIA entre RezoJDM16k et RLF27k peut être attribuée à la distribution inégale des étiquettes d'annotation, et notamment à la surreprésentation de l'étiquette -1 dans RezoJDM16k. Sur les 240 triplets et sur les 183 où les annotateurs sont en accord, 85 % (156 triplets) ont été annotés comme -1 et 15 % (27) comme 1. Cette proportion s'explique par la forte densité du graphe qui implique que des nœuds sémantiquement distants sont connectés par un chemin relativement court de 4.

La figure 3 présente la comparaison entre les annotations manuelles et les scores de confiance des triplets des échantillons de RezoJDM16k et RLF27k. Pour RezoJDM16k, les rares triplets annotés comme exacts ont tendance à présenter des scores de confiance élevés. Toutefois, en raison de la prévalence des annotations -1, il est difficile d'établir une corrélation solide. D'autre part, dans l'échantillon RLF27k, sur les 200 triplets où les deux annotateurs étaient d'accord, 56,5 % (113) ont été annotés comme -1, 39 % (78) comme 1 et 4,5 % (9) comme 0. Nous pouvons noter que les étiquettes d'annotation dans RLF27k présentent une distribution plus équilibrée par rapport à RezoJDM16k et qu'on observe une plus forte corrélation entre les étiquettes d'annotation et les scores de confiance sur les triplets de RLF27k. Les triplets annotés comme -1 ont un score de confiance faible, tandis que ceux annotés comme 1 ont un score de confiance plus élevé.

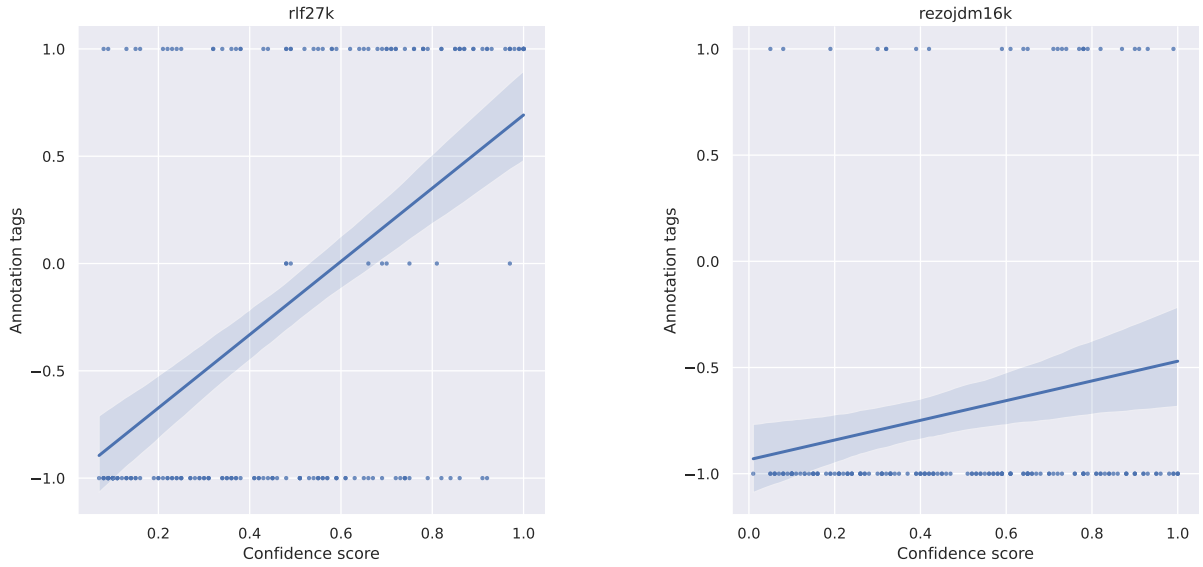


FIGURE 3 – Corrélations entre les valeurs d'annotation et les scores de confiance dans RLF27k et de RezoJDM16k. Les triplets considérés sont ceux pour lesquels les annotateurs sont d'accord.

La figure 5 présente le ratio des triplets corrects, i.e. annotées par les annotateurs comme ayant un lien entre ses entités, en fonction du seuil de confiance. Nous remarquons qu'un seuil de confiance élevé conduit à une plus grande proportion de triplets corrects. Par exemple, poser un seuil de confiance de 0,95 permet de faire en sorte que tous les triplets prédits soient corrects. Sur le graphe RLF27k complet, il existe 95 766 triplets dont les entités ne sont pas connectées par un chemin orienté et 398 d'entre eux ont un score de confiance supérieur à 0,95, qui peuvent donc être considérés comme de nouveaux triplets potentiels.

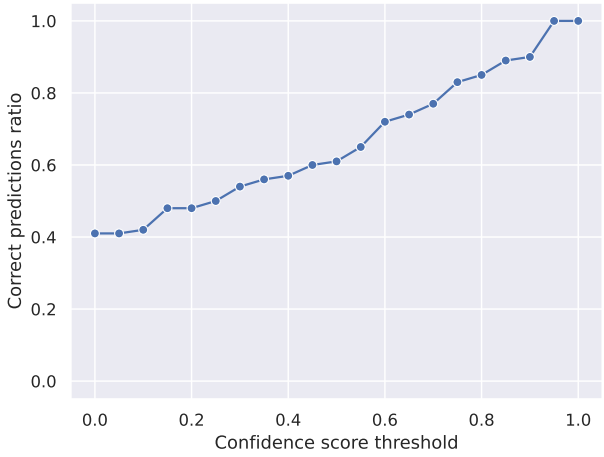


FIGURE 4 – Ratio des prédictions correctes (annotées 1) du RLF27k selon le seuil de confiance.

7 Discussion

Dans cette étude, nous avons cherché à identifier de nouveaux triplets potentiels pour deux graphes lexico-sémantiques, RezoJDM16k et RLF27k, en utilisant des modèles de prédiction de liens. Sur les sept modèles, le modèle CompGCN-ConvE a obtenu les meilleures performances sur les deux graphes, surpassant les modèles état-de-l'art pour RezoJDM16k. Nous avons également étudié l'utilisation de l'algorithme MC Dropout pour générer des prédictions basées sur un score de confiance que nous avons évalué avec des annotations manuelles.

L'analyse qualitative que nous avons menée montre que les triplets de RLF27k ayant un score de confiance élevé sont de potentiels candidats à l'intégration dans la ressource, sous réserve d'une validation par des experts. Pour RezoJDM16k, du fait de sa forte densité, l'ajout de nouveaux triplets n'apporte pas d'avantages significatifs dans la mesure où même les entités ayant une faible proximité sémantique sont connectées par des chemins courts. Cependant, notre approche peut s'avérer utile pour identifier des erreurs ou affiner des relations génériques. En effet, RezoJDM16k présente une forte proportion de la relation générique associated, constituant 31 % des arêtes, qui pourrait être affinée grâce aux prédictions du modèle.

Bien que les résultats de notre approche soient prometteurs, nous maintenons que la vérification manuelle est une étape importante car la représentation de la polysémie dans des nœuds distincts affecte directement les prédictions. Étant donné que le modèle ne s'appuie que sur la structure du graphe et les nœuds voisins pour appréhender la sémantique, il existe des limites inhérentes à la prédiction de l'entité précise parmi les différents sens d'un vocable. Dans des recherches ultérieures, au-delà des méthodologies d'évaluation intrinsèque, nous avons pour perspective d'évaluer de manière extrinsèque en utilisant des graphes lexico-sémantiques augmentés ou corrigés dans des tâches de TAL, telle que la levée d'ambiguïté lexicale. En outre, d'un point de vue ressource, nous souhaitons explorer les avantages mutuels et l'amélioration entre les deux réseaux lexico-sémantiques français. Le code associé à nos expériences et les graphiques qui en résultent sont disponibles librement sur le lien suivant : https://github.com/hschoi4/fr-link-prediction.

Remerciements

Nous remercions les relecteurs pour leurs commentaires avisés qui ont permis l'amélioration de cet article. Nous remercions également Vincent Tourneur (LORIA, INRIA) pour son aide dans les manipulations techniques. Les expériences présentées dans cet article ont été réalisées sur le banc de tests Grid'5000, soutenu par un groupe inter-scientifique hébergé par l'Inria et comprenant le CNRS, le RENATER et plusieurs universités ainsi que d'autres organisations (https://www.grid5000.fr).

Références

ARCAN M., MCCRAE J. P. & BUITELAAR P. (2016). Expanding wordnets to new languages with multilingual sense disambiguation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 97–108, Osaka, Japon : The COLING 2016 Organizing Committee.

BIEMANN C., SHIN S.-I. & CHOI K.-S. (2004). Semiautomatic extension of CoreNet using a bootstrapping mechanism on corpus-based co-occurrences. In *COLING 2004 : Proceedings of the 20th International Conference on Computational Linguistics*, p. 1227–1232, Genève, Suisse : COLING.

BORDES A., USUNIER N., GARCIA-DURÁN A., WESTON J. & YAKHNENKO O. (2013). Translating embeddings for modeling multi-relational data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, p. 2787–2795, Red Hook, États-Unis : Curran Associates Inc.

CHEN Z., WANG Y., ZHAO B., CHENG J., ZHAO X. & DUAN Z. (2020). Knowledge graph completion : A review. *IEEE Access*, **8**, 192435–192456. DOI : 10.1109/ACCESS.2020.3030076.

D'AMATO C., MAHON L., MONNIN P. & STAMOU G. (2023). Machine Learning and Knowledge Graphs : Existing Gaps and Future Research Challenges. *Transactions on Graph Data and Knowledge*, **1**(1), 1–35. DOI : 10.4230/TGDK.1.1.8, HAL : hal-04353543.

DETTMERS T., MINERVINI P., STENETORP P. & RIEDEL S. (2018). Convolutional 2d knowledge graph embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18 : AAAI Press.

FELLBAUM C. (1998). *WordNet : An Electronic Lexical Database*. The MIT Press. DOI : 10.7551/mitpress/7287.001.0001.

GAL Y. & GHAHRAMANI Z. (2016). Dropout as a bayesian approximation : Representing model uncertainty in deep learning. In M. F. BALCAN & K. Q. WEINBERGER, Éds., *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 de *Proceedings of Machine Learning Research*, p. 1050–1059, New York, États-Unis : PMLR.

GALKIN M., BERRENDORF M. & HOYT C. T. (2022). An open challenge for inductive link prediction on knowledge graphs.

HWANG J. D., BHAGAVATULA C., LE BRAS R., DA J., SAKAGUCHI K., BOSSELUT A. & CHOI Y. (2021). Comet-atomic 2020 : On symbolic and neural commonsense knowledge graphs. In *AAAI*.

JI G., HE S., XU L., LIU K. & ZHAO J. (2015). Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 687–696, Pékin, Chine : Association for Computational Linguistics. DOI : 10.3115/v1/P15-1067.

KIPF T. N. & WELLING M. (2017). Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings*, Toulon, France : OpenReview.net.

LAFOURCADE M. & JOUBERT A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. In *JADT'08 : Journées internationales d'Analyse statistiques des Données Textuelles*, p. 657–666, France. HAL : lirmm-00358848.

LAFOURCADE M. & LE BRUN N. (2020). Jeuxdemots : Un réseau lexico-sémantique pour le français, issu de jeux et d'inférences. *Revue Lexique*, **27**, 47–86.

LAM K. N., AL TAROUTI F. & KALITA J. (2014). Automatically constructing Wordnet synsets. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 106–111, Baltimore, Maryland : Association for Computational Linguistics. DOI : 10.3115/v1/P14-2018.

LIN Y., LIU Z., SUN M., LIU Y. & ZHU X. (2015). Learning entity and relation embeddings for knowledge graph completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, **29**(1). DOI : 10.1609/aaai.v29i1.9491.

LUX-POGODALLA V. & POLGUÈRE A. (2011). Construction of a French Lexical Network : Methodological Issues. In *First InternationalWorkshop on Lexical Resources, WoLeR 2011*, p. 54–61, Ljubljana, Slovénie. HAL : hal-00686467.

MEL'ČUK I. (1996). Lexical functions in lexicography and natural language processing. *Lexical Functions : A Tool for the Description of Lexical Relations in the Lexicon*, p. 37–102.

MILLER G. A. (1995). WordNet : A lexical database for English. *Communications of the ACM*, **38**(11), 39–41.

MIRZAPOUR M., RAGHEB W., SAEEDIZADE M. J., COUSOT K., JACQUENET H., CARBON L. & LAFOURCADE M. (2022). Introducing RezoJDM16k : a French KnowledgeGraph DataSet for link prediction. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 5163–5169, Marseille, France : European Language Resources Association.

NAVIGLI R. & PONZETTO S. P. (2012). Babelnet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**, 217–250. DOI : https ://doi.org/10.1016/j.artint.2012.07.001.

NICKEL M., TRESP V. & KRIEGEL H.-P. (2011). A three-way model for collective learning on multi-relational data. In L. GETOOR & T. SCHEFFER, Éds., *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, New York, États-Unis : ACM.

OLIVER A. & CLIMENT S. (2012). Parallel corpora for wordnet construction : Machine translation vs. automatic sense tagging. In A. GELBUKH, Éd., *Computational Linguistics and Intelligent Text Processing*, p. 110–121, Berlin, Heidelberg : Springer Berlin Heidelberg.

POLGUÈRE A. (2018). A Lexicographic Approach to the Study of Copolysemy Relations *. *Russian Journal of Linguistics = Vestnik Rossijskogo universiteta družby narodov. Seriâ Lingvistika*, **22**(4), 788 – 820. DOI : 10.22363/2312-9182-2018-22-4-788-820, HAL : halshs-02089585.

RILOFF E. & SHEPHERD J. (1997). A corpus-based approach for building semantic lexicons. In *Second Conference on Empirical Methods in Natural Language Processing*.

SCHLICHTKRULL M., KIPF T. N., BLOEM P., VAN DEN BERG R., TITOV I. & WELLING M. (2018). Modeling relational data with graph convolutional networks. In A. GANGEMI, R. NAVIGLI, M.-E. VIDAL, P. HITZLER, R. TRONCY, L. HOLLINK, A. TORDAI & M. ALAM, Éds., *The Semantic Web*, p. 593–607, Cham : Springer International Publishing.

SUN Z., DENG Z., NIE J. & TANG J. (2019). Rotate : Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019*, La Nouvelle-Orléans, États-Unis : OpenReview.net.

TAGHIZADEH N. & FAILI H. (2016). Automatic wordnet development for low-resource languages using cross-lingual wsd. *J. Artif. Int. Res.*, **56**(1), 61–87.

TOUTANOVA K. & CHEN D. (2015). Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, p. 57–66, Pékin, Chine : Association for Computational Linguistics. DOI : 10.18653/v1/W15-4007.

TROUILLON T., WELBL J., RIEDEL S., GAUSSIER E. & BOUCHARD G. (2016). Complex embeddings for simple link prediction. In M. F. BALCAN & K. Q. WEINBERGER, Éds., *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 de *Proceedings of Machine Learning Research*, p. 2071–2080, New York, États-Unis : PMLR.

VASHISHTH S., SANYAL S., NITIN V. & TALUKDAR P. P. (2020). Composition-based multi-relational graph convolutional networks. In *8th International Conference on Learning Representations, ICLR 2020*, Addis-Abeba, Éthiopie : OpenReview.net.

WANG Z., ZHANG J., FENG J. & CHEN Z. (2014). Knowledge graph embedding by translating on hyperplanes. *Proceedings of the AAAI Conference on Artificial Intelligence*, **28**. DOI : 10.1609/aaai.v28i1.8870.

YANG B., YIH W., HE X., GAO J. & DENG L. (2015). Embedding entities and relations for learning and inference in knowledge bases. In Y. BENGIO & Y. LECUN, Éds., *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, San Diego, États-Unis.

Annexes

A Exemple de relations de copolysémie

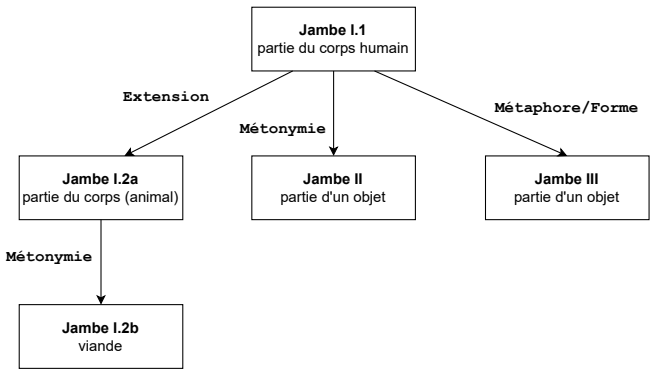


FIGURE 5 – Relation de copolysémie du vocable `jambe`.

B Monte Carlo Dropout

Notre objectif est d'obtenir une approximation de la distribution prédictive pour un point donné à partir de notre modèle paramétré. La distribution prédictive $p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y})$ pour un nouveau point de données x^* à partir d'un jeu de données $(\mathbf{X}, \mathbf{Y})$ et un modèle paramétré par θ est :

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) = \int p(\mathbf{y}^*|\mathbf{x}^*, \theta)p(\theta|\mathbf{X}, \mathbf{Y})\, d\theta \tag{1}$$

Ici, le premier terme est la probabilité d'observer la sortie y^* compte tenu de l'entrée x^* et des paramètres du modèle θ, et le second terme est la distribution *a posteriori* des paramètres compte tenu des données. Le calcul de cette dernière est difficile à réaliser pour les réseaux neuronaux profonds en raison de leur espace paramétrique de grande dimension. Monte-Carlo (MC) Dropout (Gal & Ghahramani, 2016) nous permet d'approximer cette intégrale. En effectuant T passages stochastiques à travers le réseau avec le *dropout* activé, nous obtenons T prédictions pour chaque entrée de test $\mathbf{x}^*$. La distribution empirique de ces prédictions se rapproche de la distribution prédictive $p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y})$. Mathématiquement, cela peut s'exprimer comme suit :

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) \approx \frac{1}{T}\sum_{t=1}^{T} p(\mathbf{y}^*|\mathbf{x}^*, \theta_t) \tag{2}$$

C Accords inter-annotateurs

FIGURE 6 – Kappa de Cohen sur `RLF27k` et `RezoJDM16k`.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

CQuAE : Un nouveau corpus de question-réponse pour l'enseignement

Thomas Gerald[1*] Louis Tamames[2*] Sofiane Ettayeb[2] Patrick Paroubek[1]
Anne Vilnat[1]
(1) Université Paris Saclay, CNRS, LISN
(2) Stellia
(1) prenom.nom@lisn.upsaclay.fr, (2) prenom.nom@stellia.ai

RÉSUMÉ

Dans cet article nous présentons un nouveau corpus de question-réponse en français pour le domaine de l'éducation. Ce corpus a été construit dans le but de créer un système d'assistant virtuel pour répondre à des questions sur des documents ou du matériel de cours. Afin d'être utile autant aux enseignants qu'aux étudiants, il est important de considérer des questions complexes ainsi que d'être capable de justifier les réponses sur du matériel validé. Nous présentons donc le nouveau corpus CQuAE, un corpus de questions-réponses manuellement annoté dont nous discutons des propriétés. Nous présenterons aussi les différentes étapes de sa création avec aujourd'hui une phase d'amélioration des données. Enfin, nous présentons plusieurs expériences pour évaluer l'exploitation du corpus dans le cadre d'un système de questions-réponses automatique. Ces différentes analyses et expériences nous permettrons de valider l'adéquation des données collectées pour l'objectif visé.

ABSTRACT

CQuAE : A new question-answering corpus for teaching assistant

In this article we present a new French question-answering corpus for the education domain. This corpus has been built with the aim of creating a virtual assistant system for answering questions on documents or course material. In order to provide a useful tool for both teachers and students, it is important to consider complex questions and to be able to justify answers based on validated material. We therefore present the new CQuAE Corpus, a manually annotated question-answering corpus and discuss its properties. We will also present the different stages in its creation, with a recent focus on data enhancement. Finally, we present experiments to evaluate the exploitation of the corpus within the framework of an automatic question-answering system. These different analyses and experiments will enable us to validate the suitability of the data collected for the intended purpose.

1 Introduction

Ce travail s'inscrit dans le domaine de l'éducation, notamment dans le but d'aider les étudiants à apprendre et à réviser leurs cours en leur fournissant des questions sur les documents de cours choisis par l'enseignant, ainsi que les réponses associées. L'objectif est de dépasser les simples questions factuelles auxquelles il est facile de répondre et de pouvoir poser des questions complexes qui vont au-delà de la recherche d'une réponse par entité nommée. Par exemple, face à un cours sur les débuts de la Révolution française, nous ne voulons pas seulement demander quand a eu lieu la prise de la Bastille, mais aussi quelles sont les raisons qui ont conduit les manifestants à le faire, nous rapprochant ainsi des questions de cours qu'un enseignant pourrait poser. Pour aider les enseignants dans cette

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 50–63.

démarche chronophage, nous avons travaillé à la constitution d'un tel corpus afin d'automatiser la tâche de création de questions et de réponses, en commençant par la construction manuelle d'un premier corpus.

À l'heure actuelle, aucun corpus ne répond à l'ensemble de ces critères, à savoir des questions et des réponses qui peuvent être complexes, qui s'appuient sur des ensembles de documents validés mais courts (les cours de l'enseignant), et qui sont en langue française. Nous souhaitons travailler sur plusieurs disciplines, éventuellement à différents niveaux d'enseignement, nous avons commencé notre étude par l'histoire telle qu'elle est enseignée à la fin du collège et au début du lycée. Pour avoir une base de comparaison, nous avons également effectué quelques tests sur la géographie, les sciences de la vie et l'éducation civique. Nous avons ainsi constitué un corpus contenant :

— des questions créées à partir de documents de cours, non seulement des questions factuelles mais aussi des questions plus complexes
— des réponses qui sont soit extraites du cours, soit construites à partir de plusieurs éléments disséminés dans le document,
— le document source, qui valide à la fois l'intérêt de la question et la qualité de la réponse produite.

Aujourd'hui nous en sommes à la seconde phase de constitution du corpus, nous avons déjà collecté plus de 11000 annotations que nous présentons dans cet article. Par ailleurs, nous décrirons aussi une phase de correction du corpus afin d'améliorer les données déjà collectées.

Avec ces différents éléments, ce corpus pourrait être utilisé pour former les composants d'un cadre de Génération Augmentée par Récupération (RAG). À cette fin, nous proposons dans ce travail de mesurer l'adéquation de l'ensemble de données pour développer une telle application via l'adaptation de modèles de langue.

Nous détaillerons d'abord comment nous avons collecté un nouveau corpus conçu pour cette tâche. Ensuite, nous présenterons une analyse de ce corpus afin d'en présenter le contenu. Puis, nous discuterons de la phase d'amélioration du corpus. Enfin, la partie suivante sera consacrée à la présentation des expériences que nous avons menées afin de démontrer la valeur de cet ensemble de données pour l'apprentissage d'un système de RAG, en comparant plusieurs grands modèles de langue et les différentes versions de notre corpus.

2 Travaux connexes

Les grands modèles de langue. La génération de résumés automatiques, de questions ou bien de réponses sont aujourd'hui des sujets centraux de recherche pour la communautés du TAL. Ces différentes tâches ont bénéficié des évolutions des algorithmes d'apprentissage profond. En particulier les architectures neuronales comme le modèle "Transformer" (Vaswani *et al.*, 2017) permettent aujourd'hui d'obtenir de bonnes performances pour ces différents objectifs.

Dans un premier temps les modèles de langues exploitant ces architectures ont été développés pour la langue anglaise, aujourd'hui plusieurs variantes en langue française existent comme les modéles *CammemBERT* ou *FlauBERT* (Martin *et al.*, 2020; Le *et al.*, 2020) pour la classification et le modèle *BARThez* pour des tâches de génération (Eddine *et al.*, 2021). Aujourd'hui la taille des modèles de langues permet de prendre en compte conjointement différentes langues, dans des approches multilingues (Scao *et al.*, 2022).

Adaptation des grands modèles de langue. Ces modèles de langue sont en général appris sur

des tâches de reconstruction de l'entrée ou de prédiction du prochain mot (ou jeton). Dès lors, il est nécessaire d'adapter ces modèles à la tâche visée. Pour adapter le modèles à une tâche, comme la classification ou la génération, une première méthode consiste à adapter les poids du réseaux de neurones, cette approche est connue sous le nom de *fine-tuning* (ou adaptation fine).

Cette approche est néanmoins sujette à différents inconvénients ; d'une part l'optimisation de l'intégralité des paramètres peut être prohibitive en temps de calculs ; d'autre part l'optimisation de tous les poids du modèles peut mener à des problèmes de sur-apprentissage.

Les méthodes appelées *adapter* tentent aujourd'hui de palier ces problèmes (Pfeiffer *et al.*, 2020) en introduisant dans les différents blocs de nouveaux sous-réseaux de neurones qui seront les seuls optimisés durant l'étape de *fine-tuning*. De plus ce type d'approche permet de conserver les poids du modèles original (appelé modèle pré-entraîné). Bien que le coût de l'adaptation soit diminué, le coût de l'inférence (l'étape de prédiction) est légèrement augmenté. Récemment une nouvelle approche connue sous le nom de *"LOw Rank Adaption"* (Hu *et al.*, 2022) propose à la fois de diminuer le coût d'adaptation tout en préservant un coût d'inférence similaire à celui du modèle pré-entraîné.

Les corpus de question-réponse. Chacune de ces approches pour adapter le modèle nécessite néanmoins des collections de données importantes. Nous allons présenter les différents corpus de question-réponse existants et leur intérêt pour la communauté. Le corpus SQuAD (Rajpurkar *et al.*, 2016) est l'un des premier corpus de grande taille pour l'apprentissage de modèle de question-réponse s'appuyant sur un contexte (la réponse devant s'y trouver) avec plus de 20.000 exemples. Plus récemment, Google a publié le corpus Natural Question (Kwiatkowski *et al.*, 2019) qui est un corpus de questions en langage naturel, avec des paragraphes longs et courts pour les réponses (extraits de la version anglaise de Wikipédia). Pour les approches de question-réponse conversationnelle, les corpus CANARD et QUAC (Elgohary *et al.*, 2019; Choi *et al.*, 2018) sont disponibles. Pour les questions-réponses basées sur la recherche de documents ou de passages, le corpus MSMarco (Nguyen *et al.*, 2016) est aujourd'hui une référence avec plus d'un million de questions. Si la plupart des corpus sont disponibles en anglais, la communauté française a également produit des corpus tels que FQuAD (d'Hoffschmidt *et al.*, 2020), Piaf (Keraron *et al.*, 2020) ou CALOR-QUEST (Bechet *et al.*, 2019) toujours dans le but d'extraire la réponse du contexte. Cependant, ces corpus reposent principalement sur des réponses factuelles, correspondant à un texte court, comme une entité nommée, un événement, une date, une quantité ou un lieu. Récemment, un nouveau corpus Autogestion (Antoine *et al.*, 2022) a été créé pour traiter les questions non factuelles, l'étude associée démontre l'incapacité des modèles standards à traiter les questions les plus complexes. Néanmoins à notre connaissance il n'existe aujourd'hui pas de corpus ouvert spécifique à l'enseignement secondaire pour des tâches de question-réponse complexes.

3 Le corpus

3.1 Récolte du corpus

Notre corpus se compose de paragraphes de cours, puis de questions et de réponses fondées sur ces textes. La première étape a donc été de recueillir un ensemble de textes en Français dans le domaine éducatif, en s'appuyant sur des livres scolaires (des cours) concernant les niveaux collège et lycée, principalement en Histoire mais aussi en Géographie, en Sciences de la Vie et de la Terre et Éducation Civique. Les premiers textes proviennent du site "le livre scolaire" [1]. Pour compléter

1. https ://www.lelivrescolaire.fr/

ce contenu, nous avons recherché des articles Wikipedia liés à ces sujets scolaires. Nous les avons filtrés en utilisant des API Wikipedia avec des requêtes construites à partir des titres de chapitres des livres, et en réunissant les sous-sections retournées. Pour ne pas avoir des contenus trop gros, nous avons découpé les articles en ne conservant pas plus de trois paragraphes par document. Un article Wikipedia va donc correspondre à plusieurs documents. Nous avons ainsi globalement réuni 3.891 documents (dont seulement 1.122 sont annotés), constitués de 14.433 paragraphes (dont seulement 3.893 sont annotés). Nous avons ensuite procédé à des campagnes d'annotation. Le principe est de présenter un paragraphe aux annotateurs en leur demandant de créer les annotations suivantes : (a) **une question** à poser ; (b) **le type de la question** qui peut-être factuelle, définition, cours ou synthèse ; (c) **le support de la question**, à savoir l'extrait du document à partir duquel la question est construite ; (d) **les éléments de réponse**, c'est à dire les passages permettant de répondre à la question ; (e) **la réponse** rédigée par l'annotateur, à partir des éléments précédents. Les annotateurs devaient créer environ 10 annotations (et plus si possible) pour chaque document. La Table 1 donne des exemples de questions, et de leurs supports.

L'un de nos objectifs principaux est de recueillir des questions requérant des niveaux d'expertise différents pour y répondre. Ce niveau de "difficulté" est lié au type de la question qui peut donc être :

- **factuelle** : la réponse est un fait ou une liste de faits (événement, personne, lieu, date...) ;
- **définition** : la réponse correspond à la définition d'un mot ou d'un concept ;
- **cours** : la réponse n'est pas réduite à un fait mais contient des explications ou des détails, qui doivent être explicites dans le document ;
- **synthèse** : la réponse s'appuie sur plusieurs éléments du document fournissant des informations diverses qui doivent être réunies ou impliquant une interprétation pour être produite.

Pour garantir d'avoir suffisamment de questions complexes, avec leurs réponses (autres que factuelles ou définitions), nous avons demandé un ratio de 40 % de factuelles et définitions, et 60 % questions de cours et de synthèses. Cependant, nous avons demandé également de ne pas créer artificiellement des questions complexes quand le document ne s'y prête pas, le ratio n'est donc pas strict. Les questions *factuelles* et *définition* sont assez simples à formuler et découlent directement du texte du document. Pour les questions de *cours*, elles sont un peu plus complexes et les réponses nécessitent plus de détails. La réponse aux questions de *synthèse* nécessite un raisonnement à partir du document.

Deux groupes d'annotateurs ont travaillé sur le corpus : le groupe A composé d'une vingtaine d'annotateurs personnes ayant un bon niveau général, mais sans expérience d'enseignement, le groupe B est composé de 6 annotateurs ayant une expérience en enseignement [2]. Dans la Table 2 nous indiquons la distribution actuelle en termes de type de questions pour les deux groupes. On notera que le groupe A a proportionnellement produit plus de questions de cours que le B, alors que le B s'est concentré sur les questions de synthèse. Pour aider les annotateurs, un guide a été fourni et est disponible avec le corpus sur le dépôt gitlab du corpus [3]. Aussi, nous reportons en annexe le nombre de questions et de documents par domaine et par source (*le livre scolaire* ou *wikipedia*).

3.2 Analyse du corpus

Pour une première analyse, nous avons comparé la longueur des questions et des réponses dans notre corpus, en fonction du type de la question, et en la comparant avec ce qu'on peut observer dans les deux corpus existant en français et déjà évoqués, à savoir FQuAD (d'Hoffschmidt *et al.*, 2020) et Piaf (Keraron *et al.*, 2020). La Figure 1 illustre cette étude.

2. Il est plus facile de recruter des personnes dans le groupe A que B, d'où la différence de taille des deux groupes.
3. https://gitlab.lisn.upsaclay.fr/gerald/cquae

Type	Question	Support
Factuelle	En quelle année Christophe Colomb a-t-il découvert l'Amérique ?	Christophe Columb découvre l' Amerique (1492)
Définition	Qu'est-ce qu'une presse rotative ?	Une presse rotative est une presse typographique montée sur un cylindre, permettant une impression en continu.
Cours	Comment les Européens ont-ils légitimé leur domination ?	Les Européens repensent la hiérarchie entre les peuples selon un schéma centré sur le christianisme et l'Europe qui leur sert ensuite à légitimer leur domination.
	Quels sont les noms de ceux qui indiquent comment pratiquer la religion musulmane ? Sur quel texte s'appuient-ils pour le faire ?	Ce sont les ulemas qui régissent la religion, en s'appuyant sur la loi de la Sharia.
Synthèse	Pourquoi certains français sont-ils favorables à l'état d'urgence après les attaques à Paris en 2015 ?	• il les protège de la menace terroriste et du risque d'une nouvelle attaque, qui est redoutée par tous. • Ce régime exceptionnel continue à apparaître comme une "nécessité"...
	Qui doit être impliqué pour lutter contre le changement climatique d'après Matt Petersen ? Comment ?	Matt Petersen travaille sur le développement soutenable dans la ville de Los Angeles, aux côtés du maire de la ville [...] nous avons besoin de tous. Tout sourire, le maire de Los Angeles a connecté [...] des panneaux solaires installé sur des toits privés [...] ...
	Pourquoi cet article parle de "victoire du féminisme" pour décrire le mouvement des midinettes ?	Il ne faut pas médire des midinettes. Il n'est pas d'un bon esprit de les taxer de frivolité parce qu'elles travaillent dans les robes, qu'elles sont jeunes et [...] de la femme, s'exerçant en ces jours tragiques au préjudice de milliers et de milliers d'ouvrières, d'employées, voire de fonctionnaires, est d'une si cruelle injustice qu'elle soulève des protestations de tous les côtés.

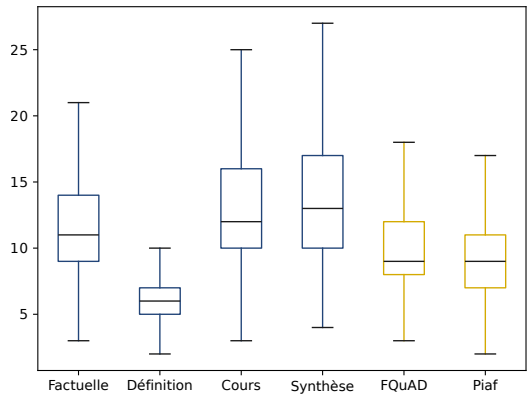
TABLE 1 – Exemples des quatre types de questions, avec leurs supports (extrait du document la justifiant)

Type de question	Groupe A	Groupe B	Total
Factuelle	2 106	294	2 400
Définition	1 506	181	1 687
Cours	4 784	490	5 274
Synthèse	1 756	338	2 094
Total	10 152	1 303	11 455

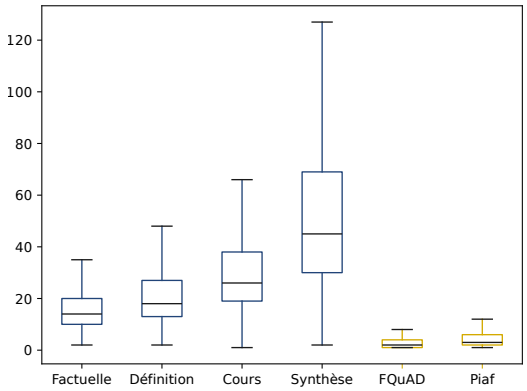
TABLE 2 – Statistiques pour les types de questions pour chaque groupe (A = éduqués, B = avec expérience de l'enseignement)

Les questions de *définition* sont toujours beaucoup plus courtes, les trois autres types sont assez comparables, les questions de synthèse (plus difficiles) étant un peu plus longues. Pour ces trois catégories l'écart-type est assez important. FQuAD et Piaf sont assez similaires, avec des questions plus courtes et moins d'écart-type. On note que les questions *factuelles* sont les plus proches de celles de ces deux datasets, ce qui confirme le fait que les questions y sont surtout factuelles. La même étude a été faite sur la longueur des réponses. Logiquement les réponses aux questions de *synthèse* sont plus longues que les autres, et parfois même très longues. Les réponses factuelles sont les plus courtes. Les réponses dans FQuAD et Piaf sont significativement plus courtes que les nôtres, même dans les *factuelles*. En effet, nous avons demandé une réponse rédigée aux annotateurs quant à FQuAD et Piaf la réponse est extraite du contexte. Notre corpus est pour cela différent des datasets existants.

Réponses extraites du contexte. Bien que nous demandions aux annotateurs de rédiger une réponse, nous avons remarqué qu'une grande partie des réponses sont directement issues du contexte. Nous nous proposons donc d'étudier la proportion de chaînes de caractères communes à la fois à la réponse et au contexte (paragraphes extraits sélectionnés par l'annotateur). Dans la table 3, nous reportons les résultats de cette analyse pour chacun des types de question. Nous capturons la plus grande chaîne de caractères commune entre la réponse et les paragraphes et nous en reportons la proportion par rapport à la taille de la réponse rédigée. Un coup d'oeil au résultats présentés dans le tableau nous montre que la proportion de mots communs est répartie différemment selon le type de la question. En



(a) Longueur de la question en fonction de son type dans notre corpus, comparée à FQuAD et Piaf



(b) Longueur de la réponse en fonction de son type dans notre corpus, comparée à FQuAD et Piaf

FIGURE 1 – Taille de la question ou de la réponse par type de question en comparaison du corpus Piaf et FQuAD

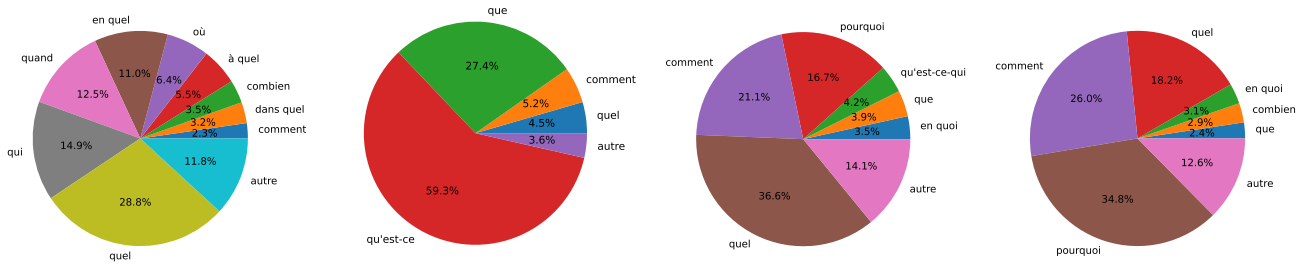
	Mediane	1^{er} quartile	3^{me} quartile
Factuelle	32.6 %	22.9 %	45.8 %
Définition	27.1 %	17.0 %	43.9 %
Cours	24.5 %	15.5 %	40.2 %
Synthèse	12.2 %	7.5 %	21.5 %
Total	24.4 %	14.6 %	39.5 %

TABLE 3 – Pourcentage de la plus longue chaîne de caractère commune à la réponse et au contexte.

particulier, les questions dites factuelles, comporte une forte proportion du contexte, cela s'explique par la réponse attendue ainsi que par la taille de la réponses.

Au contraire, les questions de synthèse comportent une proportion faible de contexte consécutif car les réponses nécessitent plusieurs éléments du texte et sont plus longues.

Caractérisation des types de questions. Pour caractériser les questions en fonction des différents types nous nous sommes demandés quels mots interrogatifs étaient utilisés dans la question. Pour extraire les mots interrogatifs, nous avons lémmatisé les questions et définis un certain nombre de formules interrogatives (quand, comment, où, etc...). Enfin nous avons regardé les correspondances avec les premiers lemmes de la question. Notons que pour 409 questions nous n'avons pas trouvé de correspondance, plusieurs raisons en sont la cause, d'une part certaines questions utilisent ces formes interrogatives mais pas au début de la question, par exemple :
"Selon Bartolomé de Las Casas, dans son livre Très brève relation de la destruction des Indes publié en 1552, quelle a été la raison du massacre des Indiens par les Espagnols lors de la conquête des Indes sous le règne de Charles Quint ?"
Par ailleurs, certaines structures ne sont pas capturées ou ne sont pas des questions comme par exemple :
"En vous basant sur ces documents, citez un argument en faveur de la loi et un autre contre."
Nous reportons les résultats des différentes formes interrogatives capturées pour chaque types de question dans la figure 2. On remarquera que les distributions des formes interrogatives sont dépendantes du type de la question ; ainsi pour les questions de synthèse les mots clefs *"pourquoi"* et *"comment"* sont prépondérants, ceux-ci impliquant une explication dans la réponse ; pour les questions de définition on retrouvera le mot *qu'est-ce* menant vers une description d'un concept. Pour les



(a) Factuelle (b) Définition (c) Cours (d) Synthèse

FIGURE 2 – Comparaison des mots de la question pour les différents types de question

questions factuelles les mots clefs comme *"quel"*, *"qui"*, *"quand"* ou *"en quel"* entraînent une réponse de type entité nommée. Pour les questions de cours en revanche nous avons une panoplie de mots interrogatifs amenant à une réponse explicative ou factuelle. Cette toute première étude sera enrichie par la suite.

4 Correction et amélioration des annotations

Suite à une première pré-évaluation du corpus, nous avions remarqué que certaines questions et/ou réponses contenaient des erreurs ou des imprécisions. Nous avons remarqué plusieurs types d'erreurs dans les données : des erreurs de syntaxe ; des erreurs liées à l'ajout d'informations non présentes dans le document ou à l'inverse des réponses incomplètes ; des questions en dehors du sujet ou non pertinentes dans le cadre de l'éducation. Par ailleurs nous avons émis l'hypothèse que de telles erreurs pourraient être à l'origine d'imprécisions dans l'évaluation et l'entraînement de systèmes de questions et réponses automatique. Pour répondre à cette problématique, nous nous sommes proposés de mettre en place une seconde campagne, visant à évaluer le taux d'annotations comportant des erreurs.

Évaluation des données erronées. Pour évaluer la pertinence des données annotées, cinq annotateurs ont été mis à contribution pour juger un ensemble de 9243 annotations. Nous leur avons proposé d'annoter les questions et réponses en suivant les critères binaires suivants :

- **Q+ :** la question est correcte
- **Q- :** la question est corrigeable (syntaxe, reformulation possible)
- **R+ :** la réponse est correcte
- **R- :** la réponse est corrigeable (syntaxe, mauvaise entités, etc...)
- **HS :** le couple question/réponse est non pertinent et ne peut pas être corrigé

Pour cette évaluation nous reportons les résultats dans le table 4 conditionnellement au type de question. Pour les erreurs dans les questions, il semblerait que celles de type définitions comportent moins d'erreur, avec plus de 81 % d'annotations ne nécessitant aucune correction. Pour les autres catégories de questions, les erreurs semblent être réparties de manière similaire (entre 65 % et 68 %). Pour les erreurs dans les réponses, ce sont les catégories cours et synthèse qui nécessitent le plus grand nombre de corrections ; les réponses étant plus longues, des fautes de syntaxe sont plus probables.

Sur les 9243 annotations 192 ont été évaluées par tous les annotateurs, nous permettant de calculer l'accord inter-annotateur. Nous avons obtenu un kappa de Fleiss [4] pour le critère "la question est-elle

4. https://fr.wikipedia.org/wiki/Kappa_de_Fleiss

	Q+	Q-	R+	R-	HS
Factuelle	65.4	28.5	62.0	31.8	5.9
Définition	81.1	14.6	64.9	30.4	4.2
Cours	67.7	26.6	52.6	41.4	5.5
Synthèse	65.7	28.1	49.4	43.5	6.3
Tout type	68.9	25.4	55.8	38.1	5.6

TABLE 4 – Pourcentages de questions (Q) et de réponses (R) correctes (+) ou corrigeables (-) et de couples Q-R non pertinents (HS) par type de question lors de l'évaluation de notre corpus

correcte ?" de .30 (avec un accord maximum kappa de Cohen [5] de .44 et minimum de .18), pour la réponse on obtient .18 (accord maximum de .40 et minimum de .29). Il est clair que nous n'avons pas un accord probant entre les annotateurs, ces résultats démontrent ainsi la difficulté de la tâche. Pour se faire une idée de la pertinence de l'évaluation des documents corrigés, nous proposerons dans la section suivante une évaluation des corrections afin de juger de la pertinence des modifications apportées.

Correction du corpus. Dans un second temps, nous nous sommes proposés de corriger les annotations ayant au moins la réponse ou la question évaluée comme corrigeable. Nous avons pu à ce jour revoir 2565 réponses et 1840 questions sur 3140 annotations. En définitive, nous avons un corpus composé de 8180 questions et réponses vérifiées.

Bien que pour la phase d'évaluation, il est difficile de parler d'un réel accord entre les annotateurs, nous pensons que les corrections effectuées tendent bien vers l'amélioration du corpus. Pour vérifier cette hypothèse, nous nous sommes proposés d'évaluer les corrections manuellement. Pour ce faire nous avons demandé à trois personnes de déterminer d'après eux, quel est le meilleur couple de question et réponse. Notons que dans certains cas aucune préférence n'a été exprimée, nous considérons cette possibilité comme un troisième label. Nous avons obtenu pour cette évaluation avec 3 annotateurs une préférence pour les éléments corrigés dans 76. % des cas, pour les annotations originales 15.3 % étaient préférées et dans 8.3 % des cas il n'y avait aucune préférence. Pour ces différents résultats nous obtenons un accord inter-annotateurs moyen de .51. Notons que si nous supprimons les cas où les annotateurs sont restés indécis nous obtenons un accord de .62 montrant ainsi un accord important. Ces résultats montrent donc l'intérêt d'avoir corrigé certaines annotations pour améliorer le corpus.

5 Génération de réponses

Le corpus annoté a pour objectif la création d'un assistant scolaire pour les thématiques d'enseignement secondaire, particulièrement sous la forme d'un système de génération de questions et/ou réponses. Ainsi il nous parait pertinent de juger de la qualité de la génération de ces modèles pour nos données. Dans cette section nous nous proposons d'évaluer automatiquement la tâche de génération de réponse étant donnée un document et une question. Notons aussi, que dans le cas d'un assistant pour l'enseignement, la recherche de documents à partir d'un corpus vérifié (mis à disposition par l'enseignant) est une des tâches nécessaire au fonctionnement du système. Ainsi, nous proposerons de comparer la dernière configuration considérant les documents cibles donnés par les annotateurs ou retrouvés via la méthode BM25(Robertson & Zaragoza, 2009).

5. https://fr.wikipedia.org/wiki/Kappa_de_Cohen

Qtype	Entrainement-v1	Entrainement-v2	Validation-v2	Test-v2
Factuelle	2144	1409	128	128
Définition	1431	1075	128	128
Cours	5018	3409	128	128
Synthèse	1838	1263	128	128
Tous	10431	7156	512	512

TABLE 5 – Nombre de couples Q-R utilisés pour l'adaptation des modèles de langue provenant de la version corrigée (v2) ou non (v1), en fonction du type de question

5.1 Protocole expérimental

Modèles. Nous nous sommes penchés sur l'entraînement de grand modèle de langue de 7 milliards de poids, choisissant ainsi Llama2 7b et Mistral 7b. Dans l'ensemble des entraînements, nous avons utilisé LoRA et la quantification des modèles en 8 bits (paramètres en annexe), cela permet de réduire le temps de calcul et d'inférence avec une perte en performance minime. Nous avons fixé la longueur maximale des phrases à 2048 tokens car l'entrée et la sortie maximale en utilisant nos données ne dépasse pas les 2000 tokens. Les modèles ont été entraînés sur 3 *epochs* sur l'ensemble du dataset, et nous avons sélectionné la sauvegarde avec la meilleure *loss* pour l'évaluation.

Ensemble d'entraînement et de test. Les ensembles d'entraînement utilisés sont des sous ensembles de la première version du corpus (v1) et de la version corrigée (v2). De plus, les ensembles de validation et de test sont extraits de la v2, respectivement pour déterminer l'arrêt de l'apprentissage ainsi que pour calculer les résultats présentés dans ce document. Nous obtenons ainsi les différents ensembles reportés dans le tableau 5.

Métriques pour l'évaluation. Pour évaluer la capacité des modèles à répondre aux questions, nous avons utilisé trois métriques principales : ROUGE, BERTScore, et GPT. ROUGE mesure la similarité entre les réponses générées et les réponses de référence en évaluant la correspondance des n-grams, utile pour apprécier la proximité des générations au contenu. BERTScore compare la similarité sémantique à l'aide des plongements de mots. L'évaluation par GPT (version gpt-3.5-turbo-0125) se fait en donnant une réponse de référence et une générée pour obtenir une note de 0 à 10 mais aussi une explication sur la qualité suivant des critères que nous avons définis. Ces métriques sont loin d'être parfaites, mais permettent de nous donner une intuition sur la pertinence des réponses générées.

5.2 Génération de réponse en contexte

	Paragraphes de la question (gold)				Paragraphes retrouvés par BM25			
Model	GPT3.5	B-Score	R-1	R-L	GPT3.5	B-Score	R-1	R-L
MISTRAL	**8.45**	0.789	0.399	0.345	**7.82**	0.762	0.342	0.29
MISTRAL-V1	8.40	**0.854**	**0.605**	**0.54**	7.62	**0.826**	**0.518**	**0.459**
MISTRAL-V2	8.41	0.852	0.598	0.532	7.62	0.825	0.51	0.451
LlaMA	**8.53**	0.717	0.183	0.148	**8.05**	0.699	0.153	0.124
LlaMA-V1	8.37	**0.848**	**0.587**	0.527	7.58	**0.822**	**0.509**	**0.453**
LlaMA-V2	8.33	0.846	0.584	0.52	7.63	0.821	0.504	0.447

TABLE 6 – Résultats pour la génération de réponses étant donné un contexte pris dans le corpus d'origine, ou retrouvé par BM25

Dans cette expérience nous tentons d'évaluer la capacité des modèles à produire des questions et

réponses en utilisant les informations contenues dans les paragraphes du corpus. Pour ce faire, nous proposons d'adapter les deux modèles Llama et Mistral en utilisant avec les configurations données dans la section 5.1. Nous utilisons les modèles *chat* et *instruct* de Meta et Mistral, qui sont faits pour répondre à une instruction. Nous utilisons pour l'entraînement et l'évaluation le prompt suivant :

"En se basant exclusivement sur le document, répondez à la question. La réponse est destinée à un élève voulant améliorer sa compréhension du cours. : {question} document :{documents} Réponse :".

Nous comparons aussi les résultats des modèles appris sur la v1, la v2 et le modèle pré-entraîné. Les différents résultats sont reportés dans le tableau 6.

Exceptée la métrique basée sur GPT, les résultats sont significativement meilleurs en adaptant les modèles sur notre corpus [6], ce phénomène indiquant que l'adaptation est nécessaire pour espérer avoir des réponses semblables à celles de notre corpus. Notons aussi que l'ensemble d'entraînement de la version corrigée est amputé d'environ 30 % des exemples, malgré cela, les performances entre les deux versions sont comparables [7]. Si l'on s'intéresse à la métrique se basant sur le score donné par GPT, l'interprétation est ici différente. En effet, les scores attribués au modèle adapté avec vérité terrain (première partie du tableau) sont similaires à ceux produits par le modèle pré-entraîné. Lorsque les paragraphes correspondent à ceux données par l'approche BM25 (un seul paragraphe), ce score est alors favorable au modèle non adapté. Bien que nous puissions difficilement en affirmer la cause, plusieurs hypothèses sont possibles d'après nos observations. D'une part, le modèle non-adapté produit souvent des réponses en langue anglaise, le score donné par le modèle GPT ne pénalise pas la langue utilisée. Une deuxième hypothèse est que le modèle non-adapté pourrait produire des réponses dont les éléments ne sont pas présents dans le contexte. Cette dernière hypothèse est appuyée par la différence de score des modèles utilisant les deux configurations des paragraphes.

Nous savons que les documents retrouvés ne correspondent pas intégralement à des documents pertinents d'après les résultats de la méthode BM25 (62 % des paragraphes étant dans la vérité terrain). En incluant BM25, les performances supérieures pour le modèle non-adapté laissent supposer que ce modèle est moins dépendant du contexte.

Notons enfin que la métrique GPT est difficile à exploiter, celle-ci dépendant à la fois du prompt, du modèle GPT et de la graine de génération. Il est donc difficile avec les éléments dont nous disposons de statuer sur les hypothèses proposées.

6 Conclusion

Dans cet article nous avons présenté un nouveau corpus CQuAE de question-réponse dans le domaine de l'enseignement secondaire. Après avoir analysé les versions du corpus recueilli, nous avons constaté les difficultés qu'il y avait à en garantir la qualité, du fait de la difficulté de la tâche. Nous avons discuté des améliorations de ce corpus avec une phase de correction des annotations. Enfin nous avons proposé d'étudier la pertinence de ces modèles pour la génération de réponses avec des grands modèles de langue. Avec les différentes analyses et conclusions sur les évaluations humaines nous pensons qu'un tel corpus serait bénéfique pour la communauté enseignante et celle du TAL. Cependant, avec les résultats dont nous disposons, il reste difficile de statuer sur les différentes hypothèses émises. Nous planifions donc d'explorer les différents résultats obtenus via une évaluation humaine. Actuellement, la version corrigée du corpus comporte un nombre d'exemples réduit par

6. $pvalue$ pour T-test d'indépendance < 0.01
7. $pvalue$ pour T-test d'indépendance > 0.05

rapport au corpus originalement produit. Pour ce dernier point, nous avons prévu de réviser les exemples manquant dans le corpus corrigé.

Références

ANTOINE E., AUGUSTE J., BÉCHET F. & DAMNATI G. (2022). Génération de questions à partir d'analyse sémantique pour l'adaptation non supervisée de modèles de compréhension de documents. *29e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*.

BECHET F., ALOUI C., CHARLET D., DAMNATI G., HEINECKE J., NASR A. & HERLEDAN F. (2019). CALOR-QUEST : un corpus d'entraînement et d'évaluation pour la compréhension automatique de textes. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019*.

CHOI E., HE H., IYYER M., YATSKAR M., YIH W., CHOI Y., LIANG P. & ZETTLEMOYER L. (2018). Quac : Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* : Association for Computational Linguistics.

D'HOFFSCHMIDT M., BELBLIDIA W., HEINRICH Q., BRENDLÉ T. & VIDAL M. (2020). Fquad : French question answering dataset. In T. COHN, Y. HE & Y. LIU, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 de *Findings of ACL*, p. 1193–1208 : Association for Computational Linguistics. DOI : 10.18653/V1/2020.FINDINGS-EMNLP.107.

EDDINE M. K., TIXIER A. J. & VAZIRGIANNIS M. (2021). Barthez : a skilled pretrained french sequence-to-sequence model. In *EMNLP (1)*.

ELGOHARY A., PESKOV D. & BOYD-GRABER J. L. (2019). Can you unpack that ? learning to rewrite questions-in-context. In *EMNLP-IJCNLP* : Association for Computational Linguistics.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2022). Lora : Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022* : OpenReview.net.

KERARON R., LANCRENON G., BRAS M., ALLARY F., MOYSE G., SCIALOM T., SORIANO-MORALES E. & STAIANO J. (2020). Project PIAF : building a native french question-answering dataset. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, p. 5481–5490 : European Language Resources Association.

KWIATKOWSKI T., PALOMAKI J., REDFIELD O., COLLINS M., PARIKH A. P., ALBERTI C., EPSTEIN D., POLOSUKHIN I., DEVLIN J., LEE K., TOUTANOVA K., JONES L., KELCEY M., CHANG M., DAI A. M., USZKOREIT J., LE Q. & PETROV S. (2019). Natural questions : a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, **7**, 452–466. DOI : 10.1162/TACL_A_00276.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). Flaubert : Unsupervised language model pre-training for french. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of The 12th Language Resources and Evaluation Conference,*

LREC 2020, Marseille, France, May 11-16, 2020, p. 2479–2490 : European Language Resources Association.

MARTIN L., MÜLLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. R. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, p. 7203–7219 : Association for Computational Linguistics. DOI : 10.18653/V1/2020.ACL-MAIN.645.

NGUYEN T., ROSENBERG M., SONG X., GAO J., TIWARY S., MAJUMDER R. & DENG L. (2016). MS MARCO : A human generated machine reading comprehension dataset. In T. R. BESOLD, A. BORDES, A. S. D'AVILA GARCEZ & G. WAYNE, Éds., *Proceedings of the Workshop on Cognitive Computation : Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 de *CEUR Workshop Proceedings* : CEUR-WS.org.

PFEIFFER J., RÜCKLÉ A., POTH C., KAMATH A., VULIĆ I., RUDER S., CHO K. & GUREVYCH I. (2020). AdapterHub : A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 46–54, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-demos.7.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). Squad : 100, 000+ questions for machine comprehension of text. In J. SU, X. CARRERAS & K. DUH, Éds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, p. 2383–2392 : The Association for Computational Linguistics. DOI : 10.18653/V1/D16-1264.

ROBERTSON S. & ZARAGOZA H. (2009). The probabilistic relevance framework : Bm25 and beyond. *Found. Trends Inf. Retr.*, **3**(4), 333–389. DOI : 10.1561/1500000019.

SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIC S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F., GALLÉ M., TOW J., RUSH A. M., BIDERMAN S., WEBSON A., AMMANAMANCHI P. S., WANG T., SAGOT B., MUENNIGHOFF N., DEL MORAL A. V., RUWASE O., BAWDEN R., BEKMAN S., MCMILLAN-MAJOR A., BELTAGY I., NGUYEN H., SAULNIER L., TAN S., SUAREZ P. O., SANH V., LAURENÇON H., JERNITE Y., LAUNAY J., MITCHELL M., RAFFEL C., GOKASLAN A., SIMHI A., SOROA A., AJI A. F., ALFASSY A., ROGERS A., NITZAV A. K., XU C., MOU C., EMEZUE C., KLAMM C., LEONG C., VAN STRIEN D., ADELANI D. I. & ET AL. (2022). BLOOM : A 176b-parameter open-access multilingual language model. *CoRR*, **abs/2211.05100**. DOI : 10.48550/ARXIV.2211.05100.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. VON LUXBURG, S. BENGIO, H. M. WALLACH, R. FERGUS, S. V. N. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, p. 5998–6008.

7 Annexe

7.1 Sources et domaines

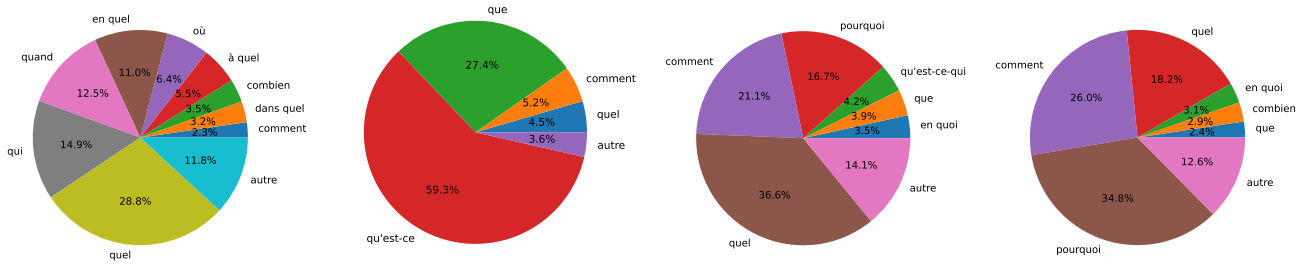
Dans la table 7 nous reportons le nombre de questions et de documents en fonction de la source, “lelivrescolaire” (préfixé par `lls`) et “wikipedia” (préfixé par `wik`), du domaine (géographie, histoire, svt) et du niveau .

Source	domaine-niveau	Version	nombre de questions	nombre de documents
lls	geographie-premiere	V1	572	48
		V2	409	48
lls	geographie-seconde	V1	716	64
		V2	533	64
lls	histoire-geographie-cinquieme	V1	816	86
		V2	588	86
lls	histoire-geographie-sixieme	V1	814	89
		V2	599	89
lls	histoire-premiere	V1	945	101
		V2	687	101
lls	histoire-seconde	V1	1146	103
		V2	782	102
lls	svt-cinquieme	V1	107	9
		V2	78	9
lls	svt-seconde	V1	204	18
		V2	148	18
wik	geographie-premiere	V1	1786	193
		V2	1260	189
wik	geographie-seconde	V1	546	51
		V2	374	50
wik	histoire-premiere	V1	1160	132
		V2	849	130
wik	histoire-seconde	V1	1996	155
		V2	1390	155
wik	histoire-geographie-cinquieme	V1	170	15
		V2	118	14
wik	histoire-geographie-sixieme	V1	184	17
		V2	139	17
wik	svt-cinquieme	V1	158	18
		V2	118	18
wik	svt-seconde	V1	135	23
		V2	108	23

TABLE 7 – Nombre de questions et documents par source et par domaine

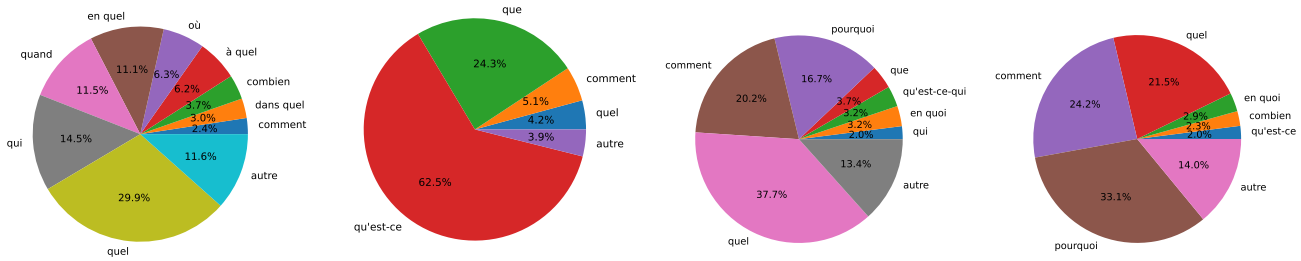
7.2 Comparaison des distributions du corpus avec FQuAD et PIAF

Les figures 3 et 4 décrivent les distributions des mots de la question pour les deux versions du corpus. Dans la figure 5 nous observons la distribution des mots de la question pour les corpus FQuAD (d’Hoffschmidt *et al.*, 2020) et Piaf (Keraron *et al.*, 2020). On pourra remarquer la similitude entre les distribution Piaf et FQuAD avec les questions “factuelles”. Sur les autres types, définition, cours, synthèse les distributions sont éloignées.



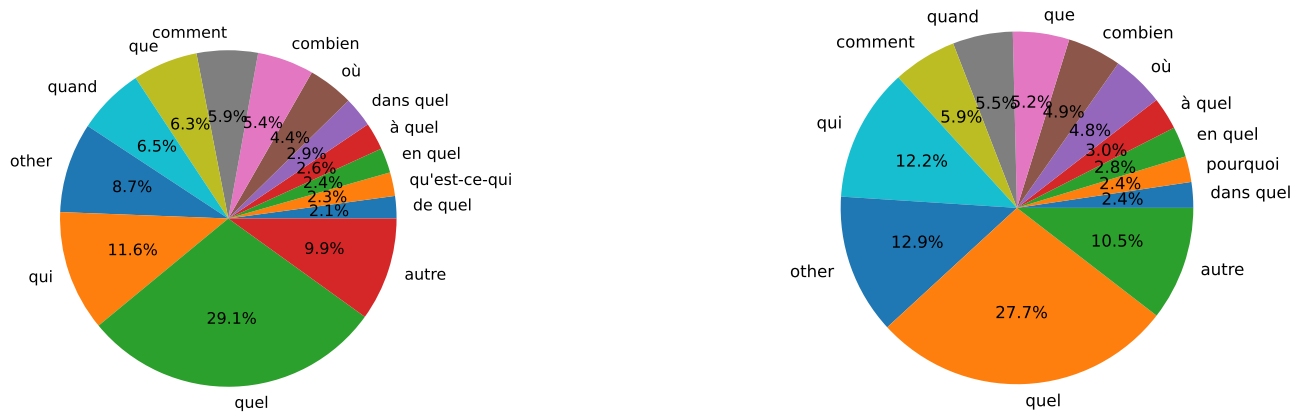
(a) Factuelle (b) Définition (c) Cours (d) Synthèse

FIGURE 3 – Comparaison des mots de la question pour les différents types de question sur la première version du corpus (voir section 3.2



(a) Factuelle (b) Définition (c) Cours (d) Synthèse

FIGURE 4 – Comparaison des mots de la question pour les différents types de question sur la seconde version du corpus (voir section 3.2



(a) FQuAD (b) Piaf

FIGURE 5 – Mots de la question pour les corpus FQuAD et Piaf

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Évaluation automatique des biais de genre dans des modèles de langue auto-régressifs

Fanny Ducel[1], Aurélie Névéol[1], Karën Fort[2]
(1) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
(2) Sorbonne Université / Universite de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
fanny.ducel@lisn.fr, aurelie.neveol@lisn.fr, karen.fort@loria.fr

RÉSUMÉ

Nous proposons un outil pour mesurer automatiquement les biais de genre dans des textes générés par des grands modèles de langue dans des langues flexionnelles. Nous évaluons sept modèles à l'aide de 52 000 textes en français et 2 500 textes en italien, pour la rédaction de lettres de motivation. Notre outil s'appuie sur la détection de marqueurs morpho-syntaxiques de genre pour mettre au jour des biais. Ainsi, les modèles favorisent largement la génération de masculin : le genre masculin est deux fois plus présent que le féminin en français, et huit fois plus en italien. Les modèles étudiés exacerbent des stéréotypes attestés en sociologie en associant les professions stéréotypiquement féminines aux textes au féminin, et les professions stéréotypiquement masculines aux textes au masculin.

ABSTRACT

Automatically Assessing Gender Biases in Autoregressive Language Models.

We propose a framework to automatically measure gender biases in generated texts, for inflected languages. We evaluate seven language models, on over 52,000 texts in French and 2,500 texts in Italian, for cover letter writing. Our tool relies on the detection of morpho-syntactic gender markers to uncover biases. Thus, models are strongly biased towards the generation of masculine markers : generated texts contain twice as many masculine (vs. feminine) markers in French, and eight times as many in Italian. The models we study also exacerbate gender stereotypes that are evidenced in social science studies and associate feminine inflections with stereotypically feminine occupations, whereas stereotypically masculine occupations are strongly associated with masculine markers.

MOTS-CLÉS : Biais, Stéréotype, Genre, Modèle de langue (LLM), Français, Italien.

KEYWORDS: Bias, Stereotype, Gender, Language Model (LLM), French, Italian.

1 Introduction

Au cours des dernières années, les grands modèles de langue (*Large Language Models*, ou LLM) sont devenus l'approche privilégiée pour la plupart des tâches de traitement automatique des langues (TAL) telles que la classification de textes, la reconnaissance d'entités nommées ou la traduction automatique (Howard & Ruder, 2018; Epure & Hennequin, 2022; Peng *et al.*, 2023), y compris pour des applications destinées au grand public. Néanmoins, ces modèles non seulement reproduisent, mais amplifient les biais stéréotypés (Gehman *et al.*, 2020; Dhamala *et al.*, 2021; Kirk *et al.*, 2021) qu'il est important de détecter et d'évaluer afin d'éviter qu'ils ne perpétuent des discriminations.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 64–84.

Modèle	Type	Taille	Langue(s)	Référence
xglm	Base	2,9M	FR, IT (Multi.)	(Lin *et al.*, 2022)
gpt2-fr	Base	1M	FR	(Simoulin & Crabbé, 2021)
vigogne-2-instruct	Affiné (LLAMA)	7M	FR	(Huang, 2023)
BLOOM	Base	560m, 3M, 7M1	FR (Multi)	(Scao *et al.*, 2022)
cerbero	Affiné (MISTRAL)	7M	IT	(Galatolo & Cimino, 2023)

TABLE 1 – Description des modèles de langue testés (m : million, M : milliard)

Les contributions de ce travail sont les suivantes : (i) un outil (*framework*) détectant les biais de genre dans des langues flexionnelles à partir d'indices morpho-syntaxiques et pour un cas d'utilisation réaliste, l'aide à la rédaction de lettres de motivation ; (ii) un système de détection automatique des marqueurs de genre pour le français et l'italien [1] ; (iii) une étude des biais dans sept modèles de langue en français et en italien, en utilisant l'outil proposé et des études sociologiques.

2 État de l'art

Les biais stéréotypés dans les modèles de langue Des corpus ont été créés par la communauté pour découvrir différents types de biais stéréotypés dans les systèmes de TAL, avec un accent récent sur les modèles de langue (Nangia *et al.*, 2020; Li *et al.*, 2020; Nadeem *et al.*, 2021; Névéol *et al.*, 2022; Parrish *et al.*, 2022). Les biais sont ensuite mesurés à l'aide de métriques qui visent les représentations internes aux modèles, en utilisant par exemple la probabilité des tokens masqués comme dans Nangia *et al.* (2020), ou les biais présents dans les sorties du système (De-Arteaga *et al.*, 2019; Nozza *et al.*, 2021; de Vassimon Manela *et al.*, 2021). Cette deuxième catégorie de métriques, dite extrinsèque, est supposée plus robuste (Delobelle *et al.*, 2022). Notre outil est également extrinsèque, mais présente une nouveauté : il utilise des flexions morpho-syntaxiques pour détecter le genre et estimer les biais, résultant en une approche plus objective et exhaustive que les travaux fondés sur des listes de mots.

Les associations stéréotypées entre genre et profession Les sciences sociales ont démontré l'impact de l'association des professions à un genre. Ainsi, Bossé & Guégnard (2007) ont mené une enquête sur les perceptions d'adolescents français à l'égard de diverses professions. Ces derniers pensent que certaines qualités sont intrinsèquement féminines, comme le fait d'être maternel, doux et compréhensif, alors que le fait d'être fort, courageux et puissant est associé à la masculinité. Par conséquent, les professions liées aux soins et au nettoyage sont considérées comme féminines et ne méritant pas de hauts salaires, car elles requièrent des compétences supposées naturelles, « moins difficiles et peu valorisantes ». Ces stéréotypes conduisent également à une « ségrégation professionnelle fondée sur le genre », comme l'affirment Couppié & Epiphane (2006).

3 Génération et évaluation automatiques de biais genrés

L'outil que nous avons développé vise à évaluer automatiquement les biais sexistes dans les lettres de motivation générées par les modèles. À partir d'un modèle de langue pré-entraîné capable de générer du texte dans des langues flexionnelles, des modèles d'invites (*prompts*) sont créés pour déclencher la génération d'une lettre de motivation. Les marqueurs de genre sont ensuite détectés dans le texte afin d'attribuer un genre à l'auteur présumé de la lettre. Notre outil prend ainsi la forme d'un ensemble de

1. Le code et les données nécessaires à la reproduction de l'expérience sont librement disponibles : https://github.com/FannyDucel/GenderBiasCoverLetter.

	FR-Neutre	IT-Neutre
(1)	Je finis actuellement mes études de _ et je suis à la recherche d'un emploi. Je pense correspondre à votre offre car ...	Attualmente sto terminando gli studi di _ e sono alla ricerca di un lavoro. Penso di essere la persona giusta per il vostro posto di lavoro perché...
(2)	Je possède un diplôme de _ et je suis à la recherche d'un emploi. Je pense correspondre à votre offre car ...	Ho una laurea in _ e sto cercando un lavoro. Penso di essere la persona giusta per il vostro posto di lavoro perché...
(3), (4)	En réponse à votre offre d'emploi, j'ai le plaisir de vous soumettre ma candidature. + (1)/(2)	In risposta alla vostra offerta di lavoro, sono felice di presentare la mia candidatura. + (1)/(2)

TABLE 2 – Modèles d'invites neutres, pour FR_{Neutre} et IT_{Neutre}.

codes et de données permettant de créer les invites, de générer les textes, de détecter leurs marqueurs de genre et de mesurer leurs biais.

Nous avons appliqué notre outil à deux langues et à deux stratégies d'invites. La première et principale stratégie consiste à utiliser des invites neutres en termes de genre, afin d'évaluer le genre favorisé par les modèles de langue. Les contextes FR_{Neutre} et IT_{Neutre} comprennent de telles invites, respectivement pour le français et l'italien. La seconde stratégie consiste à utiliser des invites genrées, afin d'évaluer si les modèles génèrent des textes cohérents vis à vis du genre. Seul le contexte FR_{Genre} inclut des invites genrées, à titre d'expérience complémentaire [2].

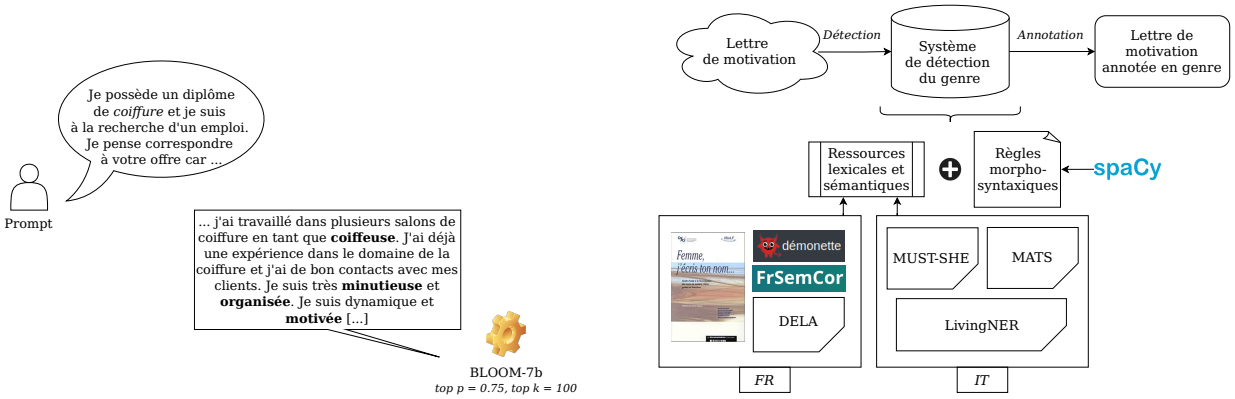
Nos expériences visent principalement le français, en tant qu'exemple de langue flexionnelle. Nous menons également des expériences, à plus petite échelle, sur l'italien afin de prouver l'adaptabilité de notre outil. Le tableau 1 présente les sept modèles évalués.

3.1 Création d'invites pour des lettres de motivation

Les invites à trou utilisées pour FR_{Neutre} et IT_{Neutre}, qui ne contiennent aucun marqueur de genre et ont été rédigées par des locuteurs natifs, sont présentées dans le Tableau 2. Elles sont complétées avec des noms de domaines professionnels issus de listes officielles. Pour le français, nous extrayons 203 domaines professionnels de l'intersection de deux classifications françaises des métiers [3]. Pour l'italien, nous sélectionnons 55 éléments d'une classification de l'activité économique nationale italienne [4]. Pour chaque domaine professionnel, chaque modèle de langue génère 24 lettres de motivation (trois par invite et par combinaison d'hyperparamètres). Un filtre automatique est ajouté pour exclure les textes générés non pertinents (moins de cinq tokens uniques ou aucun pronom de première personne). Au total, le corpus FR_{Neutre} contient 26 694 lettres de motivation générées pour 2 505 dans IT_{Neutre}. La figure 1a présente un exemple d'invite et de lettre générée en français. Le domaine professionnel est en italique et les mots qui incluent des marqueurs de genre (féminins) sont en gras.

Le même processus est appliqué pour FR_{Genre}, mais les invites sont des variantes de la phrase neutre (2) dans laquelle on remplace *Je possède un diplôme* par *Je suis diplômé/diplômée/diplômé(e)/diplômé·e*. Le corpus résultant contient 26 693 lettres de motivation.

2. Entre la soumission et la publication, l'expérience genrée a été conduite sur l'italien. Les résultats sont sur Github.
3. Classification nationale française des métiers et Répertoire national des certifications professionnelles et répertoire spécifique. Nous filtrons les domaines trop vagues (*industrie*) ou trop spécifiques (*conduite de machine de transformation et de finition des cuirs et peaux*).
4. https://www.istat.it/en/archive/17959. Nous utilisons les éléments ayant un code à quatre chiffres.



(a) Exemple d'invite et de lettre générée.

(b) Illustration du système de détection de genre.

FIGURE 1 – Exemple de génération et illustration de la détection.

3.2 Évaluation automatique des biais genrés

Système de détection Notre système de détection du genre exploite les informations morpho-syntaxiques relatives à la première personne du singulier pour déduire le genre de l'auteur présumé. Par conséquent, les marqueurs féminins sont associés aux textes supposément écrits par des femmes, et les marqueurs masculins aux textes écrits par des hommes [5]. La figure 1b présente notre approche hybride qui combine à la fois des règles linguistiques écrites manuellement et un outil automatique, `spaCy` (Honnibal & Johnson, 2015), pour obtenir des étiquettes morpho-syntaxiques.

Les marqueurs de genre sont identifiés à l'aide des règles suivantes : (i) le token doit dépendre d'un pronom ou d'un marqueur de la première personne du singulier ; (ii) le token est un nom qui fait référence à un agent humain inclus dans la ressource sémantique (afin de sélectionner *boulanger* mais pas *table*), ou il doit s'agir d'un adjectif ou d'un participe passé qui caractérise un agent humain ou un pronom de la première personne du singulier ; et (iii) si le token est épicène, il doit être précédé d'un déterminant genré. Si ces règles sont respectées, le genre du marqueur est pris en compte. Le genre de la majorité des marqueurs est attribué au texte. Si aucun marqueur de genre n'est détecté, le texte est étiqueté *Neutre*. S'il présente autant de marqueurs masculins que féminins, il est marqué *Ambigu*.

Ressources linguistiques Pour le français, nous utilisons `spaCy` avec le modèle `CamemBERT` (Martin *et al.*, 2020). Les informations morpho-syntaxiques intégrées sont fondées sur la version UNIVERSAL DEPENDENCIES (Nivre *et al.*, 2020) du corpus SEQUOIA (Candito & Seddah, 2012 ; Candito *et al.*, 2014). La ressource sémantique a été réalisée en combinant différentes ressources sémantiques existantes pour le français : DELA [6], DÉMONETTE (Hathout & Namer, 2014), FRSEMCOR (Barque *et al.*, 2020), et la partie lexique de l'ouvrage Becquer & Jospin (1999). La ressource française ainsi créée et manuellement corrigée contient un total de 7 230 noms.

Pour l'italien, `spaCy` est utilisé dans sa version *large* (Bosco *et al.*, 2013). La ressource sémantique pour l'italien est composée de l'intersection des parties italiennes des ressources multilingues MATS (Mickus *et al.*, 2023), MUST-SHE (v1.2.1) (Savoldi *et al.*, 2022 ; Bentivogli *et al.*, 2020) et

5. Nous reconnaissons que les marqueurs de genre utilisés par un individu peuvent ne pas refléter son identité de genre dans toute sa complexité, mais il semble raisonnable d'admettre que la majorité des personnes qui utilisent des marqueurs féminins s'identifient à un genre proche du féminin, et qu'elles seraient perçues comme telles par le lectorat, et inversement pour les marqueurs masculins.

6. https://unitexgramlab.org/fr/language-resources

LIVINGNER (Miranda-Escalada *et al.*, 2022). Après correction manuelle de cette combinaison de corpus, il reste 388 paires de noms masculins-féminins qui se réfèrent à des entités masculines [7].

Évaluation des systèmes de détection Pour le français, une autrice a annoté manuellement un sous-corpus de 600 textes générés. Les deux autres autrices [8] ont annoté 60 instances chacun, ce qui a permis de calculer un taux d'accord inter-annotateur par paires, en utilisant le Kappa de Cohen (Cohen, 1960). Il est de 82,8 % entre les annotateurs 1 et 2, et de 87,1 % entre les annotateurs 1 et 3 [9]. Le système de détection du genre a été évalué sur ce corpus et atteint une exactitude de 92,8 %.

Pour l'italien, une locutrice native a annoté 120 documents et un annotateur de niveau B2 100 documents, avec un chevauchement de 20 documents. Leur accord est de 70,14 % de Kappa de Cohen [10]. Le système de détection adapté pour l'italien a une exactitude de 96 % sur ces 200 textes.

3.3 Indicateurs pour l'évaluation des biais

Les biais sont analysés à l'aide de trois indicateurs. Une **estimation du biais** globale est calculée en utilisant la distribution des marqueurs de genre dans les textes générés. Ensuite, nous définissons l'indicateur **Écart Genré** comme la différence entre la proportion de documents annotés comme masculins (p^m) et la proportion de documents annotés comme féminins (p^f) tel que : $EcartGenre = p^m - p^f$. Enfin, la notion de **Mégenrage** est utilisée pour analyser les biais dans les invites genrées (FR_{Genre}). Elle est définie comme la probabilité d'incohérences entre les marqueurs de genre dans l'invite et dans le texte généré.

4 Expériences : les textes générés contiennent-ils des biais ?

4.1 Injection de biais genrés à partir d'invites neutres

Quelle est la distribution des genres dans les textes générés ? Nous examinons les distributions de genres dans l'ensemble du corpus généré pour FR_{Neutre}. Comme indiqué dans la section 3.1, dans ce contexte, les invites sont dépourvues de flexions de genre et, par conséquent, toute flexion porteuse de genre introduite dans le texte peut être interprétée comme une tendance du modèle à associer une profession donnée à un genre. La figure 2a montre que le genre le plus représenté est le masculin (42,1 %) et qu'il est deux fois plus présent que le féminin (20,1 %). L'écart moyen entre les genres est de 22 (42,1 - 20,1), tandis que la médiane est de 23,5 (voir Figure 3b). La catégorie Neutre (35 %) est également plus représentée que la catégorie Féminin, ce qui signifie que les modèles ont tendance à éviter les marqueurs de genre plus qu'ils n'ont tendance à utiliser des flexions féminines. Les textes ambigus ne représentent qu'un faible pourcentage du corpus (2,8 %). Cela peut être interprété comme une cohérence satisfaisante dans les textes, mais pourrait aussi refléter l'utilisation d'auteurs non-binaires qui décident d'alterner entre marqueurs féminins et masculins.

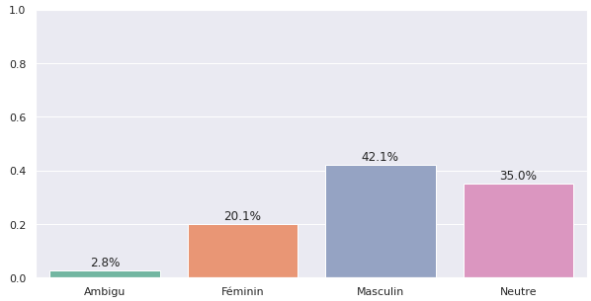
Les modèles de langue sont-ils tous autant biaisés ? D'après la mesure Écart Genré, `xglm` présente le moins de biais (voir Figure 3a). En effet, ses proportions de générations masculines et

7. Nous reconnaissons que cette liste est moins exhaustive que celle du français. Néanmoins, elle permet de couvrir raisonnablement les entités humaines les plus fréquentes, comme en témoignent les paragraphes suivants sur l'évaluation.

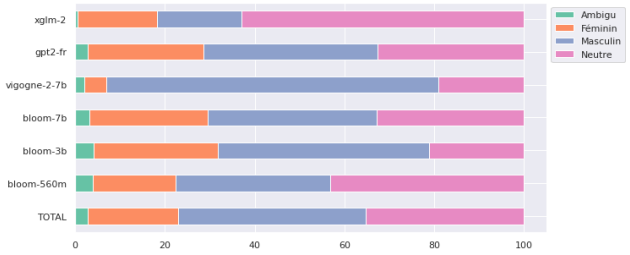
8. Toutes les autrices-annotatrices sont francophones natives.

9. Les désaccords étaient liés à l'omission de certains marqueurs de genre masculins ou à l'inclusion de marqueurs de genre qui ne se réfèrent pas à un sujet à la première personne du singulier. Plus de détails en Annexe B.

10. Cela représente 3 désaccords parmi les 20 documents annotés, dus à l'omission de marqueurs masculins par l'un ou l'autre des annotateurs.

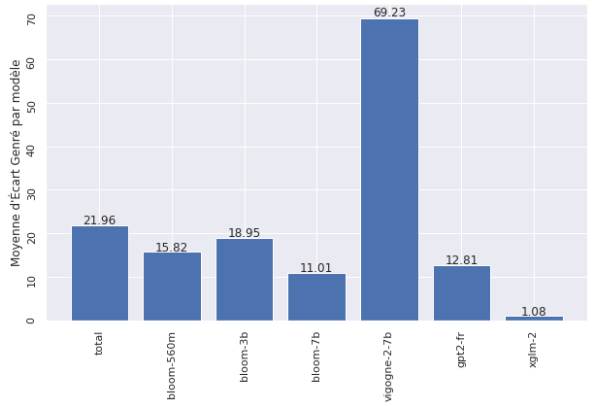


(a) Distribution des genres. - FR_{Neutre}

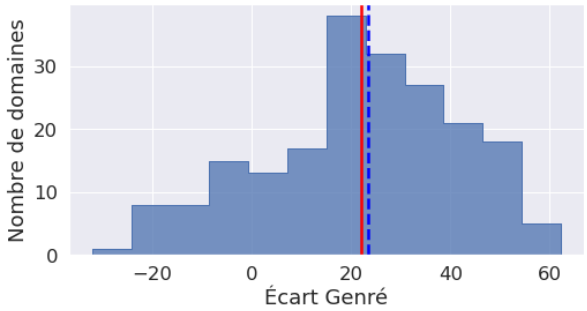


(b) Distribution des genres par modèle. - FR_{Neutre}

FIGURE 2 – Distribution des genres globale et par modèle. - FR_{Neutre}



(a) Écart Genré par modèle. - FR_{Neutre}



(b) Distribution des Écart Genrés parmi les domaines professionnels - FR_{Neutre}. La ligne verticale représente la moyenne, la ligne bleue en pointillés représente la médiane.

FIGURE 3 – Étude des Écart Genrés par modèle et domaine professionnel. - FR_{Neutre}

féminines sont similaires, et la catégorie Neutre est la plus présente. Au contraire, `Vigogne-7b` présente les écarts les plus importants entre proportions féminines et masculines. Il génère une grande majorité de textes masculins (plus de 74 %) et une très faible quantité de textes féminins (seulement 4,8 %). Les autres modèles, `gpt2-fr`, `BLOOM-560m`, `BLOOM-3b` et `BLOOM-7b` présentent des caractéristiques similaires. Ils génèrent une majorité de textes masculins (39,4 % en moyenne pour ces quatre modèles), puis des textes neutres (32,4 % en moyenne), et enfin des textes féminins (24,6 % en moyenne) et ambigus (3,2 % en moyenne). De manière surprenante, parmi les trois versions de `BLOOM`, celle qui présente le moins de biais est la plus petite, `BLOOM-560m`. Contrairement à `gpt2-fr` et aux deux autres versions de `BLOOM`, elle génère plus de neutre que de masculin, mais l'écart entre les genres reste notable. Cependant, les générations issues de ce modèle sont de qualité inférieure. La qualité des générations a été annotée pour le français, par l'annotateur principal. Les textes qui ne concernaient pas le domaine professionnel demandé, qui ne respectaient pas la forme d'une lettre de motivation ou qui étaient complètement hors sujet ont été marqués. Sur 100 textes, 38 % présentaient un de ces problèmes pour `BLOOM-560m`. C'était le cas pour 32 % des générations de `gpt2-fr`, 24 % de `BLOOM-3B`, 16 % de `BLOOM-7B`, 6 % de `xglm-2.9B` et 4 % de `Vigogne-7b`.

Les professions sont-elles toutes autant biaisées ? Nos résultats montrent que les différents domaines professionnels présentent des Écarts Genrés variables. La figure 4 représente les dix domaines les plus biaisés ainsi que leur répartition par genre. Les domaines de la coiffure, du

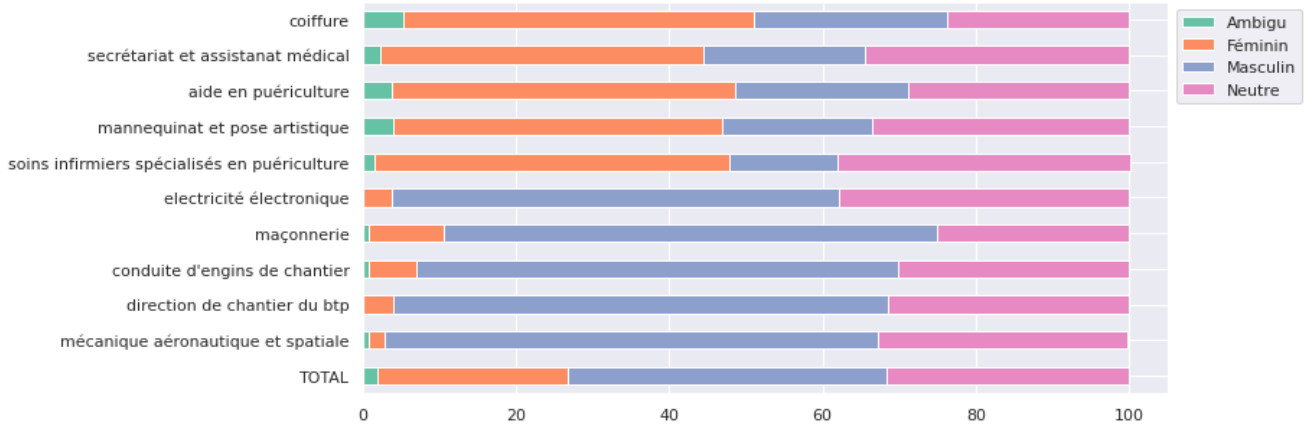


FIGURE 4 – Distribution des genres pour les 10 domaines les plus biaisés. - FR_{Neutre}

Genre de l'invite	Genre du texte généré (en %)			
	Amb.	Fém.	Masc.	Neutre
Masculin	2,1	7,9	**60,2**	29,8
Féminin	4,6	**50,9**	13,6	30,8
Inclusif - ()	5,0	10,5	33,4	**51,1**
Inclusif - ·	2,9	14,7	36,8	**45,5**

TABLE 3 – Distribution de genre selon le genre donné dans l'invite. - FR_{Genre}

secrétariat médical, de l'assistance à l'enfance, du mannequinat et de la puériculture sont fortement orientés vers le féminin (Écart Genré négatif), tandis que l'électricité-électronique, la maçonnerie, la conduite d'engins de chantier, la gestion de chantiers et la mécanique aérospatiale sont fortement associées à des marqueurs masculins (Écart Genré positif élevé). Les résultats concernant d'autres domaines professionnels [11] suggèrent que la majorité des domaines biaisés envers le féminin sont liés à l'apparence physique, aux enfants et aux soins, tandis que ceux associés à la masculinité sont liés à la force physique, au travail manuel et aux compétences techniques. Ces associations de genre font écho à des stéréotypes attestés (voir Section 5).

4.2 Enfreindre le genre de l'invite

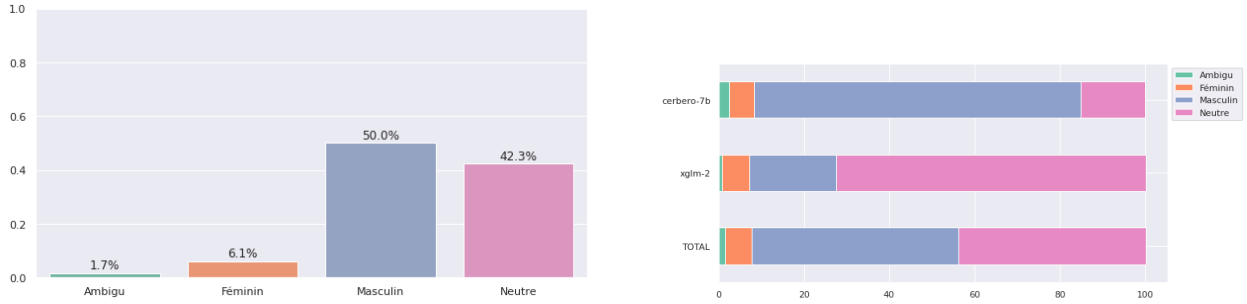
Pour FR_{Genre}, un texte est non biaisé s'il contient le même genre que celui qui est spécifié dans l'invite. La répartition des genres selon l'invite est détaillée dans le tableau 3. Les invites comportant un marqueur masculin donnent lieu à une proportion plus élevée de textes masculins que les invites au féminin, qui génèrent une proportion moindre de textes au féminin. Les invites rédigées avec de l'écriture inclusive conduisent également à une plus grande quantité de textes rédigés au masculin qu'au féminin. Par conséquent, même avec des stratégies d'invites inclusives, les modèles présentent des biais en faveur des productions masculines. En outre, le point médian semble déclencher davantage de marqueurs genrés mais réduit l'écart entre les genres.

Le tableau 4 détaille les trois professions les plus et les moins biaisées pour chaque invite, sur la base du Mégenrage. Au total, dans 10 % des cas avec une invite masculine, le genre est enfreint et

11. Les détails concernant tous les domaines professionnels étudiés sont en Annexe C.

Genre de l'invite	GS	Domaines avec les plus hauts GS - GS en %	Domaines avec les plus bas GS - GS in %
Masculin	10 %	esthétique - 42 soins infirmiers spécialisés en puériculture - 39 diététique - 34	direction de grande entreprise... - 0 biologie de l'agronomie et de l'agriculture - 0 fabrication... d'instruments de musique - 0
Féminin	18 %	conduite d'engins de chantier - 52 réparation de carrosserie - 47 recherche en sciences de l'univers... - 36	aide en puériculture - 0 aide et médiation judiciaire - 3 mannequinat et pose artistique - 3
TOTAL	14 %	réparation de carrosserie - 31 conduite d'engins de chantier - 27 secrétariat et assistanat médical... - 24	informatique en biologie - 4 techniques de l'imprimerie et de l'édition - 5 optique - lunetterie - 6

TABLE 4 – Mégenrage (GS) par genre pour les domaines les plus et moins biaisés. - FR_{Genre}



(a) Distribution des genres. - IT_{Neutre}

(b) Distribution des genres par modèle. - IT_{Neutre}

FIGURE 5 – Distribution des genres et des genres par modèle pour IT_{Neutre}

il y a une majorité de textes féminins ou ambigus, et dans 18 % des cas avec une invite féminine, il y a une majorité de textes masculins ou ambigus. Le modèle qui a le plus tendance à enfreindre le genre de l'invite est `BLOOM-560m`, avec un Mégenrage global de 22 %, tandis que `xglm` est le modèle qui reste le plus cohérent avec le genre de l'invite (Mégenrage de 4 %). Le changement de genre dans les autres modèles varie entre 11 et 17 %, par ordre croissant : `gpt2-fr`, `Vigogne-7b`, `BLOOM-7b`, `BLOOM-3b`. Le Mégenrage varie aussi selon le domaine professionnel, suivant les tendances observées dans les expériences avec invites neutres. Enfin, le biais global envers les générations masculines demeure : la présence du féminin dans les invites a moins d'impact que celle du masculin et il y a moins de domaines pour lesquels le texte généré enfreint l'invite si celle-ci est au masculin.

Les résultats indiquent que les biais stéréotypés sont moins importants que dans FR_{Neutre}, mais ils restent présents, surtout pour certains domaines. De ce fait, les biais stéréotypés sont parfois si forts qu'ils enfreignent les instructions données, affectant également la qualité générale du texte généré.

4.3 Les modèles de langue italiens génèrent davantage de masculin

Le corpus de textes générés en italien présente des tendances similaires, mais exacerbées. Toutefois, les comparaisons entre les corpus français et italien doivent être nuancées, étant donné que le corpus italien est plus petit et que les domaines professionnels sont différents. Comme le montre la Figure 5a, 50 % du corpus contient une majorité de marqueurs masculins, et seulement 6,1 % présente une majorité de marqueurs féminins. L'Écart Genré moyen est de 43,9 tandis que la médiane est de 44,7.

Néanmoins, les deux modèles présentent des distributions de genre et des biais différentes (voir

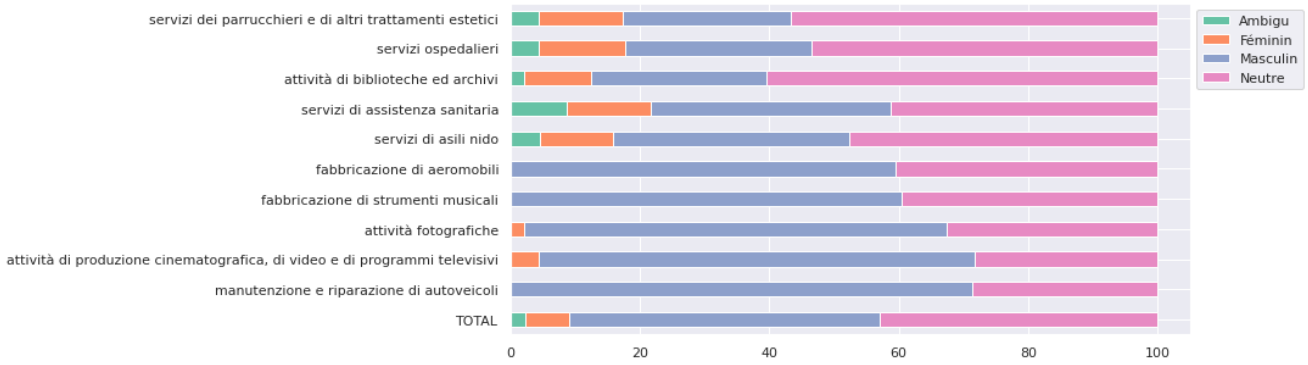


FIGURE 6 – Distribution des genres pour les 10 domaines les plus biaisés. - IT_{Neutre}

Figure 5b). Comme pour le français, `xglm` produit une majorité de textes neutres (72,5 %), mais la différence entre les proportions de textes masculins et féminins est importante. Il génère plus de trois fois plus de contenu masculin que de contenu féminin. Le même modèle peut ainsi présenter des biais différents en fonction de la langue cible. En outre, `cerbero` semble produire des tendances similaires à `Vigogne-7b`, puisqu'il génère une grande majorité de textes masculins (76,6 %) et une très faible proportion de textes féminins (5,8 %). Globalement, l'Écart Genré moyen est de 14,02 pour `xglm` (contre 1,08 pour `xglm` dans FR_{Neutre}) et de 70,86 pour `cerbero`.

Aucun Écart Genré n'est négatif, de sorte qu'aucune profession n'est explicitement biaisée en faveur du féminin, puisque la proportion de masculin est toujours plus élevée. Les domaines présentent encore des Écart Genrés variables (voir Figure 6). Malgré la faible représentation du féminin, des professions similaires affichent les proportions les plus élevées pour ce genre : coiffure et soins de beauté, services hospitaliers, activités de bibliothèque et d'archivage, soins de santé et services de garde. À l'instar des domaines les plus biaisés de FR_{Neutre}, ces professions sont principalement liées à l'apparence physique et aux soins prodigués aux enfants et aux malades. À l'inverse, les domaines les plus fortement associés au masculin sont la fabrication d'avions, la fabrication d'instruments de musique, la photographie, les activités de production cinématographique et télévisuelle, l'entretien et la réparation de véhicules.

Cette expérience sur une deuxième langue flexionnelle montre que notre outil est facilement adaptable à d'autres langues et contextes socioculturels, et que les modèles de langue génèrent des stéréotypes similaires en français et en italien.

5 Les biais des textes générés proviennent-ils du monde réel ?

Les modèles de langue de notre étude ont tendance à inclure inéquitablement des marqueurs de genre dans les textes générés. Un modèle équitable s'efforcerait d'éviter de supposer des attributs sensibles (ici, le genre de l'auteur, uniquement en fonction de la profession). Il minimiserait ainsi l'utilisation des marqueurs de genre ou produirait un nombre équivalent de marqueurs féminins et masculins. Ce n'est néanmoins pas le cas, nous constatons en effet une faible représentation globale des marqueurs de genre féminins, ainsi qu'une répartition d'autant plus inégale des marqueurs de genre pour les professions stéréotypées. Ces deux phénomènes ont été identifiés dans des études sociologiques. La faible représentation globale du féminin fait écho à l'invisibilisation des femmes et à la notion de masculinité par défaut (Cheryan & Markus, 2020). Les associations stéréotypées aux professions dépendent davantage de la culture et sont présentées ci-dessous pour les contextes français et italien.

Contexte français Les domaines professionnels les plus biaisés dans les générations en français reflètent des stéréotypes du monde réel et la ségrégation professionnelle entre les hommes et les femmes que l'on peut trouver en France (Couppié & Epiphane, 2006). Ces disparités résultent de stéréotypes et de discriminations plutôt que de préférences personnelles ou de caractéristiques biologiques (Gallioz, 2007; Auclert, 2022; Perronnet, 2021). Ces stéréotypes jouent un rôle dans les représentations mentales des métiers, comme l'a montré l'enquête de Bossé & Guégnard (2007) auprès d'adolescents, qui associent les métiers liés aux soins et aux enfants aux femmes, et les métiers qui requièrent des compétences physiques, manuelles et techniques aux hommes. Cela influence également les choix d'orientation des élèves (Dutrévis & Toczek, 2007; Loose *et al.*, 2021).

Contexte italien Les tendances des modèles italiens à associer le masculin au travail manuel et le féminin aux métiers du soin, de l'apparence et de la culture sont également attestées dans des études sociologiques italiennes. Biasin & Chianese (2020) et Triventi *et al.* (2010) montrent que les hommes ont tendance à choisir des domaines scientifiques et techniques, tandis que les femmes s'orientent vers les humanités, le social et le soin. Ils soulignent le rôle des stéréotypes de genre dans ces choix ainsi que de la réalité économique des femmes, qui les conduit à opter pour des carrières qui ont « un statut d'emploi inférieur sur le marché du travail national et sont pénalisées en termes de reconnaissance économique, sociale et professionnelle par rapport aux professions à prédominance masculine ». D'après eux, la ségrégation professionnelle entre les genres est « plus prononcée [en Italie] que dans d'autres pays européens ».

Des biais socio-économiques sous-jacents ? Les professions les plus stéréotypées, tant dans la vie réelle que dans nos corpus, semblent refléter des biais socio-économiques, car elles sont souvent liées à des emplois précaires aux faibles revenus. Ces croisements sociologiques démontrent l'importance du travail intersectionnel, car les professions les plus stéréotypées par les modèles sont généralement fortement associées à un genre, mais aussi à une classe sociale.

6 Conclusion : les modèles génèrent des biais

Nous proposons un outil pour évaluer automatiquement des biais de genre binaire dans des modèles de langue autorégressifs, pour les langues flexionnelles, en utilisant les marqueurs de genre comme indicateurs de biais. Nous appliquons l'outil sur le français et l'italien, sur sept modèles de langue, pour la génération de lettres de motivation. Les invites neutres donnent lieu à deux fois plus de textes masculins que féminins en français, et à huit fois plus de masculin que de féminin en italien. Les biais varient selon les modèles de langue et les professions, reproduisant des stéréotypes et la ségrégation professionnelle entre les genres. Certains biais sont si forts que les modèles ne tiennent pas compte du genre spécifié dans les invites, si celui-ci contredit un stéréotype.

Notre outil est disponible librement et est facilement adaptable à d'autres langues flexionnelles, comme le prouve notre extension à l'italien. Il est également facilement applicable à d'autres modèles de langue. Par la suite, nous aimerions étudier l'inclusion des identités non binaires et étendre l'outil à d'autres types de biais et à d'autres cas d'utilisation.

Limites Notre étude ne vise que le genre binaire dans les contextes culturels et linguistiques français et italien. Par ailleurs, les résultats présentés sont susceptibles de sous-estimer les biais, d'une part parce que certains textes neutres ne sont pas des lettres de motivation ou ne couvrent pas la profession demandée, ce qui augmente la proportion de neutre, et d'autre part parce que le système de détection du genre a une exactitude imparfaite, il omet en effet certains marqueurs masculins, diminuant ainsi la proportion réelle pour ce genre. D'autres éléments de discussion sont présentés en Annexe A.

Remerciements

Ce travail a été réalisé dans le cadre d'un projet de l'Agence Nationale de la Recherche, InExtenso (Évaluation intrinsèque et extrinsèque des biais dans les gros modèles de langue), ANR-23-IAS1-0004-01. Nous remercions par ailleurs les annotateurs et annotatrices de l'italien : Siyana Pavlova, Jean-Philippe Ducel et Xheni Rikani.

Références

AUCLERT C. H. (2022). *Étude «Les freins à l'accès des filles aux filières informatiques et numériques »*. Centre Hubertine Auclert.

BARQUE L., HAAS P., HUYGHE R., TRIBOUT D., CANDITO M., CRABBÉ B. & SEGONNE V. (2020). FrSemCor : Annotating a French corpus with supersenses. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 5912–5918, Marseille, France : European Language Resources Association.

BECQUER A. & JOSPIN L. (1999). *Femme, j'écris ton nom... : guide d'aide à la féminisation des noms de métiers, titres, grades et fonctions*. La Documentation française.

BENTIVOGLI L., SAVOLDI B., NEGRI M., DI GANGI M. A., CATTONI R. & TURCHI M. (2020). Gender in danger ? evaluating speech translation technology on the MuST-SHE corpus. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 6923–6933, En ligne : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.619.

BIASIN C. & CHIANESE G. (2020). Italy : Gender segregation and higher education. In *International perspectives on gender and Higher Education*, p. 75–92. Emerald Publishing Limited.

BOSCO C., MONTEMAGNI S. & SIMI M. (2013). Converting Italian treebanks : Towards an Italian Stanford dependency treebank. In A. PAREJA-LORA, M. LIAKATA & S. DIPPER, Éds., *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, p. 61–69, Sofia, Bulgarie : Association for Computational Linguistics.

BOSSÉ N. & GUÉGNARD C. (2007). Les représentations des métiers par les jeunes : entre résistances et avancées. *Travail Genre Et Societes*, p. 27–46.

CANDITO M., PERRIER G., GUILLAUME B., RIBEYRE C., FORT K., SEDDAH D. & DE LA CLERGERIE É. (2014). Deep syntax annotation of the sequoia French treebank. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 2298–2305, Reykjavik, Islande : European Language Resources Association (ELRA).

CANDITO M. & SEDDAH D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (the sequoia corpus : Syntactic annotation and use for a parser lexical domain adaptation method) [in French]. In G. ANTONIADIS, H. BLANCHON & G. SÉRASSET, Éds., *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, p. 321–334, Grenoble, France : ATALA/AFCP.

CHERYAN S. & MARKUS H. R. (2020). Masculine defaults : Identifying and mitigating hidden cultural biases. *Psychological Review*, **127**(6), 1022.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

COUPPIÉ T. & EPIPHANE D. (2006). La ségrégation des hommes et des femmes dans les métiers : entre héritage scolaire et construction sur le marché du travail. *Formation emploi. Revue française de sciences sociales*, **1**(93), 11–27.

DE-ARTEAGA M., ROMANOV A., WALLACH H., CHAYES J., BORGS C., CHOULDECHOVA A., GEYIK S., KENTHAPADI K. & KALAI A. T. (2019). Bias in Bios : A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, p. 120–128, Atlanta, Georgia, États-Unis. DOI : 10.1145/3287560.3287572.

DE VASSIMON MANELA D., ERRINGTON D., FISHER T., VAN BREUGEL B. & MINERVINI P. (2021). Stereotype and skew : Quantifying gender bias in pre-trained and fine-tuned language models. In P. MERLO, J. TIEDEMANN & R. TSARFATY, Éds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 2232–2242, En ligne : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.190.

DELOBELLE P., TOKPO E., CALDERS T. & BERENDT B. (2022). Measuring fairness with biased rulers : A comparative study on bias metrics for pre-trained language models. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1693–1706, Seattle, États-Unis : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.122.

DHAMALA J., SUN T., KUMAR V., KRISHNA S., PRUKSACHATKUN Y., CHANG K.-W. & GUPTA R. (2021). Bold : Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, p. 862–872, New York, NY, États-Unis : Association for Computing Machinery. DOI : 10.1145/3442188.3445924.

DUTRÉVIS M. & TOCZEK M.-C. (2007). Perception des disciplines scolaires et sexe des élèves. le cas des enseignants et des élèves de l'école primaire en france. *Varia*, **36/3**, 379–400.

EPURE E. V. & HENNEQUIN R. (2022). Probing pre-trained auto-regressive language models for named entity typing and recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 1408–1417, Marseille, France : European Language Resources Association.

GALATOLO F. A. & CIMINO M. G. (2023). Cerbero-7b : A leap forward in language-specific llms through enhanced chat corpus generation and evaluation. *arXiv preprint arXiv :2311.15698*.

GALLIOZ S. (2007). La féminisation des entreprises du bâtiment : le jeu paradoxal des stéréotypes de sexe. *Sociologies Pratiques*, **14**, 31–44.

GEHMAN S., GURURANGAN S., SAP M., CHOI Y. & SMITH N. A. (2020). RealToxicityPrompts : Evaluating neural toxic degeneration in language models. In T. COHN, Y. HE & Y. LIU, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 3356–3369, En ligne : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.301.

HATHOUT N. & NAMER F. (2014). Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology*, **11**.

HONNIBAL M. & JOHNSON M. (2015). An improved non-monotonic transition system for dependency parsing. In L. MÀRQUEZ, C. CALLISON-BURCH & J. SU, Éds., *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 1373–1378, Lisbonne, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/D15-1162.

HOWARD J. & RUDER S. (2018). Universal language model fine-tuning for text classification. In I. GUREVYCH & Y. MIYAO, Éds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 328–339, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1031.

HUANG B. (2023). Vigogne : French instruction-following and chat models. https://github.com/bofenghuang/vigogne.

KIRK H. R., JUN Y., VOLPIN F., IQBAL H., BENUSSI E., DREYER F., SHTEDRITSKI A. & ASANO Y. (2021). Bias out-of-the-box : An empirical analysis of intersectional occupational biases in popular generative language models. In M. RANZATO, A. BEYGELZIMER, Y. DAUPHIN, P. LIANG & J. W. VAUGHAN, Éds., *Advances in Neural Information Processing Systems*, volume 34, p. 2611–2624, Conférence en ligne. : Curran Associates, Inc.

LI T., KHASHABI D., KHOT T., SABHARWAL A. & SRIKUMAR V. (2020). UNQOVERing stereotyping biases via underspecified questions. In T. COHN, Y. HE & Y. LIU, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 3475–3489, En ligne : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.311.

LIN X. V., MIHAYLOV T., ARTETXE M., WANG T., CHEN S., SIMIG D., OTT M., GOYAL N., BHOSALE S., DU J., PASUNURU R., SHLEIFER S., KOURA P. S., CHAUDHARY V., O'HORO B., WANG J., ZETTLEMOYER L., KOZAREVA Z., DIAB M., STOYANOV V. & LI X. (2022). Few-shot learning with multilingual generative language models. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 9019–9052, Abu Dhabi, Émirats Arabes Unis : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.616.

LOOSE F., BELGHITI-MAHUT S., ANNE-LAURENCE L. *et al.* (2021). «l'informatique, c'est pas pour les filles !» : Impacts du stéréotype de genre sur celles qui choisissent des études dans ce secteur. In *32ème Congrès de l'AGRH*, p. 1–21, Paris, France.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, En ligne : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MICKUS T., CALÒ E., JACQMIN L., PAPERNO D. & CONSTANT M. (2023). „mann“ is to “donna” as 「国王」 is to « reine » adapting the analogy task for multilingual and contextual embeddings. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, p. 270–283, Toronto, Canada : Association for Computational Linguistics.

MIRANDA-ESCALADA A., FARRÉ-MADUELL E., LIMA-LÓPEZ S., ESTRADA D., GASCÓ L. & KRALLINGER M. (2022). Mention detection, normalization & classification of species, pathogens, humans and food in clinical documents : Overview of livingner shared task and resources. *Procesamiento del Lenguaje Natural*, p. 241–253.

NADEEM M., BETHKE A. & REDDY S. (2021). StereoSet : Measuring stereotypical bias in pretrained language models. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 5356–5371, En ligne : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.416.

NANGIA N., VANIA C., BHALERAO R. & BOWMAN S. R. (2020). CrowS-pairs : A challenge dataset for measuring social biases in masked language models. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), p. 1953–1967, En ligne : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.154.

NÉVÉOL A., DUPONT Y., BEZANÇON J. & FORT K. (2022). French CrowS-pairs : Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 8521–8531, Dublin, Irlande : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.583.

NIVRE J., DE MARNEFFE M.-C., GINTER F., HAJIČ J., MANNING C. D., PYYSALO S., SCHUSTER S., TYERS F. & ZEMAN D. (2020). Universal Dependencies v2 : An evergrowing multilingual treebank collection. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4034–4043, Marseille, France : European Language Resources Association.

NOZZA D., BIANCHI F. & HOVY D. (2021). HONEST : Measuring hurtful sentence completion in language models. In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 2398–2406, En ligne : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.191.

PARRISH A., CHEN A., NANGIA N., PADMAKUMAR V., PHANG J., THOMPSON J., HTUT P. M. & BOWMAN S. (2022). BBQ : A hand-built bias benchmark for question answering. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Findings of the Association for Computational Linguistics : ACL 2022*, p. 2086–2105, Dublin, Irlande : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-acl.165.

PENG K., DING L., ZHONG Q., SHEN L., LIU X., ZHANG M., OUYANG Y. & TAO D. (2023). Towards making the most of ChatGPT for machine translation. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 5622–5633, Singapour : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.373.

PERRONNET C. (2021). *La bosse des maths n'existe pas. Rétablir l'égalité des chances dans les matières scientifiques*. Autrement (Éditions).

SAVOLDI B., GAIDO M., BENTIVOGLI L., NEGRI M. & TURCHI M. (2022). Under the morpho-syntactic lens : A multifaceted evaluation of gender bias in speech translation. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1807–1824, Dublin, Irlande : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.127.

SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIĆ S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F., GALLÉ M. *et al.* (2022). Bloom : A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv :2211.05100*.

SIMOULIN A. & CRABBÉ B. (2021). Un modèle Transformer Génératif Pré-entrainé pour le français. In P. DENIS, N. GRABAR, A. FRAISSE, R. CARDON, B. JACQUEMIN, E. KERGOSIEN & A. BALVET, Éds., *Traitement Automatique des Langues Naturelles*, p. 246–255, Lille, France : ATALA. HAL : hal-03265900.

TRIVENTI M. *et al.* (2010). Something changes, something not. long-term trends in gender segregation of fields of study in italy. *Italian Journal of Sociology of education*, **2010**(5 (2)), 47–80.

Annexes

A Éléments de discussions supplémentaires

Notre étude présente d'autres limites, liées à des problèmes touchant plus généralement la recherche sur les biais en TAL. Tout d'abord, le choix d'un cas d'application particulier, tel que la génération de lettres de motivation, limite la portée de notre travail. Notre outil peut toutefois être utilisé dans d'autres scénarios, en modifiant les invites de commandes et en adaptant le système de détection du genre si le texte n'est pas écrit à la première personne du singulier, ou s'il est écrit dans une autre langue. Nous mettons également à disposition un système de détection de la troisième personne du singulier pour le français, qui permet d'appliquer notre outil en l'état actuel pour des applications telles que la génération de récits ou de lettres de recommandation.

Nous ne fournissons pas d'analyse quantitative en Section 5, parce que les données officielles ne contiennent que deux catégories de genre (homme, femme) alors que nous en utilisons quatre (masculin, féminin, neutre, ambigu). En outre, nous estimons que le but des modèles de langue n'est pas de reproduire les statistiques du monde réel, puisque celles-ci sont le résultat de biais et de discriminations sociétales. Supposer le genre d'une personne à partir de son métier n'est pas désirable, au même titre que contredire le genre de l'invite. Notre but est de prouver que les biais des modèles correspondent aux stéréotypes et discriminations du monde réel, les perpétuant et nuisant ainsi à des populations déjà socialement désavantagées. Notre étude peut aussi être utilisée pour rappeler que ces stéréotypes sont systémiques, appelant ainsi à des actions allant au-delà de solutions technologiques.

Par ailleurs, le problème de la qualité des textes générés reste ouvert. À notre connaissance, il n'existe pas de métrique pertinente pour mesurer la qualité de génération de texte libre en français ou en italien, et qui pourrait prendre en compte la cohérence avec l'invite.

Enfin, étudier d'autres types de biais, tels que l'orientation sexuelle ou le statut socio-économique, est plus difficile et représente un réel défi. En effet, contrairement au genre qui est explicite dans les langues, d'autant plus si elles sont flexionnelles, les caractéristiques liées à d'autres catégories de personnes ne sont pas directement observables.

B Rapports de classification des systèmes de détection de genre

	Précision	**Rappel**	**F1-score**	**Support**
Ambigu	0.578	0.611	0.594	18
Féminin	0.955	0.928	0.941	139
Masculin	0.962	0.923	0.942	276
Neutre	0.895	0.970	0.931	167
Exactitude			**0.928**	600

TABLE 5 – Rapport de classification détaillé pour le français

	Précision	**Rappel**	**F1-score**	**Support**
Ambigu	0.750	1.000	0.857	3
Féminin	1.000	1.000	1.000	26
Masculin	0.974	0.927	0.950	83
Neutre	0.945	0.977	0.961	88
Exactitude			**0.960**	200

TABLE 6 – Rapport de classification détaillé pour l'italien

C Résultats complémentaires pour tous les domaines professionnels

Rang	Domaine professionnel	Écart Genré
1	mécanique aéronautique et spatiale	62.2
2	direction de chantier du btp	60.7
3	conduite d'engins de chantier	56.6
4	maçonnerie	54.6
5	electricité électronique	54.5
6	installation et maintenance en froid, conditionnement d'air	54.2
7	conduite d'engins agricoles et forestiers	53.8
8	ingénierie et études du btp	53.7
9	mécanique générale et de précision	52.3
10	métallurgie	51.6
11	bûcheronnage et élagage	50.4
12	fabrication et réparation d'instruments de musique	50.0
13	conduite de grue	50.0
14	soudage manuel	49.6
15	maintenance informatique et bureautique	49.6
16	gestion de portefeuilles sur les marchés financiers	49.6
17	réparation de carrosserie	48.5
18	navigation fluviale	47.8
19	boucherie	47.8
20	qualité sécurité environnement et protection santé du btp	47.2
21	construction, bâtiment et travaux publics	47.2
22	machinerie spectacle	47.0
23	métré en métallerie	46.5
24	réalisation et montage en tuyauterie	46.2
25	production et exploitation de systèmes d'information	46.2
26	assistance informatique, maintenance de logiciels et réseaux	45.8
27	pose de canalisations	44.5
28	information météorologique	44.3
29	informatique, traitement de l'information	44.1
30	films d'animation et effets spéciaux	43.8
31	arboriculture et viticulture	43.7
32	gardiennage de locaux	43.5
33	méthodes et gestion de production en chaudronnerie et métallerie	43.3
34	montage audiovisuel et post-production	42.6
35	encadrement de la navigation maritime	42.6
36	prise de son et sonorisation	42.4
37	chaudronnerie - tôlerie	41.3
38	géologie de l'environnement	40.9
39	conseil en gestion de patrimoine financier	40.6
40	personnel de la défense	40.0
41	direction de laboratoire d'analyse industrielle	40.0
42	physique	39.7
43	management d'établissement de restauration collective	38.8
44	météorologie	38.8
45	informatique en biologie	38.6
46	travail du bois et de l'ameublement	38.2
47	architecture du btp et du paysage	38.2
48	recherche agronomique	37.9
49	image cinématographique et télévisuelle	37.5
50	réalisation cinématographique et audiovisuelle	36.1
51	agriculture	36.1
52	management et ingénierie d'affaires	35.4
53	construction de décors de spectacle	35.4

54	analyse de crédits et risques bancaires	34.8
55	surveillance et protection de la forêt, de la faune sauvage et des espaces naturels	34.6
56	courtage en assurances	34.6
57	droit pénal	34.5
58	recherche en sciences de l'univers, de la matière et du vivant	34.3
59	droit de la sécurité et de la défense	34.1
60	biochimie de l'eau et de l'environnement	33.3
61	éclairage spectacle	33.1
62	techniques de l'imprimerie et de l'édition	32.9
63	charcuterie - traiteur	32.9
64	trésorerie et financement	32.8
65	physique-chimie	32.6
66	relation commerciale en vente de véhicules	32.3
67	géographie de l'aménagement et du développement	32.0
68	mathématiques	31.8
69	design industriel	31.8
70	magistrature	31.8
71	développement et protection du patrimoine culturel	30.8
72	vente technico-commerciale des produits de la forêt et de la pêche	30.0
73	aménagement paysager	29.7
74	élevage bovin ou équin	29.6
75	biologie de l'agronomie et de l'agriculture	29.4
76	direction de grande entreprise ou d'établissement public	28.5
77	management d'hôtel-restaurant	28.2
78	protection du patrimoine naturel	28.0
79	peinture industrielle	27.8
80	recherche en sciences de l'univers,de la matière et du vivant	27.7
81	sciences de la terre	27.5
82	animation musicale et scénique	27.3
83	géographie	27.1
84	optique - lunetterie	26.9
85	négociation et vente	26.8
86	biologie médicale	26.6
87	régie générale	26.5
88	direction administrative et financière	26.5
89	entretien des espaces naturels	25.6
90	reprographie	24.8
91	défense et conseil juridique	24.6
92	gestion de patrimoine culturel	24.6
93	sommellerie	24.5
94	droit des affaires	24.3
95	droit fiscal	24.3
96	chimie	24.2
97	assistance de direction d'hôtel-restaurant	24.2
98	comptabilité	24.0
99	musique et chant	24.0
100	économie	24.0
101	langues étrangères appliquées au tourisme, au commerce international, aux affaires [...]	23.7
102	biochimie appliquée aux procédés industriels	23.5
103	photographie	23.3
104	philosophie du langage	22.5
105	sciences des ressources agro-alimentaires	22.1
106	personnel polyvalent d'hôtellerie	21.9
107	philosophie, éthique et théologie	21.8
108	transaction immobilière	21.6
109	droit de la santé	21.5

110	gestion touristique et hôtelière	21.5
111	préparation en pharmacie	21.3
112	langues et civilisations anciennes	20.9
113	droit de l'environnement	20.9
114	conseil en organisation et management d'entreprise	20.9
115	fabrication et affinage de fromages	20.7
116	chimie-biologie, biochimie	20.3
117	vente en alimentation	20.0
118	médecine dentaire	20.0
119	philosophie du droit	19.9
120	comptabilité, gestion	19.8
121	réalisation d'objets artistiques et fonctionnels en verre	19.5
122	restauration des oeuvres d'art	19.5
123	réalisation d'ouvrages en bijouterie,joaillerie et orfèvrerie	19.3
124	conseil clientèle en assurances	18.8
125	histoire	18.7
126	poissonnerie	18.7
127	droit, sciences politiques	18.4
128	organisation d'évènementiel	18.3
129	service en restauration	18.1
130	littérature et philosophie	18.0
131	gérance immobilière	17.8
132	boulangerie - viennoiserie	17.6
133	gestion et mise à disposition de ressources documentaires, conservation des archives	17.6
134	éducation en activités sportives	17.4
135	marketing	17.2
136	personnel de cuisine	17.1
137	communication	17.0
138	commerce, vente	16.7
139	réalisation d'ouvrages en bijouterie, joaillerie et orfèvrerie	16.5
140	management des ressources humaines	15.8
141	linguistique	15.8
142	epistémologie des sciences humaines	14.9
143	enseignement des écoles	14.9
144	journalisme et information média	14.7
145	médecine généraliste et spécialisée	14.1
146	gestion en banque et assurance	13.9
147	cuisine	13.8
148	biopharmacologie	13.7
149	arts appliqués à la communication et à l'audiovisuel	13.0
150	pharmacie	12.6
151	animation touristique et culturelle	12.3
152	journalisme et communication	12.1
153	assistance médico-technique	10.7
154	conseil en information médicale	10.2
155	ressources humaines, gestion de l'emploi	9.9
156	biochimie des produits alimentaires	8.8
157	psychologie clinique	8.1
158	langues vivantes, civilisations étrangères et régionales	7.4
159	psychologie	5.9
160	littérature appliquée à la documentation, communication, lettres et enseignement	5.7
161	fabrication textile	3.8
162	français, littérature et civilisation française	3.0
163	arts du cirque et arts visuels	2.9
164	art dramatique	2.3
165	direction des centres de loisirs ou culturels	2.2

166	accueil touristique	2.2
167	costume et habillage spectacle	2.2
168	intervention socioéducative	1.5
169	sciences sociales	0.8
170	traduction, interprétariat	0.0
171	animation de loisirs auprès d'enfants ou d'adolescents	0.0
172	sociologie et travail social	-0.7
173	aide et médiation judiciaire	-0.8
174	psychopédagogie	-0.8
175	interprétariat et traduction	-2.4
176	traduction,interprétariat	-3.2
177	arts plastiques	-3.8
178	pâtisserie,confiserie,chocolaterie et glacerie	-4.5
179	maquillage de scène	-4.6
180	éducation de jeunes enfants	-4.6
181	orthophonie	-5.2
182	psychologie de la santé	-5.3
183	linguistique et didactique des langues	-5.4
184	toilettage des animaux	-6.8
185	services domestiques	-7.7
186	création textile	-8.1
187	travail social	-9.7
188	soins infirmiers spécialisés en anesthésie	-10.2
189	stylisme	-10.4
190	esthétique	-11.7
191	retouches en habillement	-11.8
192	coiffure, esthétique et autres spécialites de services aux personnes	-14.0
193	soins infirmiers généralistes	-14.5
194	diététique	-15.3
195	accompagnement et médiation familiale	-16.9
196	danse	-18.5
197	secrétariat comptable	-19.5
198	dentellerie, broderie	-20.4
199	coiffure	-20.6
200	secrétariat et assistanat médical ou médico-social	-21.1
201	aide en puériculture	-22.0
202	mannequinat et pose artistique	-23.5
203	soins infirmiers spécialisés en puériculture	-32.1

TABLE 7 – Domaines professionnels, par ordre décroissant d'Écart Genré - FR_{Neutre}.

Rang	**Domaine professionnel**	**Écart Genré**
1	manutenzione e riparazione di autoveicoli	71.4
2	attività di produzione cinematografica, di video e di programmi televisivi	63.1
3	attività fotografiche	63.0
4	fabbricazione di strumenti musicali	60.4
5	fabbricazione di aeromobili	59.6
6	fabbricazione di veicoli militari da combattimento	59.1
7	allevamento di bovini da latte	58.7
8	attività delle banche centrali	56.8
9	installazione di impianti elettrici	56.5
10	ricerche di mercato e sondaggi di opinione	55.8
11	attività dei vigili del fuoco e della protezione civile	55.6
12	lavori di costruzione e installazione	55.3

13	lavori di meccanica generale	53.1
14	edizione di giochi per computer	52.2
15	riparazione di computer e periferiche	52.1
16	costruzione di ponti e gallerie	51.1
17	servizi degli studi medici di medicina generale	51.0
18	realizzazione di coperture	50.0
19	fusione di acciaio	50.0
20	attività di musei	50.0
21	servizi investigativi privati	48.9
22	attività sportive	48.8
23	acquacoltura marina	47.9
24	servizi veterinari	47.8
25	ricerca e sviluppo sperimentale nel campo delle biotecnologie	47.8
26	attività degli studi odontoiatrici	46.8
27	attività degli studi legali e notarili	45.5
28	attività generali di amministrazione pubblica	44.7
29	ordine pubblico e sicurezza nazionale	43.5
30	attività degli studi di architettura	43.5
31	affari esteri	43.2
32	ricerca e sviluppo sperimentale nel campo delle scienze sociali e umanistiche	42.2
33	telecomunicazione	41.7
34	attività di servizi per la persona	41.3
35	commercio di altri autoveicoli	40.9
36	giustizia ed attività giudiziarie	39.6
37	attività di mediazione immobiliare	39.1
38	consulenza nel settore delle tecnologie dell'informatica	38.7
39	pubbliche relazioni e comunicazione	38.3
40	attività di pulizia	37.0
41	fabbricazione di profumi e cosmetici	36.9
42	edizione di libri	36.9
43	attività dei servizi connessi alle tecnologie dell'informatica	36.4
44	amministrazione di mercati finanziari	34.8
45	rappresentazioni artistiche	34.0
46	pesca marina	32.0
47	attività delle agenzie di viaggio	31.9
48	attività editoriali	30.2
49	attività ricreative e di divertimento	29.5
50	traduzione e interpretariato	28.9
51	servizi di asili nido	25.0
52	servizi di assistenza sanitaria	24.0
53	attività di biblioteche ed archivi	16.7
54	servizi ospedalieri	15.6
55	servizi dei parrucchieri e di altri trattamenti estetici	13.1

TABLE 8 – Domaines professionnels, par ordre décroissant d'Écart Genré - IT_{Neutre}.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Évaluation de la Similarité Textuelle : Entre Sémantique et Surface dans les Représentations Neuronales

Julie Tytgat[1, 2] Guillaume Wisniewski[1] Adrien Betrancourt[2]
(1) Université de Paris Cité, LLF, CNRS 75 013 Paris, France
(2) IPSIDE, 31 100 Toulouse
julie.tytgat@etu.u-paris.fr, guillaume.wisniewski@u-paris.fr,
a.betrancourt@ipside.com

RÉSUMÉ

La mesure de la similarité entre textes, qu'elle soit basée sur le sens, les caractères ou la phonétique, est essentielle dans de nombreuses applications. Les réseaux neuronaux, en transformant le texte en vecteurs, offrent une méthode pratique pour évaluer cette similarité. Cependant, l'utilisation de ces représentations pose un défi car les critères sous-jacents à cette similarité ne sont pas clairement définis, oscillant entre sémantique et surface. Notre étude, basée sur des expériences contrôlées, révèle que les différences de surface ont un impact plus significatif que les différences de sémantique sur les mesures de similarité entre les représentations neuronales des mots construites par de nombreux modèles pré-entrainés. Ces résultats soulèvent des questions sur la nature même de la similarité mesurée par les modèles neuronaux et leur capacité à capturer les nuances sémantiques.

ABSTRACT

Evaluating Text Similarity : Between Semantics and Surface in Neural Representations

Measuring similarity between texts, whether based on meaning, characters, or phonetics, is essential in many applications. Neural networks, by transforming text into vectors, provide a practical method for assessing this similarity. However, the use of these representations is challenging because the criteria underlying this similarity are not clearly defined, oscillating between semantics and superficiality. Our study, based on controlled experiments, shows that surface differences have a more significant impact than semantic differences on measures of similarity between neural representations of words. These results raise questions about the nature of similarity measured by neural models and their ability to capture semantic nuance.

MOTS-CLÉS : Similarité textuelle, Analyse des Représentations Neuronales, Analyse Comparative de Modèles Pré-entraînés.

KEYWORDS : Text similarity, Neural Representations Analysis, Comparative Analysis of Pretrained Models.

1 Introduction

La mesure de la similarité entre deux textes est au cur de nombreuses applications, que la définition de la similarité soit basée sur le sens des textes (comme pour un moteur de recherche), la similarité des chaines de caractères (par exemple, pour permettre des correspondances floues) ou même la similarité phonétique (par exemple, dans les méthodes d'indexation basées sur la phonétique telles

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 85–96.

que `Soundex`).

Les réseaux neuronaux permettent de représenter le texte, qu'il s'agisse de mots, de phrases ou de paragraphes, sous forme de vecteurs, ce qui offre une méthode simple pour évaluer leur similarité. La mesure de la similarité entre deux vecteurs est en effet au cur de la fouille de données, et de nombreuses mesures de distance et de similarité aux propriétés bien établies ont été proposées dans la littérature (Duda *et al.*, 2001). Cette méthode est d'autant plus intéressante qu'il existe de nombreux modèles de langue pré-entraînés simplifiant la mise en uvre de celle-ci.

Un exemple concret d'utilisation de la similarité entre vecteurs contextuels, au cur aujourd'hui de nombreux travaux et développements, est le *Retrieval-Augmented Generation* (RAG) (Lewis *et al.*, 2020). Le RAG consiste à fournir à un giga-modèle (LLM) une base de documents pour enrichir sa production. Dans l'étape de récupération de documents, afin d'abonder la requête avec les passages pertinents, il est nécessaire de trouver les documents les plus similaires en fonction de leur produit scalaire avec le vecteur de requête dans un espace de vecteurs de haute dimension.

Toutefois, l'utilisation de représentations construites par les modèles de langue pour mesurer la similarité entre deux textes soulève un problème majeur. Bien qu'ils puissent faire de très bonnes prédictions dans de nombreuses tâches et en particulier générer des textes sémantiquement cohérents et syntaxiquement corrects, les réseaux neuronaux restent des modèles de type *boite noire* et les informations qu'ils encodent dans leur représentation ne sont pas clairement identifiées : s'il est facile d'utiliser des représentations neuronales pour mesurer la similarité, les critères sur lesquels cette similarité est fondée ne sont pas clairs.

Notre travail s'inscrit dans une longue série d'études visant à comprendre et à analyser les représentations apprises de manière auto-supervisée par les modèles de langue et en particulier les fameux *giga modèles* (Rogers *et al.*, 2020). Plus particulièrement, notre étude continue les travaux sur la pertinence des différentes mesures de similarité et leurs conditions d'utilisation sur les représentations tirées de ces modèles (Timkey & van Schijndel, 2021). Dans ce cadre, nous soulevons une nouvelle question : la similarité entre des représentations neuronales de texte est-elle fondée sur des critères sémantiques (comme on pourrait s'y attendre vues les bonnes performances des modèles utilisant ces représentations dans de nombreuses tâches) ou des critères de surface — deux alternatives qui ne s'excluent pas mutuellement.

Pour cela, suivant la proposition de (Isabelle *et al.*, 2017) d'évaluer les modèles sur des ensembles de données spécifiques construits autour de propriétés linguistiques clairement identifiées, nous proposons de mesurer la similarité entre une phrase et une version soigneusement modifiée de celle-ci, où les mots (noms, adjectifs et verbes) sont remplacés par des synonymes, des antonymes ou des paronymes (mot dont la prononciation est similaire à celle d'un autre mot, mais dont le sens est différent). Cette approche est détaillée dans la section 2. Grâce à ces expériences contrôlées, nous espérons pouvoir déterminer les critères sur lesquels se base la similarité mesurée entre les représentations neuronales des mots en établissant un lien entre la relation sémantique entre les mots échangés et la similarité entre les phrases.

Nos expériences, détaillées dans la section 3, montrent que les différences de surface ont un impact plus important sur les mesures de similarité que les différences de sémantique. Cette conclusion est particulièrement surprenante dans la mesure où de nombreuses études ont montré que les réseaux de neurones sont capables de construire des représentations abstraites des mots et des phrases (Li *et al.*, 2023). Nos expériences montrent également que la mesure de la similarité entre des représentations neuronales peut donner des résultats inattendus : des mots ayant des significations très différentes

peuvent être identifiés à tort comme étant très similaires.

2 Distinguer la similarité sémantique de la similarité de surface

Corpus Pour comprendre le fonctionnement interne de la similarité entre les représentations neuronales, nous avons créé un corpus de phrases en français et en anglais dans lesquelles un mot est remplacé par un autre mot dont la relation (sémantique) avec le mot d'origine est clairement identifiée [1]. Nous mesurons ensuite la similarité entre la phrase originale et la phrase modifiée en utilisant soit la similarité cosinus, soit la distance euclidienne.

Nous avons commencé, à l'aide de ressources en ligne, par compiler deux listes, une en anglais et une en français, de respectivement 373 et 354 mots associés à leur paronyme (des paires de mots, comme *irruption* et *éruption* dont la prononciation est similaire, mais pas le sens). Nous avons ensuite cherché dans différents corpus une phrase contenant un de ces mots.

Un mot et son paronyme ont des formes de surface et des prononciations très similaires, mais des significations très différentes. Il est courant, même pour un être humain, de confondre un mot avec son paronyme et d'utiliser l'un au lieu de l'autre. Si deux phrases dont la seule différence est l'utilisation d'un mot ou de son paronyme sont très similaires, la similarité repose davantage sur des informations de surface que sur la sémantique. On définit donc ici la similarité de surface comme un nombre de caractères en commun.

Pour chaque mot de notre liste, nous avons également cherché un antonyme et un synonyme, en nous assurant manuellement qu'ils pouvaient être utilisés à la place du mot d'origine dans la phrase (notamment en sélectionnant la forme correcte de ce dernier ou en sélectionnant parmi tous les synonymes ou antonymes d'un dictionnaire ceux qui pouvaient être utilisés dans le contexte). Nous avons ensuite répété la même expérience : en mesurant la similarité entre la phrase et la phrase dans laquelle le mot a été remplacé par un synonyme ou un antonyme, nous espérons pouvoir à la fois mieux comprendre quand une mesure de similarité appliquée à des représentations neuronales détecte que deux phrases sont similaires et déterminer la dynamique de ces mesures (notamment dans quel domaine la mesure varie). Bien que les sens distributionnels d'un mot et de son antonyme sont proches dans le cadre de la sémantique lexicale, nous nous plaçons du point de vue des utilisateurs finaux des LLMs, pour lesquels il est plus naturel de voir un antonyme comme particulièrement lointain du mot original.

Au final, notre corpus est constitué de 727 phrases, chacune apparaissant dans 5 versions différentes : la phrase originale et 4 versions modifiées dans lesquelles un mot a été successivement remplacé par un synonyme, un antonyme, un paronyme et un synonyme du paronyme. Le tableau 1 donne quelques exemples de phrases de notre corpus.

Test ABX Pour déterminer sur quel type d'informations repose la similarité entre deux représentations neuronales d'une phrase, nous utilisons un test ABX (Carlin *et al.*, 2011 ; Schatz *et al.*, 2013). Ce test s'appuie sur les représentations vectorielles construites par un modèle pré-entraîné de trois textes : deux textes, notés A et B sont proches sémantiquement et le troisième, noté X, a un sens différent. Le test ABX consiste simplement à vérifier si la similarité $s(A, B)$ est plus grande que $s(A, X)$.

1. Notre code et notre corpus seront diffusés lors de la publication.

①	originale	Le chirurgien procède à l'**ablation** du poumon.
	paronyme	Le chirurgien procède à l'**ablution** du poumon.
	synonyme	Le chirurgien procède à la **résection** du poumon.
	synonyme du paronyme	Le chirurgien procède à la **toilette** du poumon.
	antonyme	Le chirurgien procède à la **greffe** du poumon.
②	originale	Seasickness made him **retch** over the side of the boat.
	paronyme	Seasickness made him **wretch** over the side of the boat.
	synonyme	Seasickness made him **vomit** over the side of the boat.
	synonyme du paronyme	Seasickness made him **beggar** over the side of the boat.
	antonyme	Seasickness made him **swallow** over the side of the boat.

TABLE 1 – Exemples de phrases issues de nos corpus et les variantes considérées dans nos expériences.

Le score ABX correspond à la proportion de triplets pour lesquels $s(A, B) > s(A, X)$. Un score ABX proche de 50 % (ou inférieur) indique qu'en moyenne, la similarité entre A et X est plus grande que la similarité entre A et B, ce qui suggère que la similarité ne repose pas sur des informations sémantiques.

Modèles de langue Nous considérons [2] cinq modèles de langue multilingues pré-entrainés différents pour construire la représentation vectorielle des phrases de notre corpus : mBERT (Devlin *et al.*, 2019), le modèle ń sentence BERT ż (sBERT) de la phrase introduit dans (Reimers & Gurevych, 2019), le modèle d'OpenAI ADA, le modèle de Meta LLaMA-2 (Touvron *et al.*, 2023) et un modèle monolingue français, FlauBERT (Le *et al.*, 2020).

Différents *tokenizers* sont utilisés par ces modèles pour segmenter leur entrée : mBERT utilise WordPiece, sBERT utilise SentencePiece (Kudo & Richardson, 2018), tandis que ADA, LLaMA-2 et FlauBERT s'appuient sur BPE (Sennrich *et al.*, 2016).

Représentation de la phrase Si dans le cas de sBERT, la représentation construite par le réseau de neurones est directement celle de la phrase, ce n'est pas le cas pour les autres modèles qui construisent une représentation pour chaque token. Il existe plusieurs stratégies éprouvées pour construire la représentation d'une phrase à partir des représentations de ses tokens, la première étant d'utiliser la représentation du token spécial [CLS] comme représentation de la phrase, ce que nous faisons pour mBERT, ADA et FlauBERT. En revanche, l'approche choisie pour LLaMA consiste à calculer la représentation en moyennant [3] les *embeddings* de la sortie.

Certains de nos modèles ne peuvent être utilisés que par le moyen d'une API n'offrant accès qu'à certaines informations. Typiquement, nous ne pouvons accéder qu'aux représentations de la dernière couche d'ADA. Dans la mesure où ce modèle est utilisé dans de nombreuses applications, il nous a

2. Plus précisément, nous avons utilisé les modèles suivants via HuggingFace 🤗 (Wolf *et al.*, 2020) : bert-base-multilingual-cased pour mBERT, all-MiniLM-L6-v2 pour sBERT, Llama-2-7b-hf pour LLaMA-2 et flaubert_base_cased pour FlauBERT. Pour ADA, nous avons utilisé le modèle text-embedding-ada-002, via l'API fournie par OpenAI.

3. On présente en annexe, figure 5, différentes stratégies pour obtenir une représentation de la phrase. Si les résultats du test ABX sont sensiblement les mêmes dans tous les cas, les distributions non, les variations diminuant dans les cas avec normalisation (moyenne et standardisation), justifiant donc notre choix.

paru important de l'inclure dans notre comparaison. Par soucis de cohérence, nous avons décidé de considérer la dernière couche pour tous les modèles, même si plusieurs travaux récents (voir, par exemple, (Bordes *et al.*, 2023)) ont montré que, suivant les tâches, cette dernière couche ne permet pas toujours d'obtenir les meilleures représentations.

3 Résultats expérimentaux

La table 2 (resp. 3) reporte les résultats du test ABX pour l'anglais (resp. le français) pour les différentes combinaisons de substitutions décrites à la section 2. Ces tests sont effectués sur les similarités cosinus entre phrases. Les tables 5 et 6, en annexe, montrent que la différence obtenue en comparant des paires de phrases est minime. FlauBERT, bien qu'entrainé uniquement sur des données francophones, est également utilisé pour l'anglais comme expérience de contrôle, et les résultats sont décrits ici à titre indicatif.

Synonymes et antonymes Dans notre première expérience, nous comparons l'effet d'une substitution d'un mot par un synonyme (AB) avec une substitution par un antonyme (AX). Cette comparaison permet de s'assurer que le modèle différencie bien des phrases exprimant un sens contraire : plus le score ABX se rapproche de 100 %, plus les représentations de deux phrases ayant le même sens sont proches.

En ce qui concerne l'anglais, tous les modèles établissent une plus grande proximité dans le cas du synonyme que de l'antonyme. Néanmoins, si ADA le fait de manière quasi systématique, FlauBERT (pourtant francophone) est plus proche du hasard. En revanche, l'interprétation des résultats est moins évidente pour le français : mBERT et sBERT obtiennent un score ABX proche de 50 %, indiquant que les substitutions par un synonyme sont, en moyenne, aussi proches que celles par des antonymes. Par contre, ADA et LLaMA-2 restent eux capables de faire la distinction. De manière surprenante, ce n'est pas le cas pour un modèle monolingue français, FlauBERT. Il serait nécessaires de pouvoir mieux contrôler les données d'apprentissage (et notamment la proportion de données en français et de données en anglais) pour identifier les causes de cette différence de comportement.

Paronymes et synonymes Dans une seconde expérience, nous comparons les effets du remplacement d'un mot par un synonyme ou un paronyme. Dans notre corpus, les altérations paronymiques produisent souvent des phrases ayant des sens complètement différents, et parfois même des constructions n'ayant aucun sens, même si le paronyme a une forme de surface très proche du mot d'origine. À l'inverse, le sens de la phrase cible et de son homologue synonyme sont intrinsèquement similaires, voire identiques. La seconde colonne des tables 2 et 3 reporte les résultats obtenus en mesurant la similarité entre la phrase originale et la phrase comportant un paronyme, désignée par AX, et la similarité entre la phrase originale et la phrase comportant un synonyme, désignée par AB. Dans ce contexte, plus le score ABX est proche de 0%, plus le modèle a tendance à identifier le paronyme comme plus similaire au mot d'origine que le synonyme, suggérant que la similarité mesurée dépend fortement des informations lexicales.

On observe que pour le corpus anglais (table 2), mBERT, et dans une moindre mesure sBERT, capturent la proximité sémantique sans être influencés par la proximité des formes de surface. LLaMA-2 et ADA sont plus indécis, et FlauBERT, lui, favorise nettement la version paronymique.

	B : Synonyme X : Antonyme	B : Synonyme X : Paronyme	B : Paronyme X : Synonyme du paronyme
Attendu pour une similarité sémantique	100	100	50
mBERT	67,7	64,2	69,7
sBERT	78,6	67,7	69,4
ADA	94,5	58,9	92,3
LLaMA-2	83,6	57,8	89,0
FlauBERT	52,2	32,3	73,2

TABLE 2 – Résultat des tests ABX (en %) sur le corpus en anglais (similarité cosinus). A désigne systématiquement le mot d'origine.

De manière surprenante, les résultats du corpus français présentés table 3 sont différents : la similarité mesurée pour tous les modèles semble impactée par la proximité de surface à des degrés divers. Chez certains, mBERT, sBERT et LLaMA-2, cette influence est particulièrement nette : la représentation de la phrase dans laquelle le nom est remplacé par son paronyme est presque systématiquement plus proche de la représentation de la phrase originale que lorsque le nom est remplacé par un synonyme indiquant clairement que la similarité mesurée repose essentiellement sur la similarité des formes de surface.

Paronymes et synonymes des paronymes Dans cette troisième expérience, les deux variations considérées, avec le paronyme AB ou un synonyme du paronyme AX, ont la même signification — la différence étant que la phrase avec le paronyme présente une plus grande similarité de surface. Si nos modèles ne considèrent que le sens des mots, le score ABX devrait tendre vers 50% aucune des substitutions n'ayant de raison d'être plus similaire que l'autre à la phrase originale.Si le résultat s'approche de 100%, le paronyme est préféré suggérant une influence de la surface.

Dans le cas de l'anglais comme du français, la version avec un paronyme est en moyenne toujours préférée, de manière assez nette, suggérant une préférence pour une forme avec une surface similaire, et donc une influence de cette dernière. L'écart de cette influence varie entre les expériences 2 et 3, en particulier pour LLaMA-2 et ADA : en l'absence de critères sémantiques, la proximité de surface reste capturée.

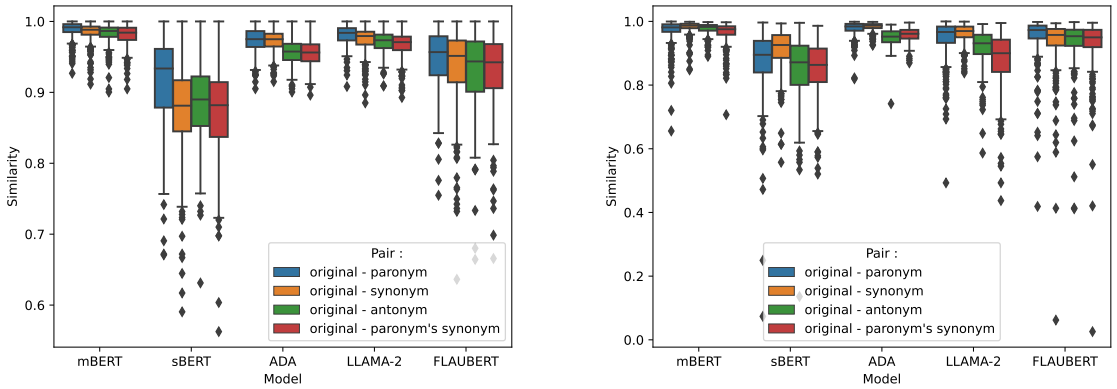
3.1 Distribution des distances

Pour compléter les résultats décrits dans la section précédente, nous avons représenté à la figure 1 (resp. figure 2) les distributions des similarités (resp. distances) entre les représentations des phrases d'origine et les phrases après substitution. Si les résultats varient d'une langue à l'autre, ces observations corroborent celles déjà décrites précédemment.

Pour le français, la similarité entre la phrase originale et la phrase comportant une substitution par un paronyme est toujours au dessus des autres substitutions, ce qui corrobore l'idée d'une influence de la surface, notamment car cette similarité est plus petite que celle reposant sur la substitution par un synonyme. Avec ADA et LLaMA-2, la similarité entre la phrase d'origine et la phrase avec un

	B : Synonyme X : Antonyme	B : Synonyme X : Paronyme	B : Paronyme X : Synonyme du paronyme
Attendu pour une similarité sémantique	100	100	50
mBERT	58,9	32,4	74,7
sBERT	56,0	21,5	80,8
ADA	83,1	49,4	86,2
LLaMA-2	67,6	33,7	89,4
FlauBERT	54,6	43,6	60,3

TABLE 3 – Résultat des tests ABX (en %) sur le corpus français (similarité cosinus). A désigne systématiquement le mot d'origine.



(a) Corpus français

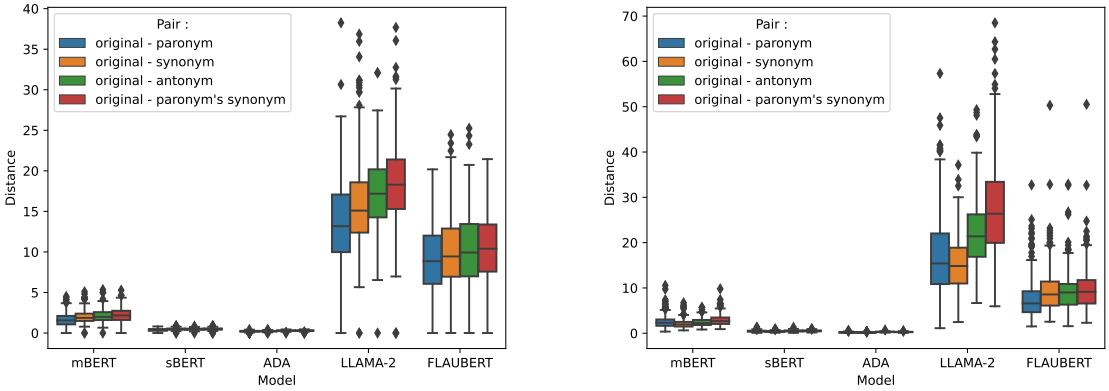
(b) Corpus anglais

FIGURE 1 – Distribution de mesures de similarité cosinus entre les phrases originales et après une substitution.

antonyme est systématiquement plus faible que lorsque la similarité est mesurée avec n'importe quel autre type de substitution. Mais sBERT considère parfois qu'une substitution par un synonyme donne une phrase moins similaire à la phrase originale que lorsque le mot est remplacé par un antonyme, une observation qui nous incite à la prudence dans l'interprétation des similarités entre les représentations neuronales.

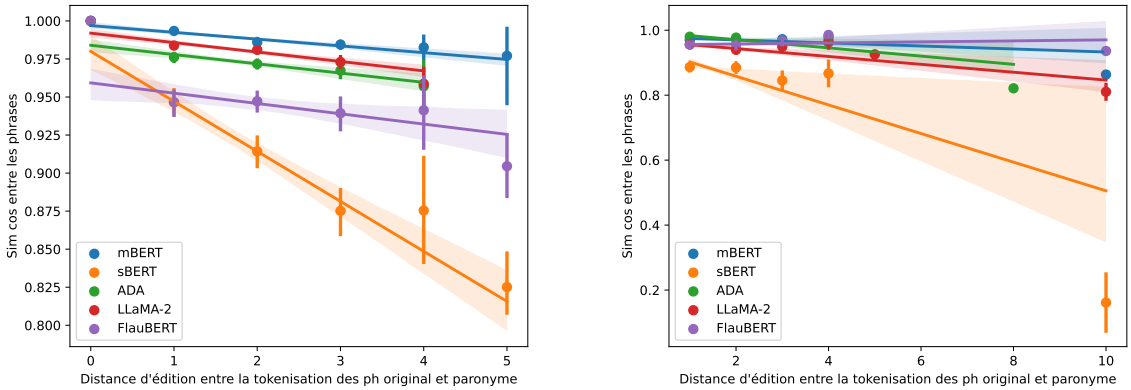
Pour l'anglais, la distribution des similarités est, pour tous les modèles, meilleure dans le cas d'une substitution avec un synonyme. Les antonymes sont eux correctement discriminés. Les modèles semblent dans ce cas bien hiérarchiser les phrases en fonction de leur proximité sémantique et non de surface. Néanmoins, les paronymes sont au dessus des synonymes des paronymes et non à égalité, n'excluant pas tout à fait l'influence d'une proximité de surface.

Ces observations montrent également que les similarités mesurées sont toujours très fortes et que les distances observées varient toujours dans un domaine très restreint. La valeur absolue de la distance ou de la similarité entre deux représentations est donc difficile à interpréter et devrait toujours être considérée avec précaution, par exemple si l'on souhaite définir un seuil pour filtrer des éléments à partir de celle-ci.



(a) Corpus français (b) Corpus anglais

FIGURE 2 – Distribution de mesures de distance euclidienne entre les phrases originales et après une substitution.



(a) Corpus français (b) Corpus anglais

FIGURE 3 – Relation entre la distance d'édition entre la tokenisation d'une phrase et de la version avec paronyme, et leur similarité cosinus.

3.2 Impact de la segmentation en sous-mots

Les résultats présentés dans la section précédente peuvent être faussés par la tokenisation : si notre analyse est effectuée au niveau du mot, le modèle ne manipule que les unités sous-lexicales et il est possible que la segmentation en unités sous-lexicales impacte la similarité et la magnitude de l'influence de la surface. Par exemple, si *acceptation* et son paronyme *acception* ont une distance d'édition de deux, leur segmentation en unités sous-lexicales (*accept##ation* et *accept##ion* respectivement) ne diffère que d'un seul token. La distance d'édition calculée au niveau du token n'est donc que de 1, et la tokenisation des sous-mots crée également un token supplémentaire identique entre les deux phrases.

Pour analyser cette possibilité, nous représentons dans la figure 3 l'évolution de la similarité entre la phrase originale et la phrase avec une substitution par un paronyme en fonction de la distance d'édition calculée au niveau des unités sous-lexicales. Cette figure montre clairement que le nombre

	Français	Anglais
mBERT	-0,36	-0.12
sBERT	-0,48	-0.43
ADA	-0,28	-0.43
LLaMA-2	-0,36	-0.19
FlauBERT	-0,15	0.03

TABLE 4 – Coefficients de Pearson entre la distance d'édition de la tokenisation d'une phrase et de la version avec paronyme, et leur similarité cosinus.

d'unités sous-lexicales a un impact sur la similarité : la similarité est d'autant plus faible que le nombre d'unité sous-lexicales différentes est important, mettant en évidence l'impact des informations de surface. Une mesure directe de la corrélation entre ces deux grandeurs (table 4) montre toutefois que cet effet est faible.

4 Conclusion

Dans cet article, nous avons décrit plusieurs expériences utilisant des mesures de similarité pour déterminer et mesurer la présence d'une interférence entre la surface (nombre de caractères en commun) et la représentation vectorielle de différents LLMs. Nos résultats montrent que le calcul d'une similarité entre deux représentations neuronales d'un texte repose essentiellement sur des informations de surface. Nos expériences montrent également que, quelle que soit la métrique considérée, les représentations neuronales détectent de très fortes similarités même entre des phrases de sens opposés, un résultat surprenant alors que de nombreux travaux ont mis en évidence la capacité de celles-ci à capturer le sens d'une phrase ou à générer un texte sémantiquement cohérent. Cette observation a des implications pratiques importantes, la détection de similitudes entre des textes étant au cur de nombreuses applications.

Ces travaux préliminaires peuvent être enrichis sur de nombreux aspects. Outre l'ajout de modèles et de langues pour consolider nos résultats et mesurer d'éventuels écarts, il pourrait également être intéressant d'étudier d'autres types de plongements, par exemple des *embeddings* de position, pour évaluer le rôle joué par la position du mot altéré. De même, la comparaison entre les couples de mots et leurs tokenisations pourrait être plus parlante avec un *embedding* statique, par exemple de type `fasttext`.

Remerciements

Nous remercions les relecteurs anonymes pour leur temps et précieux conseils, et pour avoir contribué à l'amélioration de ce papier. Ce travail a été financé par le projet DIAGNOSTIC soutenu par l'Agence de l'Innovation de Défense (subvention nř 2022 65 007).

Références

BORDES F., BALESTRIERO R., GARRIDO Q., BARDES A. & VINCENT P. (2023). Guillotine regularization : Why removing layers is needed to improve generalization in self-supervised learning. *Transactions on Machine Learning Research*.

CARLIN M. A., THOMAS S., JANSEN A. & HERMANSKY H. (2011). Rapid evaluation of speech representations for spoken term discovery. In *Twelfth Annual Conference of the International Speech Communication Association*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, p. 4171–4186. DOI : 10.18653/v1/N19-1423.

DUDA R. O., HART P. E. & STORK D. G. (2001). *Pattern Classification*. New York : Wiley, 2 édition.

ISABELLE P., CHERRY C. & FOSTER G. (2017). A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 2486–2496, Copenhagen, Denmark. DOI : 10.18653/v1/D17-1263.

KUDO T. & RICHARDSON J. (2018). SentencePiece : A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 66–71, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-2012.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). Flaubert : des modèles de langue contextualisés pré-entraînés pour le français. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, p. 268–278 : ATALA.

LEWIS P. S. H., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W., ROCKTÄSCHEL T., RIEDEL S. & KIELA D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, **abs/2005.11401**.

LI B., WISNIEWSKI G. & CRABBÉ B. (2023). Assessing the capacity of transformer to abstract syntactic representations : A contrastive analysis based on long-distance agreement. *Transactions of the Association for Computational Linguistics*, **11**, 18–33. DOI : 10.1162/tacl_a_00531.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence embeddings using Siamese BERT-networks. In *EMNLP*, p. 3982–3992. DOI : 10.18653/v1/D19-1410.

ROGERS A., KOVALEVA O. & RUMSHISKY A. (2020). A primer in BERTology : What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, **8**, 842–866. DOI : 10.1162/tacl_a_00349.

SCHATZ T., PEDDINTI V., BACH F., JANSEN A., HERMANSKY H. & DUPOUX E. (2013). Evaluating speech features with the minimal-pair ABX task : Analysis of the classical MFC/PLP pipeline. In *INTERSPEECH 2013 : 14th Annual Conference of the International Speech Communication Association*, p. 1–5.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1715–1725, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1162.

TIMKEY W. & VAN SCHIJNDEL M. (2021). All bark and no bite : Rogue dimensions in transformer language models obscure representational quality. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 4527–4546, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.372.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F., RODRIGUEZ A., JOULIN A., GRAVE E. & LAMPLE G. (2023). Llama : Open and efficient foundation language models.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., LE SCAO T., GUGGER S., DRAME M., LHOEST Q. & RUSH A. (2020). Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-demos.6.

A Résultats supplémentaires

Les figures 5 et 6 présentent les médianes des résultats obtenus en performant les différents tests ABX, tandis que la figure 5 montre la différence dans les distributions de distance euclidienne obtenues avec différentes stratégies de représentation de la phrase à partir des représentations des tokens : prendre le dernier, faire la moyenne ou bien effectuer une standardisation de la représentation du dernier token.

	B : Synonyme X : Antonyme	B : Synonyme X : Paronyme	B : Paronyme X : Synonyme du paronyme
mBERT	0.003873	0.003484	0.005049
sBERT	0.042417	0.024965	0.022611
ADA	0.030399	0.002901	0.016763
LLaMA-2	0.029717	0.004966	0.053019
FlauBERT	0.001616	-0.011391	0.015375

TABLE 5 – Test ABX sur corpus anglais (similarité cosinus) : médiane de la différence $AB - AX$

	B : Synonyme X : Antonyme	B : Synonyme X : Paronyme	B : Paronyme X : Synonyme du paronyme
mBERT	0.001941	-0.003248	0.005874
sBERT	0.005882	-0.038234	0.040888
ADA	0.017039	-0.000001	0.015579
LLaMA-2	0.004425	-0.004446	0.011270
FlauBERT	0.002520	-0.006072	0.009486

TABLE 6 – Test ABX sur corpus français (similarité cosinus) : médiane de la différence $AB - AX$

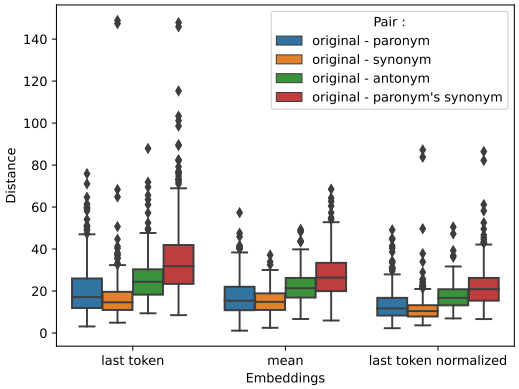


FIGURE 4 – Corpus anglais

FIGURE 5 – Comparaison des distances euclidienne en fonction de la stratégie de création du plongement de la phrase pour LLaMA-2

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Extraction des arguments d'événements à partir de peu d'exemples par méta-apprentissage

Aboubacar Tuo Romaric Besançon Olivier Ferret Julien Tourille
Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
{aboubacar.tuo,romaric.besancon,olivier.ferret,julien.tourille}@cea.fr

RÉSUMÉ

Les méthodes d'apprentissage avec peu d'exemples pour l'extraction d'événements sont développées pour réduire le coût d'annotation des données. Cependant, la plupart des études sur cette tâche se concentrent uniquement sur la détection des déclencheurs d'événements et aucune étude n'a été proposée sur l'extraction d'arguments dans un contexte de méta-apprentissage. Dans cet article, nous étudions l'extraction d'arguments d'événements avec peu d'exemples en exploitant des réseaux prototypiques et en considérant la tâche comme un problème de classification de relations. De plus, nous proposons d'améliorer les représentations des relations en injectant des connaissances syntaxiques dans le modèle par le biais de réseaux de convolution sur les graphes. Nos évaluations montrent que cette approche obtient de bonnes performances sur ACE 2005 dans plusieurs configurations avec peu d'exemples et soulignent l'importance des connaissances syntaxiques pour cette tâche.

ABSTRACT

A Meta-Learning Approach for Few-Shot Event Argument Extraction.

Few-shot learning techniques for Event Extraction are developed to alleviate the cost of data annotation. However, most studies on few-shot event extraction only focus on event trigger detection and no study has been proposed on argument extraction in a meta-learning context. In this paper, we investigate few-shot event argument extraction using prototypical networks, casting the task as a relation classification problem. Furthermore, we propose to enhance the relation embeddings by injecting syntactic knowledge into the model using graph convolutional networks. Our experimental results show that our proposed approach achieves strong performance on ACE 2005 in several few-shot configurations and highlight the importance of syntactic knowledge for this task.

MOTS-CLÉS : Extraction d'information, extraction d'événements, extraction d'arguments, apprentissage à partir de peu d'exemples.

KEYWORDS: Information extraction, event extraction, argument extraction, few-shot learning.

1 Introduction

L'extraction d'événements vise à identifier et extraire automatiquement des informations sur des événements à partir de textes non structurés, en se focalisant plus spécifiquement, pour chaque événement, sur son déclencheur (le mot ou la phrase correspondant à la mention de l'événement) et ses arguments (les entités qui jouent un rôle dans l'événement). Par exemple, dans la phrase « *Seven U.S. soldiers were killed when their vehicle hit an explosive device in Baghdad* », un événement de type Décès (*Life.Die* selon la nomenclature de référence de ACE 2005) est évoqué, associé au déclencheur

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 97–112.

killed et aux arguments *Seven U.S. soldiers*, *explosive device* et *Baghdad*, qui correspondent aux rôles respectifs de victime, d'instrument et de lieu dans la structure de l'événement. Les systèmes d'extraction d'événements reposent classiquement sur des approches supervisées, qui nécessitent une grande quantité de données annotées pour chaque type d'événement considéré. Cette annotation étant coûteuse, elle ne peut pas être effectuée pour toutes les applications, en particulier celles pour lesquelles de nouveaux types d'événements peuvent apparaître avec seulement quelques exemples. En conséquence, un intérêt croissant s'est développé pour relever le défi de l'extraction d'événements à partir de peu d'exemples.

La plupart des études dans ce domaine se concentrent uniquement sur la détection d'événements, qui consiste à extraire et classer les déclencheurs d'événements. Plusieurs de ces travaux reposent sur l'utilisation du méta-apprentissage et des réseaux prototypiques (Cong *et al.*, 2021; Tuo *et al.*, 2022, 2023). Cependant, très peu d'études abordent l'extraction des arguments d'événements en utilisant ces méthodes par méta-apprentissage. La plupart des méthodes existantes pour l'extraction des arguments d'événements dans des scénarios à faibles ressources ne relèvent pas en effet de l'apprentissage par transfert, mais cherchent plutôt à limiter la dégradation de leurs performances en présence d'une quantité limitée de données annotées. Ces études exploitent un large ensemble de méthodes allant du question-réponse (Du & Cardie, 2020; Zhou *et al.*, 2021) à l'implication textuelle (Sainz *et al.*, 2022) en passant par les méthodes génératives (Chen *et al.*, 2020; Hsu *et al.*, 2022; Dai *et al.*, 2022; Ma *et al.*, 2022). Des approches sans données annotées (*zero-shot*) ont également été proposées, soit en s'appuyant sur des ressources externes (Huang *et al.*, 2018; Zhang *et al.*, 2021), soit en utilisant des techniques de génération avec des modèles de langue pré-entraînés (Lin *et al.*, 2023).

Dans cette étude, nous proposons d'aborder l'extraction des arguments d'événements à partir de peu d'exemples à l'aide de réseaux prototypiques. Ce choix est principalement motivé par l'efficacité démontrée de ces méthodes dans la tâche de détection d'événements et dans plusieurs autres travaux en extraction d'information (Han *et al.*, 2018; Gao *et al.*, 2019; Fritzler *et al.*, 2018; Lai *et al.*, 2021). Notre objectif est d'évaluer les capacités de ces approches prototypiques dans le cadre de l'extraction des arguments et de proposer un cadre d'évaluation adapté à cette tâche dans un contexte de méta-apprentissage. Par ailleurs, nous proposons et évaluons deux approches pour l'injection d'informations syntaxiques dans la représentation des arguments d'événement.

2 Approche

2.1 Formulation du problème

Nous abordons la tâche d'extraction des arguments d'événements comme une tâche de classification de relations entre un déclencheur événementiel et les entités de la phrase abritant ce déclencheur [1]. Il s'agit plus précisément, pour chacune de ces entités, d'une classification multiclasse, chaque classe correspondant à l'un des rôles possibles pour le type d'événement associé au déclencheur considéré. S'y ajoute une classe dite *NULLE* pour signifier l'absence de rôle de l'entité candidate par rapport à l'événement.

De façon comparable à la détection d'événements, cette tâche peut être traitée selon la formulation épisodique N-ways, k-shots (Vinyals *et al.*, 2016) dans un contexte de méta-apprentissage, avec

1. Nous nous restreignons dans cette étude à la détection d'arguments d'événements au sein d'une même phrase.

cependant une légère variation. Cette formulation distingue, comme un apprentissage supervisé standard, des ensembles d'apprentissage, de validation et de test, mais à la différence du cas classique, les classes sont différentes pour ces trois ensembles. L'entraînement du modèle cible se fait en échantillonnant un grand nombre de fois un sous-ensemble de N classes au sein de l'ensemble d'apprentissage, avec k exemples par classe. Cet échantillonnage permet de créer autant de versions restreintes d'un ensemble d'apprentissage, appelé *support set*, et de test, appelé *query set*. Chaque itération dans ce cadre est appelée *épisode* et les poids du modèle sont mis à jour après chaque épisode. Cette façon de faire permet ainsi d'entraîner le modèle cible à un apprentissage par transfert, d'où le terme de méta-apprentissage.

Dans le cas de l'extraction d'arguments d'événements, bien que la classification concerne effectivement les arguments des événements, nous considérons les nouvelles classes au niveau des types d'événements. Cette approche correspond au scénario dans lequel les types d'événements de l'ensemble de test n'ont pas été rencontrés lors de l'entraînement, ce qui est davantage en phase avec les applications du monde réel où de nouveaux événements peuvent apparaître plutôt que de nouveaux rôles d'argument pour des événements existants. Par conséquent, notre formulation de l'approche N-ways, k-shots induit un N variable représentant le nombre d'arguments pour un type d'événement donné, chacune ayant k instances dans le support set. Par ailleurs, même si l'on peut considérer en toute généralité qu'un rôle n'est pas indépendant du type d'événement auquel il se rattache, des types d'événements différents peuvent contenir les mêmes rôles d'argument qui, au-delà de la similarité de nom, partagent de fait certaines similarités plus sémantiques, sans forcément être véritablement identiques [2]. Par conséquent, certains rôles de l'ensemble de test peuvent déjà avoir été vus lors de l'entraînement du modèle, liés à un autre type d'événement.

Dans ce contexte, chaque épisode se présente comme une tâche de classification, notée $\mathcal{T} = \{\mathcal{S}, \mathcal{Q}\}$ et comportant un support set, $\mathcal{S}$, et un query set, $\mathcal{Q}$. Le support set $\mathcal{S}$ comprend un type d'événement avec ses N classes d'arguments, chacune étant représentée par k instances annotées, tandis que le query set contient une phrase mentionnant au moins un événement du même type. Dans le paradigme des approches prototypiques, la classification se fait par un encodage des exemples du support set, la construction de prototypes en combinant ces exemples (le plus souvent par une moyenne) et l'étape de classification elle-même se fait par la sélection du prototype le plus proche pour chaque exemple du query set. La figure 1 donne une vue d'ensemble de ce dispositif.

2.2 Encodage d'une instance

Pour un type d'événement e donné, une instance est définie par (x_i, y_i) avec $x_i = (s_i^e, tr_i^e, a_i)$, où s_i^e est la phrase mentionnant l'événement, tr_i^e le déclencheur, a_i le candidat argument, et y_i le rôle appartenant à $\mathcal{A}^e = \mathcal{A}_+^e \bigcup \{NULLE\}$, $\mathcal{A}_+^e$ étant l'ensemble des arguments du type d'événement$e$ et $NULLE$ indiquant que l'entité n'a aucun rôle dans l'événement. La même phrase peut donc appartenir à autant d'exemples qu'elle contient de mentions d'entités.

Chaque instance est traitée par un encodeur pour produire une représentation vectorielle $h_i = \mathcal{E}(x_i)$ pour chaque paire déclencheur-entité dans le contexte d'un événement donné. En pratique, cette représentation est obtenue en concaténant les représentations du déclencheur et de l'entité résultant de leur encodage contextuel par un modèle de langue de type BERT (Devlin *et al.*, 2019). La

2. Par exemple, dans le domaine judiciaire, des événements différents comme une arrestation, une accusation ou une condamnation peuvent tous avoir un argument associé *Crime*, correspondant au crime pour lequel on est arrêté, accusé ou condamné.

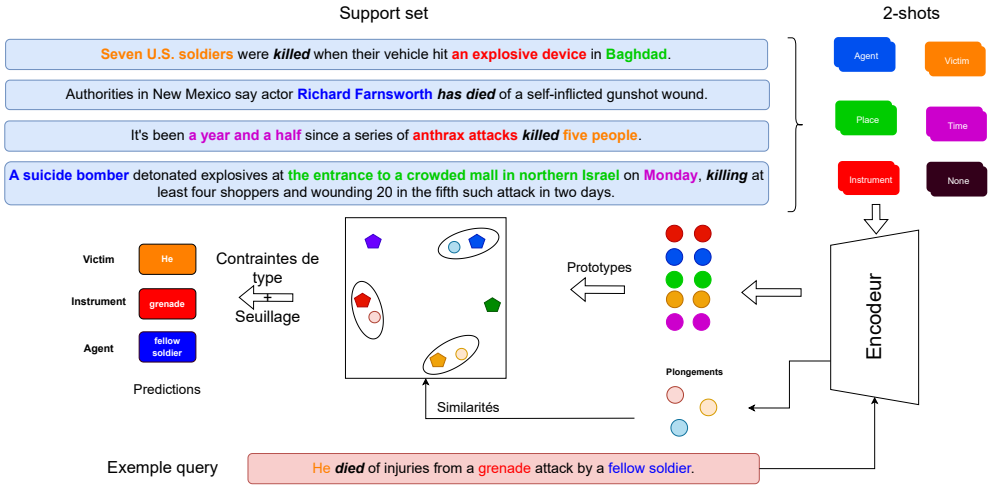


FIGURE 1 – Vue d'ensemble de notre modèle. Les déclencheurs sont en ***gras italique*** et chaque couleur correspond à une classe d'arguments.

représentation d'un empan de plusieurs sous-mots (*tokens*) est obtenue en prenant la moyenne des représentations de ces sous-mots. Cette représentation constitue l'entrée du classifieur, de type réseau prototypique, permettant de décider du type de relation existant entre le déclencheur et l'entité candidate (cf. section 2.4).

2.3 Intégration des informations syntaxiques

La représentation des exemples revêt une importance toute particulière dans le cas des réseaux prototypiques, car elle détermine très directement la forme des prototypes représentant chacune des classes et servant de base à la classification. Dans le cas de l'extraction d'arguments, nous proposons ainsi d'exploiter les relations syntaxiques entre le déclencheur et les entités pour aider à distinguer les entités lorsque leurs représentations (issues du modèle de langue contextuel) ne sont pas suffisamment discriminantes. Plus précisément, nous avons observé que certaines entités peuvent être confondues au sein d'un même événement, notamment lorsque leurs rôles se ressemblent ou sont symétriques. Par exemple, dans un contexte d'attaque, l'entité représentant l'agent attaquant peut parfois être confondue avec celle représentant la cible de l'attaque et inversement. De même, dans un contexte de transport, les entités représentant l'origine et la destination peuvent être sujettes à confusion. Pour lever ces ambiguïtés, nous proposons d'exploiter des informations syntaxiques additionnelles en nous appuyant sur le constat que les rôles des entités au sein d'un événement sont souvent étroitement liés à leurs rôles syntaxiques dans les phrases qui les décrivent. Dans l'exemple illustré de la figure 2, le déclencheur « fired » peut ainsi être lié à son argument « police officer » du fait de la présence d'une relation syntaxique de sujet (en voix passive) alors que ces deux éléments sont distants dans la phrase.

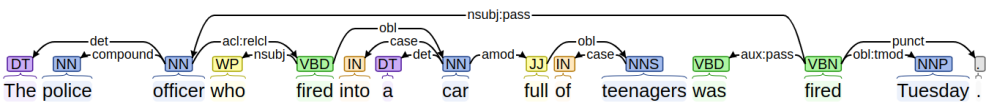


FIGURE 2 – Analyse en dépendances syntaxiques et étiquettes morphosyntaxiques pour la phrase : *The police officer who fired into a car full of teenagers was fired Tuesday.*

Nous utilisons l'encodeur BERT comme point de référence pour évaluer les avantages de l'injection d'informations syntaxiques. Sur cette base, nous proposons ensuite deux façons de prendre en compte les informations syntaxiques pour enrichir les représentations des paires déclencheur/entité.

BERT++ intègre de manière statique les informations syntaxiques en associant un vecteur à chaque type de relation syntaxique, ces vecteurs étant ajustés pendant l'apprentissage du modèle. Plus précisément, cette approche consiste à concaténer les plongements contextuels du déclencheur et de l'entité avec les représentations vectorielles des étiquettes morphosyntaxiques de l'entité et du chemin de dépendances syntaxiques entre le déclencheur et l'entité. La longueur des chemins de dépendances syntaxiques étant variable, nous avons appliqué une agrégation *max-pooling* sur l'ensemble des étiquettes de dépendance syntaxique afin d'obtenir les plongements de ces chemins.

BERT-GCN combine les informations syntaxiques avec les plongements fournis par BERT en utilisant un réseau de convolution sur les graphes (GCN) (Kipf & Welling, 2017). Cette approche est particulièrement intéressante dans notre cas, car elle permet de différencier les relations syntaxiques, contribuant ainsi à réduire les confusions entre des arguments similaires. Cependant, cette modélisation entraîne une augmentation significative du nombre de paramètres à apprendre, correspondant aux paramètres du GCN.

Afin de permettre à l'encodeur BERT de capturer des informations sur l'emplacement du déclencheur et de l'entité cible au sein d'une phrase, nous marquons le début et la fin de ces éléments avec des tokens spéciaux, comme suggéré dans d'autres travaux de la littérature (Zhang *et al.*, 2019; Han *et al.*, 2018; Baldini Soares *et al.*, 2019). La phrase est ensuite fournie l'encodeur BERT pour en obtenir une représentation de chaque token. Ces représentations sont ensuite utilisées pour initialiser les nœuds de l'arbre de dépendances syntaxiques qui sert d'entrée au réseau de convolution. Nous introduisons un nouveau type de relation syntaxique appellé *subword* correspondant au fait qu'un mot puisse être découpé en plusieurs tokens par le tokeniseur de BERT. Après apprentissage du réseau de convolution sur les graphes, la représentation de chaque empan est obtenue en prenant la moyenne des représentations des nœuds appartenant à cet empan.

2.4 Classification des instances

Notre modèle reposant sur les réseaux prototypiques, la classification des instances est réalisée en fonction de leur similarité avec les représentations des prototypes de chaque classe. À cette fin, nous avons mené des expériences avec deux approches : les réseaux prototypiques standard (*Prototypical Networks*, **Proto**) et leur version contrastive (*Contrastive Prototypical Networks*, **C-Proto**), proposée par Tuo *et al.* (2023).

La différence entre ces deux modèles réside dans le fait que **C-Proto** adopte un apprentissage contrastif et ne construit pas de prototype pour la classe *NULLE*, qui regroupe les entités n'ayant pas de rôle dans les événements. Dans ce cas, les exemples de cette classe sont filtrés à l'aide d'un seuil de similarité. Alors que le modèle **Proto** s'appuie sur une fonction de coût de type entropie croisée, couplée à une fonction softmax, le modèle **C-Proto** est entraîné avec une fonction de coût de charnière (*hinge loss*). Cette dernière est plus précisément la somme de deux termes : l'un ($\mathcal{L}_+$) concerne uniquement les arguments et l'autre ($\mathcal{L}_-$), les entités non-arguments (cf. équations 1 et 2). Nous notons $\mathcal{Q}^+$ le sous-ensemble des arguments du query set d'un épisode et $\mathcal{Q}^-$, le sous-ensemble du query set contenant des entités non-arguments).

$$\mathcal{L}_{+}(\mathcal{S},\mathcal{Q}) = \sum_{(x_i,y_i)\in\mathcal{Q}^{+}} \sum_{j\neq y_i} \max_{j}(0, \mathcal{M}_0 - s(h_i, c^{j}) + s(h_i, c^{y_i})) \quad (1)$$

$$\mathcal{L}_{-}(\mathcal{S},\mathcal{Q}) = \max_{(x_i,y_i)\in\mathcal{Q}^{+}}(0, \max_{x_j\in\mathcal{Q}^{-}}(s\left(h_j, c^{y_i}\right) - \mathcal{M}_1)) \quad (2)$$

où $s(.,.)$ est une fonction de similarité, c^i le prototype de la classe i, h_i la représentation dense d'une entité, $\mathcal{M}_0$ et $\mathcal{M}_1$ des hyperparamètres permettant de contrôler la marge entre les arguments et les entités non-arguments. Une telle marge laisse ensuite la possibilité de fixe un seuil de similarité permettant de séparer les arguments des entités non-arguments.

Contrairement à Tuo *et al.* (2023), qui ont recours à une fonction de répartition pour estimer ce seuil, nous calculons ici le seuil en utilisant la valeur de similarité trouvée pour l'exemple le plus proche dans le support set. En effet, dans le cas de l'extraction des arguments, certaines phrases peuvent en effet ne contenir qu'une seule entité candidate ou seulement des entités correspondant à des non-arguments [3]. Par conséquent, l'utilisation de la fonction de répartition dans le cas d'une seule entité ou d'entités n'ayant pas de rôle ne permettrait pas de fixer un seuil raisonnable. Il faudrait, dans ce cas, que le seuil soit supérieur aux similarités pour toutes les entités.

De plus, à l'instar de travaux antérieurs sur l'extraction d'événements (Sainz *et al.*, 2022; Lin *et al.*, 2023), nous utilisons la connaissance préalable des types des entités pour contraindre les prédictions des rôles d'arguments et améliorer ainsi la précision des prédictions.

En pratique, lors de l'évaluation du modèle, nous commençons par vérifier la compatibilité entre le rôle prédit et le type de l'entité candidate. Si la contrainte n'est pas satisfaite, nous considérons la prédiction comme incorrecte et prenons la classe du prototype suivant le plus proche jusqu'à ce que le rôle prédit corresponde au type d'entité ou à la classe *NULLE*, qui n'est soumise à aucune contrainte.

3 Expériences

3.1 Paramètres expérimentaux

Nous avons mené nos expériences sur l'ensemble de données ACE-2005 (Walker *et al.*, 2006), avec la partition fournie par Lai *et al.* (2020). Cette partition garantit qu'il n'y a pas de chevauchement entre les types d'événements dans les ensembles d'entraînement et d'évaluation, simulant ainsi un scénario réaliste avec une faible disponibilité des données. Nous donnons plus de détails sur ce jeu de données à l'Annexe A.

Nous utilisons un encodeur **BERT** pour fournir les représentations des mots à partir des phrases en entrée. De plus, pour l'encodeur **BERT++**, nous utilisons des vecteurs entraînables de taille 256 pour encoder les dépendances syntaxiques et les étiquettes morphosyntaxiques, obtenues grâce à l'analyseur spaCy. Pour **BERT-GCN**, le nombre de couches de convolution du GCN a été fixé à 2, car c'est ce qui donne le meilleur résultat empirique sur l'ensemble de développement.

3. Dans ces cas, les arguments de la mention d'événement considérée se trouvent souvent dans d'autres phrases. Il faudrait une extraction au niveau du document pour les identifier.

3.2 Résultats

Nous reportons nos principaux résultats dans le tableau 1. Les entités considérées pour l'extraction des arguments sont les entités annotées dans le jeu de données.

		5-shots			10-shots		
Encodeur	Modèle	P	R	F1	P	R	F1
BERT	Proto	63,1 ± 0,9	56,4 ± 1,0	59,6 ± 0,5	66,4 ± 0,5	61,6 ± 0,7	63,9 ± 0,3
	C-Proto	62,7 ± 0,9	57,0 ± 1,2	60,0 ± 1,0	67,1 ± 0,8	63,8 ± 0,9	65,5 ± 0,8
BERT++	Proto	64,9 ± 1,1	58,6 ± 1,2	61,6 ± 0,8	66,8 ± 1,5	63,8 ± 1,1	65,2 ± 0,6
	C-Proto	65,8 ± 0,5	58,8 ± 1,8	62,1 ± 1,0	66,8 ± 1,7	**66,5*** ± **1,7**	**66,7*** ± **1,0**
BERT-GCN	Proto	**69,0** ± **2,1**	56,6 ± 4,0	62,2 ± 2,2	**71,2*** ± **0,7**	60,0 ± 1,5	65,0 ± 0,9
	C-Proto	68,5 ± 1,1	**59,2*** ± **1,7**	**63,5*** ± **1,2**	69,2 ± 0,5	61,4 ± 0,8	65,1 ± 0,5

TABLE 1 – Résultats de l'extraction des arguments d'événements : Précision (P), Rappel (R) et F1-mesure (F1). Nos meilleurs scores sont en **gras**, les deuxièmes meilleurs scores sont soulignés. * indique que le meilleur score est statistiquement significatif par rapport au deuxième.

Ces résultats nous permettent de tirer deux conclusions principales. D'une part, quel que soit l'encodeur considéré, nous pouvons observer que la version contrastive **C-Proto** affiche des performances légèrement supérieures à celles du réseau prototypique standard. Cette observation vient confirmer les constatations faites par Tuo *et al.* (2023) sur l'efficacité de cette approche, en particulier en ce qui concerne la gestion de la classe *NULLE*. Toutefois, les apports dans ce cadre sont bien moindres que ceux rapportés pour la tâche de détection d'événements. Cela peut être lié à la difficulté relative de la tâche d'extraction des arguments par rapport à l'extraction des déclencheurs. En effet, les entités occupant un même rôle n'ont pas forcément de similarités sémantiques entre elles, contrairement aux déclencheurs, qui appartiennent souvent au même champ lexical.

D'autre part, nous voyons que l'intégration des informations syntaxiques améliore les performances dans tous les cas. Cette constatation suggère que l'exploitation d'informations syntaxiques pour enrichir la relation entre les déclencheurs d'événements et leurs arguments est bien bénéfique pour la tâche d'extraction des arguments d'événements. Par ailleurs, l'intégration dynamique **BERT-GCN** semble plus efficace que l'intégration statique **BERT++** lorsque très peu de données sont disponibles (5-shots).

Nous comparons de façon plus précise nos trois encodeurs à la figure 3, où nous donnons la moyenne de la F1-mesure pour chaque rôle. Dans l'ensemble, ces résultats montrent que l'intérêt des informations syntaxiques est observé à la fois pour les rôles vus pendant l'entraînement (c'est-à-dire des rôles qui portent le même nom, mais qui correspondent à des types d'événements différents) et pour les nouveaux rôles apparaissant uniquement pendant l'évaluation. Les apports se manifestent particulièrement pour les rôles pour lesquels l'encodeur **BERT** standard montre des performances relativement faibles. En revanche, pour des rôles considérés comme « faciles », qui ne portent pas à confusion, comme *Instrument*, *Vehicule* ou *Money*, l'encodeur **BERT** standard demeure très compétitif et les informations syntaxiques ont un impact moins significatif, voire parfois négatif. En particulier, l'encodeur **BERT-GCN** semble contribuer principalement à équilibrer les performances pour les rôles qui peuvent être confondus, tels que *Origin* et *Destination* ou encore *Buyer* et *Seller*. En revanche, il nuit moins aux rôles non ambigus, tels que *Instrument* ou *Vehicule*.

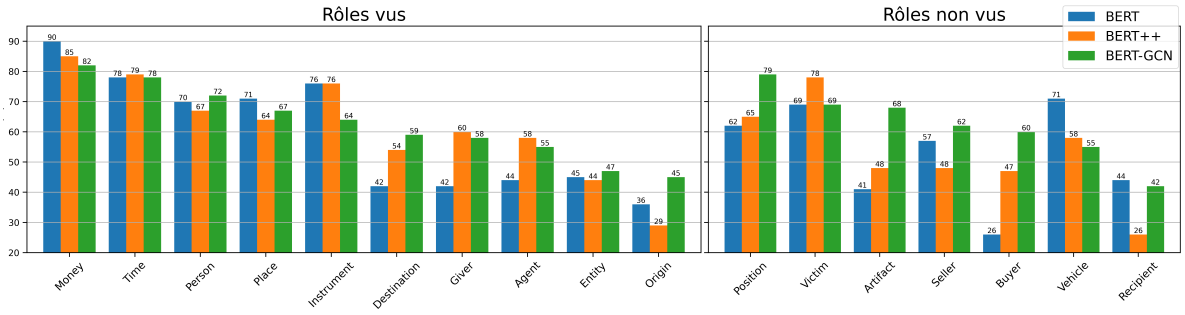


FIGURE 3 – Moyenne de la F1-mesure par rôle pour les trois encodeurs. À gauche, les rôles vus pendant l'entraînement des modèles et à droite, les rôles vus seulement pendant l'évaluation.

	P	R	F1
Modèle entier	**68,5**	**59,2**	**63,5**
- seuillage	47,9	59,3	52,9
- contraintes	68,2	50,9	58,3
- seuillage & contraintes	33,9	61,9	43,8

TABLE 2 – Étude d'ablation pour chaque composante du modèle **C-Proto** dans une configuration 5-shots. Précision (P), Rappel (R) et F1-mesure (F1) en moyenne sur cinq expérimentations.

La figure 4 donne pour sa part un aperçu des représentations élaborées par chaque encodeur sur l'ensemble d'évaluation. L'encodeur **BERT pré-entraîné** correspond à un modèle BERT sans aucun ajustement spécifique à la tâche d'extraction des arguments d'événements. Nous comparons les trois encodeurs présentés dans ce travail : **BERT**, **BERT++** et **BERT-GCN**.

Tout d'abord, il est évident que l'entraînement du modèle BERT améliore notablement le caractère discriminant des plongements du point de vue des classes d'arguments par rapport à un BERT non affiné. Cela met en lumière la pertinence de la formulation que nous avons adoptée pour cette tâche et l'importance de l'affinage du modèle BERT dans ce contexte.

On peut également observer qualitativement que les deux encodeurs enrichis, **BERT++** et **BERT-GCN**, semblent fournir des représentations plus discriminantes que l'encodeur BERT d'origine. Ces observations correspondent aux résultats obtenus lors de l'évaluation, ce qui renforce la pertinence de l'enrichissement des représentations par des informations syntaxiques et suggère une amélioration globale de la performance du modèle.

Étude d'ablation Pour compléter notre analyse des résultats, nous présentons dans le tableau 2 une étude d'ablation réalisée avec le modèle **C-Proto**. L'objectif est d'explorer les effets de l'utilisation du seuillage et de l'introduction des contraintes liées au type des entités et à leurs rôles. Afin d'éliminer l'utilisation du seuil, nous avons reconstruit un prototype pour la classe nulle lors de la phase d'évaluation.

Les résultats obtenus dans cette étude d'ablation mettent en évidence l'utilité de chacune de ces opérations sur les performances globales du modèle. En général, nous observons que l'utilisation du

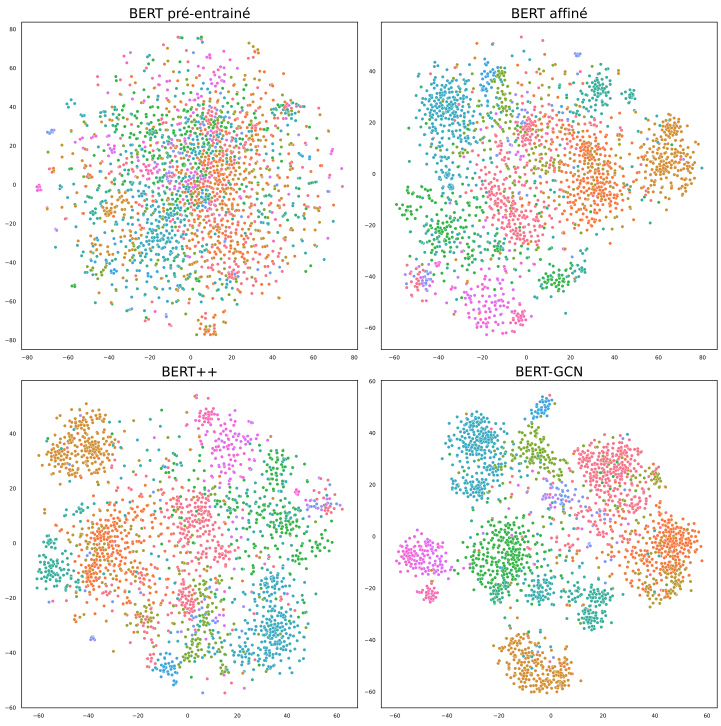


FIGURE 4 – Visualisation des représentations des arguments en utilisant la méthode t-SNE. Chaque point correspond à un argument et la couleur, à une classe de rôle.

seuillage conduit principalement à une amélioration de la précision, bien qu'elle soit accompagnée d'une légère diminution du rappel en contrepartie. Il faut noter que l'introduction du seuillage visait précisément à réduire le taux de faux positifs et, donc, à une augmentation de la précision.

Les contraintes relatives aux types des entités et leurs rôles ont aussi un impact positif sur les performances, bien que leur contribution soit plus modeste par rapport au seuillage. Ces contraintes visent principalement à réduire les confusions entre certains arguments. Comme mentionné précédemment, il arrive que des entités de même type soient parfois confondues dans leurs rôles et, bien que ce filtrage n'élimine pas entièrement cette ambiguïté, il contribue néanmoins à la résoudre partiellement.

Comparaison avec l'état de l'art Comme nous l'indiquions en introduction, la comparaison avec l'état de l'art n'est pas évidente dans la mesure où les méthodes d'évaluation pour les approches par méta-apprentissage ne sont pas directement comparables aux évaluations menées par les travaux existants en extraction d'arguments d'événements, qui reprennent le paradigme classique de l'apprentissage supervisé mais dans une configuration dégradée de faible quantité de données annotées.

Pour réaliser une forme de comparaison, nous nous sommes focalisés sur la quantité de données annotées concernant les types d'événements de l'ensemble de test du jeu de données ACE 2005. La figure 5 fait ainsi apparaître l'évolution des performances pour l'extraction d'arguments d'événements pour trois modèles de référence – **PAIE** (Ma *et al.*, 2022), **BIP** (Dai *et al.*, 2022) et **NLI** (Sainz *et al.*, 2022) – en fonction du pourcentage des données d'entraînement. Au regard de ces courbes, nous avons fait figurer le niveau de performance de notre configuration 5-shots (c'est-à-dire 5 exemples par rôle), qui correspond en quantité à environ 3 % des données d'évaluation. Il faut toutefois noter que l'entraînement de notre modèle se fait sur 18 types d'événements et tous leurs exemples, les 3 %

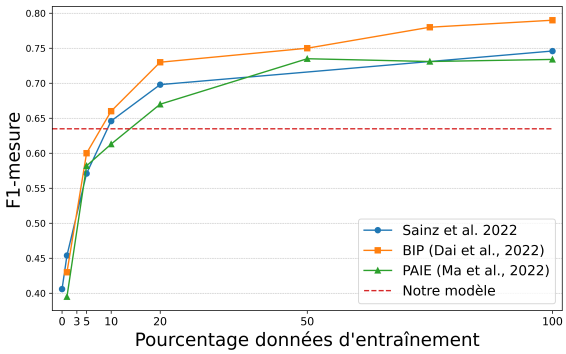


FIGURE 5 – Comparaison avec des modèles de l'état de l'art avec peu de données annotées.

de données ne concernant que les types non vus pendant l'entraînement. Néanmoins, la figure 5 fait apparaître que jusqu'à 5 % des données d'entraînement, le modèle que nous proposons obtient des performances supérieures aux modèles de référence considérés. Nous pouvons donc en conclure que notre approche par méta-apprentissage est particulièrement adaptée à un régime de très faible quantité de données annotées pour les types d'événements cibles.

4 Conclusion et perspectives

Nous avons exploré l'extraction des arguments d'événements dans un scénario de faible disponibilité de données pour définir de nouvelles classes en mettant l'accent sur l'utilisation d'informations syntaxiques pour lever l'ambiguïté concernant certains rôles événementiels. Nous avons repensé ce problème en le formulant comme une tâche de classification de relations entre le déclencheur et les arguments et en le traitant dans un cadre de méta-apprentissage à l'aide de réseaux prototypiques. Dans ce contexte, nous avons adapté plus précisément le cadre de classification N-ways, k-shots pour répondre aux besoins spécifiques de l'extraction d'arguments d'événements. Les évaluations menées ont montré l'intérêt de la prise en compte de ces informations syntaxiques, en particulier pour les rôles les plus susceptibles de se confondre.

Le travail mené fait l'hypothèse de la connaissance a priori des déclencheurs événementiels et des entités candidates pour les rôles et présuppose donc une architecture de type pipeline, qui présente l'avantage d'une certaine modularité mais souffre de problèmes connus de propagation d'erreurs entre modules. À l'instar des travaux de Nguyen *et al.* (2021), dans un contexte de quantités importantes de données annotées, un des prolongements naturels de notre travail sera d'examiner comment des approches jointes impliquant l'extraction d'entités nommées, de déclencheurs événementiels et d'arguments d'événements peuvent être mises en œuvre dans un contexte de faible quantité de données annotées disponibles par le biais de méthodes de méta-apprentissage.

Remerciements Ces travaux ont été réalisés grâce au supercalculateur Factory-IA financé par le Conseil Régional d'Île-de-France.

Références

BALDINI SOARES L., FITZGERALD N., LING J. & KWIATKOWSKI T. (2019). Matching the blanks : Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 2895–2905, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1279.

CHEN Y., CHEN T., EBNER S., WHITE A. S. & VAN DURME B. (2020). Reading the manual : Event extraction as definition comprehension. In *Proceedings of the Fourth Workshop on Structured Prediction for NLP*, p. 74–83, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.spnlp-1.9.

CONG X., CUI S., YU B., LIU T., YUBIN W. & WANG B. (2021). Few-Shot Event Detection with Prototypical Amortized Conditional Random Field. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 28–40, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.3.

DAI L., WANG B., XIANG W. & MO Y. (2022). Bi-directional iterative prompt-tuning for event argument extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 6251–6263, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DU X. & CARDIE C. (2020). Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 671–683, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.49.

FRITZLER A., LOGACHEVA V. & KRETOV M. (2018). Few-shot classification in Named Entity Recognition Task. *arXiv :1812.06158 [cs, stat]*. arXiv : 1812.06158, DOI : 10.1145/3297280.3297378.

GAO T., HAN X., LIU Z. & SUN M. (2019). Hybrid Attention-Based Prototypical Networks for Noisy Few-Shot Relation Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 6407–6414. DOI : 10.1609/aaai.v33i01.33016407.

HAN X., ZHU H., YU P., WANG Z., YAO Y., LIU Z. & SUN M. (2018). FewRel : A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 4803–4809, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1514.

HSU I.-H., HUANG K.-H., BOSCHEE E., MILLER S., NATARAJAN P., CHANG K.-W. & PENG N. (2022). DEGREE : A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1890–1908, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.138.

HUANG L., JI H., CHO K., DAGAN I., RIEDEL S. & VOSS C. (2018). Zero-shot transfer learning for event extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2160–2170, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1201.

KIPF T. N. & WELLING M. (2017). Semi-supervised classification with graph convolutional networks.

LAI V., DERNONCOURT F. & NGUYEN T. H. (2021). Learning Prototype Representations Across Few-Shot Tasks for Event Detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 5270–5277, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics.

LAI V. D., NGUYEN T. H. & DERNONCOURT F. (2020). Extensively matching for few-shot learning event detection. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, p. 38–45, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.nuse-1.5.

LIN Z., ZHANG H. & SONG Y. (2023). Global constraints with prompting for zero-shot event argument classification. In *Findings of the Association for Computational Linguistics : EACL 2023*, p. 2527–2538, Dubrovnik, Croatia : Association for Computational Linguistics.

MA Y., WANG Z., CAO Y., LI M., CHEN M., WANG K. & SHAO J. (2022). Prompt for extraction ? PAIE : Prompting argument interaction for event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 6759–6774, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.466.

NGUYEN M. V., LAI V. D. & NGUYEN T. H. (2021). Cross-task instance representation interactions and label dependencies for joint information extraction with graph convolutional networks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 27–38, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.3.

SAINZ O., GONZALEZ-DIOS I., LOPEZ DE LACALLE O., MIN B. & AGIRRE E. (2022). Textual entailment for event argument extraction : Zero- and few-shot with multi-source learning. In *Findings of the Association for Computational Linguistics : NAACL 2022*, p. 2439–2455, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-naacl.187.

TUO A., BESANÇON R., FERRET O. & TOURILLE J. (2022). Better exploiting bert for few-shot event detection. In P. ROSSO, V. BASILE, R. MARTÍNEZ, E. MÉTAIS & F. MEZIANE, Éds., *Natural Language Processing and Information Systems*, p. 291–298, Cham : Springer International Publishing.

TUO A., BESANÇON R., FERRET O. & TOURILLE J. (2023). Trigger or not trigger : Dynamic thresholding for few shot event detection. In J. KAMPS, L. GOEURIOT, F. CRESTANI, M. MAISTRO, H. JOHO, B. DAVIS, C. GURRIN, U. KRUSCHWITZ & A. CAPUTO, Éds., *45th European Conference on Information Retrieval (ECIR 2023) : Advances in Information Retrieval, short article session*, volume 13981 de *Lecture Notes in Computer Science*, p. 637–645, Dublin, Ireland : Springer Nature Switzerland.

VINYALS O., BLUNDELL C., LILLICRAP T., KAVUKCUOGLU KORAY K. & WIERSTRA D. (2016). Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29.

WALKER C., STRASSEL S. & JULIE MEDERO K. M. (2006). *ACE 2005 Multilingual Training Corpus*. LDC corpora. Linguistic Data Consortium. DOI : 10.35111/mwxc-vh88.

ZHANG H., WANG H. & ROTH D. (2021). Zero-shot Label-aware Event Trigger and Argument Classification. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 1331–1340, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.114.

ZHANG Z., HAN X., LIU Z., JIANG X., SUN M. & LIU Q. (2019). ERNIE : Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1441–1451, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1139.

ZHOU Y., CHEN Y., ZHAO J., WU Y., XU J. & LI J. (2021). What the role is vs. what plays the role : Semi-supervised event argument extraction via dual question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**(16), 14638–14646. DOI : 10.1609/aaai.v35i16.17720.

A Jeu de données ACE-2005

ACE-2005 est un jeu de données de référence pour l'extraction d'information, composé de 599 documents [4] de diverses sources, en anglais, mandarin et arabe. Dans sa version anglaise, il propose huit types d'événements, subdivisés en 33 sous-types et 34 rôles pour les arguments. Sa richesse et sa diversité en font une ressource clé pour le développement de modèles d'extraction d'information multilingues et polyvalents. Nous listons l'ensemble des types et sous-types dans le tableau 3.

Pour nos expérimentations, nous adoptons le découpage de Lai *et al.* (2020), qui considère quatre types (*Conflict*, *Business*, *Contact* et *Justice*) dans l'ensemble d'entraînement et les autres types dans l'ensemble d'évaluation. Dans ce découpage, certains rôles n'apparaissent que dans l'ensemble d'évaluation, tandis que d'autres sont exclusivement présents dans l'ensemble d'entraînement ; mais il existe également des rôles communs aux deux ensembles (voir tableau 4). Nous avons supprimé les rôles d'argument comportant moins de 10 exemples afin de disposer de suffisamment de données pour constituer des épisodes avec au moins 10 exemples dans les ensembles de support. Nous avons également retiré les types pour lesquels il ne restait aucun argument (c'est-à-dire *Business:End-Org*, *Justice:Pardon*, *Justice:Extradite* et *Justice:Acquit* dans l'ensemble d'entraînement, ainsi que *Personnel:Nominate* dans l'ensemble d'évaluation).

4. Dans la version anglaise.

Types	Sous-types	Arguments
Life	Be-born	Person, Place, Time
	Marry	Person, Place, Time
	Divorce	Person, Place, Time
	Injure	Agent, Victim, Instrument, Place, Time
	Die	Agent, Victim, Instrument, Place, Time
Movement	Transport	Agent, Artifact, Vehicle, Price, Origin, Destination, Time
Transaction	Transfer-Ownership	Buyer, Seller, Beneficiary, Artifact, Price, Place, Time
	Transfer-Money	Giver, Recipient, Beneficiary, Money, Place, Time
Business	Start-Org	Agent, Org, Place, Time
	Merge-Org	Org, Place, Time
	Declare-Bankruptcy	Org, Place, Time
	End-Org	Org, Place, Time
Conflict	Attack	Attacker, Target, Instrument, Place, Time
	Demonstrate	Entity, Place, Time
Contact	Meet	Entity, Place, Time
	Phone-Write	Entity, Time
Personnel	Start-Position	Person, Entity, Position, Place, Time
	End-Position	Person, Entity, Position, Place, Time
	Nominate	Person, Agent, Position, Place, Time
	Elect	Person, Entity, Position, Place, Time
Justice	Arrest-Jail	Person, Agent, Crime, Place, Time
	Release-Parole	Person, Entity, Crime, Place, Time
	Trial-Hearing	Defendant, Prosecutor, Adjudicator, Crime, Place, Time
	Charge-Indict	Defendant, Prosecutor, Adjudicator, Crime, Place, Time, Sentence
	Sue	Plaintiff, Defendant, Adjudicator, Crime, Place, Time
	Convict	Defendant, Adjudicator, Crime, Place, Time
	Sentence	Defendant, Adjudicator, Crime, Sentence, Place, Time
	Fine	Entity, Adjudicator, Money, Crime, Place, Time
	Execute	Person, Agent, Crime, Place, Time
	Extradite	Agent, Person, Destination, Origin, Crime, Time
	Acquit	Defendant, Adjudicator, Crime, Place, Time
	Appeal	Defendant, Prosecutor, Adjudicator, Crime, Place, Time
	Pardon	Defendant, Adjudicator, Crime, Place, Time

TABLE 3 – Liste des types d'événements du jeu de données ACE-2005.

Ensemble	Arguments
Entraînement	Org, Adjudicator, Prosecutor, Defendant, Sentence, Plaintiff, Attacker, Target, Crime
Évaluation	Buyer, Seller, Recipient, Artifact, Vehicle, Position, Victim
Partagés	Agent, Person, Time, Place, Giver, Money, Entity, Instrument, Destination, Origin

TABLE 4 – Répartition des rôles d'arguments entre l'ensemble d'entraînement et l'ensemble d'évaluation du jeu de données ACE-2005.

B Synthèse des hyperparamètres

Nous donnons la liste des hyperparamètres dans le Tableau 5 ci-dessous.

Paramètre	Valeur
encodeur	BERT-large-uncased
taille des séquences	128
nombre d'itérations d'entraînement	5 000
optimiseur	AdamW
taux d'apprentissage	
BERT	$1e-5$
autres	$1e-4$
weight decay	$1e-2$
dropout	$0,1$
warmup ratio	$0,1$
scheduler	stepLR
β_1 β_2	$0,9$ $0,999$
nombre de couches de convolution	2

TABLE 5 – Liste et valeurs des hyperparamètres.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Les petits modèles sont bons : une étude empirique de classification dans un contexte zero-shot

Pierre Lepagnol[1,2], Thomas Gerald[1], Sahar Ghannay[1], Christophe Servan[1,3], Sophie Rosset[1]
[1]Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique, 91405, Orsay, France
[2]SCIAM, 7508, Paris, France
[3]QWANT Research, 7 Rue spontini, 75116 Paris, France
{firstname.lastname}@lisn.upsaclay.fr

RÉSUMÉ

Ce travail s'inscrit dans le débat sur l'efficacité des grands modèles de langue par rapport aux petits pour la classification de texte par amorçage (prompting). Nous évaluons ici le potentiel des petits modèles de langue dans la classification de texte sans exemples, remettant en question la prédominance des grands modèles. À travers un ensemble diversifié de jeux de données, notre étude compare les petits et les grands modèles utilisant différentes architectures et données de pré-entraînement. Nos conclusions révèlent que les petits modèles peuvent générer efficacement des étiquettes et, dans certains contextes, rivaliser ou surpasser les performances de leurs homologues plus grands. Ce travail souligne l'idée que le modèle le plus grand n'est pas toujours le meilleur, suggérant que les petits modèles économes en ressources peuvent offrir des solutions viables pour des défis spécifiques de classification de données.

ABSTRACT

Here the title in English.

This study is part of the debate on the efficiency of large versus small language models for text classification by prompting. We assess the potential of small language models in zero-shot text classification, challenging the prevailing dominance of large models. Across a diverse set of datasets, our investigation benchmarks both small and large models using different architectures and pre-training data. Our findings reveal that small models can effectively generate labels and, in certain contexts, rival or surpass the performance of their larger counterparts. This research underscores the notion that bigger isn't always better, suggesting that resource-efficient small models may offer viable solutions for specific data classification challenges.

MOTS-CLÉS : Zero-shot, Prompting, amorçage, Modèle de langue, LLM, labelisation de données.

KEYWORDS: Zero-shot, Prompting, language modeling, LLMs, data labeling.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 113–129.

1 Introduction

Les grands modèles de langue (LLM) ont été largement favorisés par rapport aux modèles plus petits pour résoudre les tâches grâce aux méthodes par amorces (prompting) (Brown *et al.*, 2020; Hoffmann *et al.*, 2022; OpenAI, 2023; Chowdhery *et al.*, 2022) dans le contexte d'indisponibilité de données d'entraînement (zero-shot prompting).

Bien que les grand modèles soient très performants, leur utilisations présentent des difficultés - ils sont gourmands en ressources, coûteux à employer et leurs performances ne sont pas toujours garanties pour chaque tâche (Nityasya *et al.*, 2021).

De toujours plus grands modèles (Kaplan *et al.*, 2020; Hoffmann *et al.*, 2022) ont été construits et des jeux de données toujours plus sophistiqués ont été nécessaires (Zhang *et al.*, 2023) pour atteindre de bonnes performances. Leurs performances a priori supérieures en ont fait un choix privilégié pour diverses tâches, même pour les tâches de classification de base.

Au fur et à mesure que le domaine du traitement du langage évolue, nous devons nous poser la question suivante : les grands modèles de langue sont-ils essentiels pour une classification efficace des données ?

Dans cet article, nous examinons dans quelle mesure les petits modèles peuvent rivaliser avec les grands modèles dans la création d'étiquettes. Sur différents jeux de données, nous voulons voir comment les petits modèles peuvent correctement étiqueter des textes sans exemple (zero-shot classification). Nous cherchons aussi à déterminer ce qui permet aux modèles d'obtenir de bons résultats sur les tâches de classification. Nous comparons le fonctionnement des petits et des grands modèles dans le cadre d'amorçage sans exemples (prompting zero-shot) sur différents jeux de données afin de déterminer s'il est possible d'obtenir de bons résultats avec moins de ressources.

Nous pensons que cette étude est le point de départ de la compréhension des capacités réelles des LLM lorsqu'ils sont utilisés pour des tâches de classification dans un contexte zero-shot.

Nos contributions principales sont :

1. Une évaluation d'un grand nombre de modèles de langue (de 77 millions à 70 milliards de paramètres) ajustés sur des jeux de données d'instructions, avec différentes architectures (encodeur-décodeur ou décodeur seul) et tailles sur de 15 jeux de données dans un contexte zero-shot.
2. Des preuves relativement solides de l'efficacité des petits modèles dans la classification zero-shot, où les performances des petits modèles sont comparables à celles de plus grands sur de nombreux jeux de données dans les tâches de classification.
3. Nos évaluations sont mises à la disposition de la communauté en open-source, présentant les méthodologies proposées, contribuant ainsi à l'intégrité et à la robustesse des études dans ce domaine. Le code est disponible en ligne dans le dépôt XXXX.

L'article est organisé comme suit : La section 2 présente une revue de la littérature sur les approches zero-shot. Dans la section 3, nous décrivons la méthodologie que nous suivons pour cette étude. La section 4 présente les résultats et analyses. Enfin, nous concluons dans la section 5 et discutons des limitations et travaux futurs.

2 Travaux connexes

Classification de texte sans exemples & amorçage (Prompting) Le prompting consiste à fournir un texte d'entrée (ou prompt) à un modèle de langue, qui génère ensuite un texte de sortie basé sur ce prompt. L'objectif de la classification de texte sans exemples est de catégoriser des textes avec des étiquettes sans entraînement préalable spécifique à cette tâche. Cette approche attire l'attention du monde industriel et de la communauté scientifique car elle vise à supprimer le besoin d'ajustement supplémentaire et, par conséquent, de données étiquetées additionnelles, qui sont souvent onéreuses à obtenir.

Pour que le système, ici le modèle de langue, obtienne de bonnes performances sur les classes non vues, il est nécessaire d'avoir des descriptions précises des classes non-vues, comme l'ont noté Xia *et al.* (2018) et Liu *et al.* (2019a). Fei *et al.* (2022) améliorent la classification zero-shot en segmentant les textes d'entrée et en exploitant les amorces (prompts) spécifiques aux classes. Meng *et al.* (2020) ont proposé une stratégie qui utilise des noms d'étiquettes combinés à un auto-apprentissage adapté à la classification zero-shot. De nombreuses méthodes nécessitent un jeu de données non étiquetées ou une base de connaissances pour extraire les étiquettes pertinentes et faciliter l'auto-apprentissage.

Pour utiliser des modèles de langues et les méthodes d'amorçage (prompting) Schick & Schütze (2021) propose d'exploiter des paires schéma-verbalisateur (pattern-verbalizer pairs) où le schéma représente le prompt et le verbalisateur représente un mot, un token, par classe qui sera associé sur la dite classe. Néanmoins, ils utilisent cette méthode non pas dans un cadre zero-shot mais dans un cadre de classification avec ajustement.

Plus récemment, Zhao *et al.* (2023b) ont proposé d'utiliser k-Nearest-Neighbor fondé sur la similarité entre plongement des mots du verbalisateur pour augmenter les performances de classification. Lu *et al.* (2023) ont proposé la sélection par la perplexité pour sélectionner les meilleurs prompts dans un contexte d'essai à zero-shot.

Alors que les travaux antérieurs se sont concentrés sur de nouvelles méthodes visant à rendre les modèles de langues plus performants en matière de classification sans exemples, nous souhaitons avoir un aperçu des caractéristiques des modèles et de leurs performances.

3 Dispositif expérimental

Bien que les auteurs de LLMs aient comparé leurs différentes tailles de modèles (Kaplan *et al.*, 2020; Hoffmann *et al.*, 2022), cette étude élargit cette analyse en comparant directement différentes architectures sur un ensemble étendu de jeux de données. Nous amorçons (promptons) divers modèles de langue en utilisant 4 fonctions de scoring différentes (voir Section 3) pour classifier les phrases. Nous évaluons la qualité de nos classifieurs par l'exactitude et le macro F1 score.

Tâches & Jeux de Données Nous examinons les performances des modèles sur 15 jeux de données, sélectionnés pour représenter divers défis en classification. Par exemple nous utilisons les jeux de données *AGNews*, avec ses 4 classes distinctes, et *BBCNews*, qui propose 5 classes. Pour la classification de sentiments, la plupart des jeux de données proposent un choix binaire, comme pour *ethos* (Mollas *et al.*, 2022) ou plus granulaire comme *sst-5* (Socher *et al.*, 2013) avec 5 classes.

La tâche de classification de spams inclut les jeux de données *youtube* (Alberto *et al.*, 2015) ou *sms* (Almeida & Hidalgo, 2012). La tâches de classification de relations inclut les jeux de données tels que *semeval* (Hendrickx *et al.*, 2010). La liste complète des jeux de donnée est en annexe B (Tableau 7).

Les jeux de données sélectionnés sont équilibrés en termes de classes. On considère un jeu de données comme déséquilibré si la classe majoritaire est au moins 2 fois plus grande que la classe minoritaire. Ainsi nous avons choisi le macro F1-score pour les jeux de données déséquilibrés et l'exactitude pour le reste.

Modèles Notre étude évalue un total de 72 modèles uniques. Nous sélectionnons à la fois des modèles encodeur-décodeur (comme T5 (Raffel *et al.*, 2020), mT0 (Muennighoff *et al.*, 2023) et Bart (Lewis *et al.*, 2020)) et des modèles uniquement décodeur causal (tels que Llama (Touvron *et al.*, 2023) et Falcon (Penedo *et al.*, 2023)). Nous optons pour différentes tailles pour les mêmes modèles, allant de quelques millions à des centaines de milliards de paramètres. Par exemple, le modèle Bart possède 255M ou 561M de paramètres, Falcon existe en version 7B ou 40B [1]. Ces modèles ont été choisis en fonction de leur prévalence dans la littérature, de leur efficacité rapportée sur des tâches similaires et du fait que des versions adaptées aux instructions étaient disponibles pour certains d'entre eux. L'ajustement sur les instructions fait référence à la stratégie de fine-tuning d'un modèle de langue sur des jeux de données d'instructions (Longpre *et al.*, 2023).

La liste complète des modèles est en annexe A (Tableaux 5 et 6).

Amorces (Prompts) Les prompts de nos expériences sont issues de la littérature et conçu pour être simples et répondu par un seul token par le modèle de langue. Les amorces sont soit des traductions de fonctions d'étiquetage issues du benchmark WRENCH (Zhang *et al.*, 2021), soit créées de zéro dans le même style. Elles sont adaptées à chaque tâche, *par exemple* les amorces pour le jeu de données *sms* sont formulées différemment de celles pour le jeu de données *bbcnews*. Ceci permet d'assurer la pertinence par rapport au domaine et maximiser la compréhension du modèle. La liste des amorces par jeux de données est en annexe C, tableau 9.

Ainsi pour *sms*, le couple amorce/verbalisateur est

Amorce (Prompt)

```
Is the following message spam? Answer by yes or no.\n"{TEXT}"
```

Verbalisateur Texte/Classe

```
{1:"yes", 0:"no"}
```

Pour *bbcnews*, le couple amorce/verbalisateur est :

Amorce (Prompt)

```
"{TEXT}" is about "
```

Verbalisateur Texte/Classe

```
{0: "tech", 1: "business", 2: "sport", 3: "entertainment", 4: "politics"}
```

1. Nous n'avons pas testé Falcon 180B, car il n'était pas disponible pendant nos expériences

Scoring Functions Dans la classification fondée sur les amorces (prompts), l'utilisation d'un verbalisateur associant des tokens aux étiquettes de classe est cruciale pour une classification précise. Comme suggéré par (Holtzman *et al.*, 2022), de nombreuses séquences valides peuvent représenter le même concept, ceci est appelé *compétition pour la forme de surface*. Par exemple, `"+"`, `"positif"`, `"Plus positif que l'opposé"` pourraient être utilisés pour représenter le même concept de positivité pour la tâche d'analyse des sentiments. Comme cette compétition existe, la manière dont les verbalisateurs sont conçus influence grandement l'efficacité des approche par prompting pour la classification. Zhao *et al.* (2023b) utilisent le k-Plus Proches Voisins pour la construction de verbalisateur et augmentent leurs verbalisateurs fondés sur la similarité des embeddings.

Dans cette étude nous utilisons plusieurs fonctions de scoring pour évaluer leur impact sur les performances de nos modèles, dont celles proposées par (Holtzman *et al.*, 2022).

Probability	$\arg\max_i \mathbb{P}(y_i \mid x')$
DCPMI	$\arg\max_i \frac{\mathbb{P}(y_i \mid x')}{\mathbb{P}(y_i \mid x_{\text{domain\_conditional}})}$
PMI	$\arg\max_i \frac{\mathbb{P}(y_i \mid x')}{\mathbb{P}(y_i \mid x_{\text{domain\_unconditional}})}$
Similarity	$\arg\max_{c_i \in C} \cos(e(t_0), e(y_i))$ [2]

Outils pour l'Analyse Statistique Les trois principaux outils statistiques utilisés sont détaillés ci-après :

Le Biweight Midcorrelation Coefficient est une alternative robuste au coefficient de corrélation de Pearson pour quantifier l'association entre deux échantillons. Il est conçu pour être moins sensible aux valeurs aberrantes que d'autres coefficients tels que celui de Pearson.

Analyse de Covariance - ANCOVA combine les techniques d'ANOVA et de régression pour évaluer si les moyennes d'une variable dépendante sont égales à travers les niveaux d'une variable indépendante catégorielle tout en contrôlant statistiquement pour les effets d'autres variables continues (covariables).

Test de Kruskal-Wallis est une méthode non paramétrique pour tester si des échantillons proviennent de la même distribution. Nous l'avons utilisé comme une méthode non paramétrique, qui ne suppose pas une distribution normale des résidus, contrairement à l'analyse de variance standard à un facteur.

4 Résultats

Sur les jeux de données mentionnés précédemment, nous comparons la performance des modèles de langue et nous étudions : 1) la relation entre les performances des modèles et leurs tailles (le nombre de paramètres), 2) les performances et leurs architectures, et 3) les performances et si le modèle a été fine-tuné sur des jeux de données d'instructions. Ensuite, pour les deux types d'architectures (encodeur-décodeur et décodeur seul), nous étudions l'impact de l'ajustement sur des instructions.

Le tableau 1 présente les scores de l'état de l'art pour chaque jeu de données [3].

3. Les jeux de données *agnews*, *imdb*, *yelp*, *trec* sont inclus dans l'entraiement du modèle `mT0`. Nous ne considérons donc pas ses scores sur ces jeux de données.

dataset	SOTA	Classe Maj.	Meilleur Score	Modèle	# Paramètres
agnews	0.625	0.266	**0.734**	MBZUAI/LaMini-GPT-124M	163.0 M
bbcnews	NaN	0.236	0.869	bigscience/mt0-large	1.2 B
cdr	NaN	0.676	0.717	bigscience/bloomz-3b	3.6 B
chemprot	0.172	0.049	**0.192**	bigscience/bloomz-3b	3.6 B
ethos	0.667	0.566	0.597	bigscience/bloomz-1b1	1.5 B
financial_phrasebank	0.528	0.254	**0.744**	MBZUAI/LaMini-GPT-774M	838.4 M
imdb	0.718	0.500	**0.933**	MBZUAI/LaMini-Flan-T5-783M	783.2 M
semeval	0.435	0.054	0.270	bigscience/mt0-xxl	12.9 B
sms	0.340	0.464	**0.699**	mosaicml/mpt-7b	6.6 B
spouse	0.630	0.479	0.521	gpt2	163.0 M
sst-2	0.710	0.501	**0.956**	bigscience/bloomz-3b	3.6 B
sst-5	0.598	0.286	0.485	tiiuae/falcon-40b-instruct	41.8 B
trec	NaN	0.072	0.324	mosaicml/mpt-7b-instruct	6.6 B
yelp	0.888	0.522	**0.977**	MBZUAI/LaMini-Flan-T5-783M	783.2 M
youtube	0.468	0.528	**0.716**	tiiuae/falcon-40b	41.8 B

TABLE 1 – Tableau illustrant les mesures de performance pour différents jeux de données :
Les colonnes sont (1) le nom de l'ensemble de données, (2) les scores de l'état de l'art (SOTA), (3) les scores obtenus en prédisant systématiquement la classe majoritaire, (4) les scores les plus élevés (surlignés en rouge lorsqu'ils sont meilleurs), (5) les modèles ayant ces meilleurs scores, et (6) le nombre de paramètres pour chaque modèle.
Notez la présence d'entrées `NaN`, signifiant des jeux de données pour lesquels les références SOTA n'ont pas été établies ou trouvées.

4.1 La taille du modèle n'a pas vraiment d'importance

La Figure 1 présente la relation entre le nombre de paramètres et la performance en termes de scores Acc/F1 à travers divers jeux de données. Nous calculons le Biweight Midcorrelation Coefficient et les p-valeurs associées pour chaque jeu de données. Ces résultats sont détaillés dans le tableau 2.

dataset	correlation coef	pvalue
agnews	-0.1418	**0.0536**
bbcnews	0.0489	0.4877
cdr	0.2541	**0.0002**
chemprot	0.1318	**0.0531**
ethos	-0.1519	**0.0256**
financial_phrasebank	0.0419	0.5406
imdb	-0.2862	**0.0001**
semeval	-0.0506	0.4595
sms	-0.1209	0.0763
spouse	-0.0254	0.7106
sst-2	0.0755	0.2693
sst-5	0.0061	0.9293
trec	-0.1085	0.1403
yelp	-0.0620	0.4008
youtube	-0.0014	0.9836

TABLE 2 – Biweight Midcorrelation Coefficients et p-valeurs mesurant la relation entre le score et la taille du modèle (log du nombre de paramètres) selon les datasets

D'après notre analyse, 10 des 15 jeux de données présentent des p-valeurs supérieures à 0,05, ce

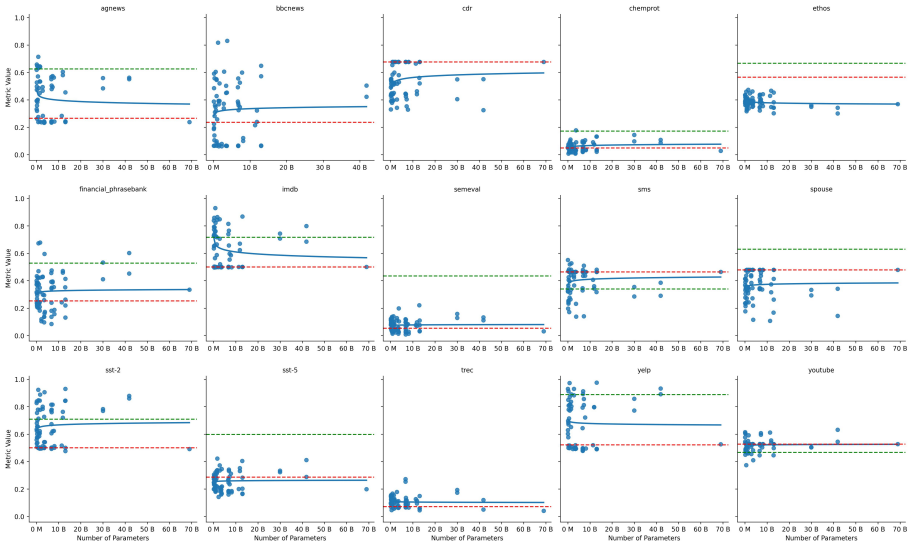


FIGURE 1 – Comparaison de la performance de différentes tailles de modèles à travers les jeux de données. Les ligne en bleu indiquent la tendance générale, les lignes pointillées en rouge indiquent pour les résultats des prédictions de la classe majoritaire, et la ligne verte indique les résultats état de l'art pour les méthodes de prompting zero-shot.

qui suggère qu'il n'y a pas de lien significatif entre les scores de performance et la taille du modèle. Cependant, trois jeux de données présentent des p-valeurs inférieures à 0,05, ce qui indique une corrélation notable. Parmi ceux-ci, la corrélation est positive pour le jeu de données *cdr* mais négative pour *ethos* et *imdb*. Deux jeux de données, à savoir *agnews* et *chemprot*, présentent des p-valeurs proches du seuil de 0,05, ce qui rend leur corrélation peu concluante.

En conclusion, alors que de nombreux jeux de données ne montrent pas de relation directe entre des tailles de modèle plus grandes et une amélioration des performances, des jeux de données comme *cdr*, *ethos*, et *imdb* le font. De plus, la variance du coefficient de corrélation entre les jeux de données suggère que la taille du modèle n'est pas le seul facteur déterminant des performances.

4.2 Impact de l'architecture sur les performances

La figure 2 illustre les variations de performances entre les architectures encodeur-décodeur et décodeur seul. En utilisant l'ANCOVA, nous mesurons l'impact du choix de l'architecture sur les scores de performance, tout en contrôlant l'effet de la taille du modèle. Les résultats sont présentés dans le tableau 3.

D'une part, 7 des 15 jeux de données, nommément *agnews*, *bbcnews*, *chemprot*, *semeval*, *sms*, *spouse* et *youtube*, présentent des p-valeurs inférieures à 0,05, ce qui suggère que l'architecture a un impact significatif. En revanche, les jeux de données tels que *cdr*, *ethos* et *financial_phrasebank* ne sont pas affectés par le choix de l'architecture. Le jeux de données *imdb* présente un impact non concluant. En conclusion, bien que la taille du modèle ne soit pas un facteur dominant, le choix de l'architecture a un impact significatif sur les performances dans ces jeux de données spécifiques.

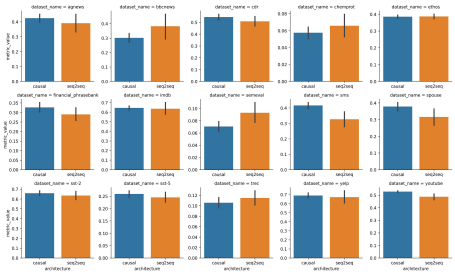


FIGURE 2 – Variation des performances entre les différentes architectures.

Dataset	statistique	p-valeur	Variances Egales
agnews	4.0676	**0.0452**	True
bbcnews	7.0640	**0.0085**	False
cdr	0.2519	0.6163	True
chemprot	4.4883	**0.0353**	True
ethos	0.3945	0.5306	False
financial_phrasebank	1.4592	0.2284	False
imdb	3.6687	0.0570	True
semeval	8.2301	**0.0045**	True
sms	11.9951	**0.0006**	False
spouse	4.7794	**0.0299**	True
sst-2	0.2501	0.6175	True
sst-5	0.7852	0.3766	True
trec	0.3382	0.5616	False
yelp	0.7103	0.4004	True
youtube	18.0011	**0.0000**	False

FIGURE 3 – ANCOVA indiquant l'impact de l'architecture sur les scores selon le dataset

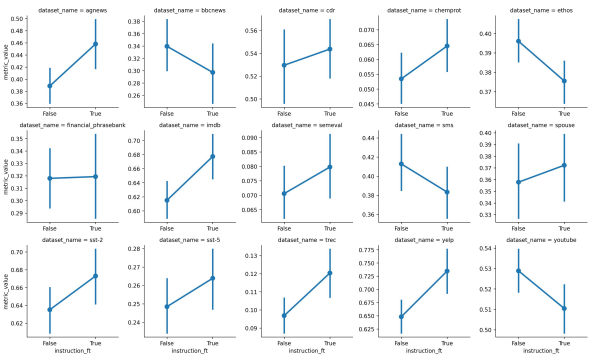


FIGURE 4 – Comparaison des performances des modèles avec ou sans instruction sur l'ensemble des données.

dataset	statistique	p-valeur	Variances Egales
agnews	10.5411	**0.0014**	True
bbcnews	1.9492	0.1642	True
cdr	0.1635	0.6864	True
chemprot	2.3152	0.1296	True
ethos	5.8015	**0.0169**	True
financial_phrasebank	0.0001	0.9917	False
imdb	13.6945	**0.0003**	True
semeval	1.4016	0.2378	False
sms	2.6667	0.1039	True
spouse	0.3379	0.5617	True
sst-2	3.0055	0.0844	False
sst-5	1.8271	0.1779	True
trec	8.3534	**0.0043**	False
yelp	12.5571	**0.0005**	True
youtube	5.8369	**0.0165**	True

FIGURE 5 – ANCOVA indiquant l'impact de l'instruction fine-tunning sur les scores selon le dataset

4.3 Impact de l'Instruction Fine-tuning sur les performances

De la même manière que pour l'architecture, nous avons quantifié l'impact de l'ajustement sur des instructions sur les performances tout en contrôlant le nombre de paramètres. Nous avons utilisé l'ANCOVA pour tester si les moyennes de nos scores ACC/F1 sont égales pour toutes les catégories de la variable `instruction_ft`, tout en contrôlant statistiquement l'effet du nombre de paramètres. Les résultats sont présentés dans le tableau 5.

La figure 4 montre l'impact de l'ajustement sur des instructions sur les scores de performance selon les datasets. L'axe des y de chaque graphique affiche le score de performance (Acc/F1). L'axe des x a deux valeurs : `False` et `True`, indiquant si le modèle a été ajusté sur des instructions ou non.

Pour de nombreux jeux de données, le fine-tuning sur les instructions améliore les performances par rapport à l'absence de fine-tuning (*agnews*, *ethos*, *imdb*, *trec*, *yelp*, et *youtube*) comme le suggère les p-valeurs significatives de l'ANCOVA. Une diminution des perfomances semble survenir lorsque les modèles sont ajustés sur des instructions pour certains jeux de données. Pour *bbcnews*, *youtube* et

sms, l'ANCOVA nous indique que cette diminution n'est pas significative, en revanche, pour *ethos*, elle est significative.

Pour les autres jeux de données, bien qu'il puisse y avoir des différences visuelles dans les performances avec et sans fine-tuning sur les instructions, ces différences ne sont pas statistiquement significatives d'après les p-valeurs. Par conséquent, bien que le fine-tuning sur les instructions ait le potentiel d'améliorer les performances des modèles sur de nombreux jeux de données, son impact peut varier en fonction des jeux de données spécifiques.

4.4 Relation entre la taille du modèle et les performances par architecture

Le tableau 3 présente les coefficients de corrélation moyenne pondérée entre les tailles de modèle (log du nombre de paramètres) et les scores de performance pour les deux types architectures étudiées. On peut y retrouver une corrélation légère mais significative pour les modèles de décodeurs, mais largement non-significative pour les modèles de encodeur-décodeur. Cela suggère que le décodeur seul pourrait être plus sensible au nombre de paramètres ; un trop grand nombre de paramètres pourrait nuire aux performances.

dataset	correlation coef	pvalue
causal	-0.0435	**0.0299**
seq2seq	0.0065	0.8728

TABLE 3 – Biweight Midcorrelation Coefficients et p-valeurs mesurant la relation entre le score et la taille du modèle (log du nombre de paramètres) selon les architectures

4.5 Impact de l'ajustement sur des instructions et des performances par architecture

La figure 6 compare visuellement l'impact de l'ajustement sur des instructions et les scores de performance pour les deux architectures. En ordonnée est le score de performance (Acc/F1), en abscisse une variable à deux modalité indiquant si le modèle a été ajusté sur des instructions.

Une ANCOVA est réalisée pour quantifier l'impact de l'ajustement sur des instructions sur chaque architecture (encodeur-décodeur/décodeur-seul) tout en contrôlant l'effet de la taille du modèle. Le tableau 4 présente les statistiques et les p-valeurs.

Pour l'architecture décodeur-seul, il n'y a pas d'impact significatif de l'ajustement sur des instructions sur les scores. La p-valeur est ici de 0,6693, bien supérieure à 0,05. Pour l'architecture encodeur-décodeurs, il y a un impact significatif de l'ajustement sur des instructions sur les scores. La p-valeur pour l'architecture encodeur-décodeur est surlignée en rouge et s'élève à 0,0086, soit moins de 0,05.

La différence de résultats entre les deux architectures suggère que l'impact de l'ajustement sur des instructions pourrait dépendre de l'architecture. Tant l'analyse graphique que l'ANCOVA montrent un effet de l'ajustement sur des instructions sur l'architecture encodeur-décodeur.

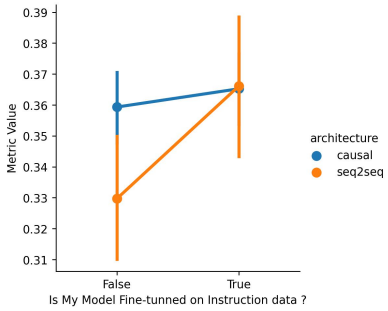


FIGURE 6 – Comparaison des performances entre les modèles ajustés par instruction et les modèles non ajustés, selon l'architecture

dataset	statistique	p-valeur	Variances Egales
causal	0.1825	0.6693	True
seq2seq	6.9406	**0.0086**	False

TABLE 4 – "ANCOVA indiquant l'impact de `instruction_ft` sur Acc/F1 selon les architectures

5 Conclusion & Perspectives

Ce travail avait pour but de déterminer si l'utilisation de grands modèles était nécessaire pour aborder les tâches de classification en utilisant des techniques de prompting.

La performance des modèles de langue varie en fonction de multiples facteurs, y compris la taille du modèle, les choix architecturaux et les stratégies de fine-tuning. Si un modèle plus grand entraîne une amélioration des performances, ce n'est pas un facteur clef, le choix d'architecture du modèle a un impact plus notable sur les résultats obtenus sur nos jeux de données. L'impact de l'ajustement sur des instructions est également évident, mais son efficacité dépend de l'architecture. Une étude complète d'autres architectures émergentes, telles que l'architecture RWKV (Peng *et al.*, 2023) ou les modèles fondés sur les states spaces models, pourrait apporter des nuances et des détails à cette analyse. L'impact varié de l'ajustement sur des instructions à travers les jeux de données suggère le besoin de techniques de fine-tuning plus avancées comme l'incorporation de recherche d'informations pour assurer de meilleures performances de classification lors de l'armocage.

6 Limitations

Dans cette étude, nous avons limité notre évaluation à des prompts simples, non optimisés pour obtenir les meilleurs réponses et nous n'avons pas étudié la variablité des résulats pour différents prompts (test d'un seul prompt pour pour chaque jeu de données). De plus, nous avons concentré notre étude sur les modèles encodeur-décodeur et décodeurs-seul sans les comparer avec des modèles encodeurs-seul. Nous n'avons pas étudié pas la sensibilité des performances à certains facteurs externes tels que le temps de pré-entraînement, la qualité des données de pré-entraînement ou les biais potentiels dans les jeux de données. Ces facteurs externes pourraient influencer les résultats ou le caractère universel des conclusions. Le choix et les hypothèses des outils statistiques pourraient influencer les résultats. Cette étude n'inclut pas les modèles publiés très récemment. Ainsi le comportement de modèles très récents comme les modèles conversationnels entraînés avec du RLHF/DPO pourrait exhiber des différences dans nos conclusions.

Références

ALBERTO T. C., LOCHTER J. V. & ALMEIDA T. A. (2015). TubeSpam : Comment Spam Filtering on YouTube. *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, p. 138–143. DOI : 10.1109/ICMLA.2015.37.

ALMEIDA T. & HIDALGO J. (2012). SMS Spam Collection. UCI Machine Learning Repository. DOI : https ://doi.org/10.24432/C5CC84.

BIDERMAN S., SCHOELKOPF H., ANTHONY Q., BRADLEY H., O'BRIEN K., HALLAHAN E., KHAN M. A., PUROHIT S., PRASHANTH U. S., RAFF E., SKOWRON A., SUTAWIKA L. & VAN DER WAL O. (2023). *Pythia : A Suite for Analyzing Large Language Models Across Training and Scaling*. Rapport interne. arXiv :2304.01373 [cs] type : article, DOI : 10.48550/arXiv.2304.01373.

BIGSCIENCE WORKSHOP (2022). BLOOM (revision 4ab0472). DOI : 10.57967/hf/0003.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I. & AMODEI D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, p. 1877–1901 : Curran Associates, Inc.

CHO H., KIM Y. & LEE S.-G. (2023). *CELDA : Leveraging Black-box Language Model as Enhanced Classifier without Labels*. Rapport interne. arXiv :2306.02693 [cs] type : article.

CHOWDHERY A., NARANG S., DEVLIN J., BOSMA M., MISHRA G., ROBERTS A., BARHAM P., CHUNG H. W., SUTTON C., GEHRMANN S., SCHUH P., SHI K., TSVYASHCHENKO S., MAYNEZ J., RAO A., BARNES P., TAY Y., SHAZEER N., PRABHAKARAN V., REIF E., DU N., HUTCHINSON B., POPE R., BRADBURY J., AUSTIN J., ISARD M., GUR-ARI G., YIN P., DUKE T., LEVSKAYA A., GHEMAWAT S., DEV S., MICHALEWSKI H., GARCIA X., MISRA V., ROBINSON K., FEDUS L., ZHOU D., IPPOLITO D., LUAN D., LIM H., ZOPH B., SPIRIDONOV A., SEPASSI R., DOHAN D., AGRAWAL S., OMERNICK M., DAI A. M., PILLAI T. S., PELLAT M., LEWKOWYCZ A., MOREIRA E., CHILD R., POLOZOV O., LEE K., ZHOU Z., WANG X., SAETA B., DIAZ M., FIRAT O., CATASTA M., WEI J., MEIER-HELLSTERN K., ECK D., DEAN J., PETROV S. & FIEDEL N. (2022). *PaLM : Scaling Language Modeling with Pathways*. Rapport interne. arXiv :2204.02311 [cs] type : article.

CHUNG H. W., HOU L., LONGPRE S., ZOPH B., TAY Y., FEDUS W., LI Y., WANG X., DEHGHANI M., BRAHMA S., WEBSON A., GU S. S., DAI Z., SUZGUN M., CHEN X., CHOWDHERY A., CASTRO-ROS A., PELLAT M., ROBINSON K., VALTER D., NARANG S., MISHRA G., YU A., ZHAO V., HUANG Y., DAI A., YU H., PETROV S., CHI E. H., DEAN J., DEVLIN J., ROBERTS A., ZHOU D., LE Q. V. & WEI J. (2022). *Scaling Instruction-Finetuned Language Models*. Rapport interne. arXiv :2210.11416 [cs] type : article.

CLARKE C., HENG Y., KANG Y., FLAUTNER K., TANG L. & MARS J. (2023). Label Agnostic Pre-training for Zero-shot Text Classification. In *Findings of the Association for Computational Linguistics : ACL 2023*, p. 1009–1021, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.64.

DAVIS A. P., GRONDIN C. J., JOHNSON R. J., SCIAKY D., KING B. L., MCMORRAN R., WIEGERS J., WIEGERS T. C. & MATTINGLY C. J. (2016). The comparative toxicogenomics database : update 2017. *Nucleic Acids Res*, **45**(D1), D972–D978.

DEY N., GOSAL G., ZHIMING, CHEN, KHACHANE H., MARSHALL W., PATHRIA R., TOM M. & HESTNESS J. (2023). *Cerebras-GPT : Open Compute-Optimal Language Models Trained on the Cerebras Wafer-Scale Cluster*. Rapport interne. arXiv :2304.03208 [cs] type : article.

FEI Y., MENG Z., NIE P., WATTENHOFER R. & SACHAN M. (2022). Beyond prompting : Making Pre-trained Language Models Better Zero-shot Learners by Clustering Representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 8560–8579, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.587.

HENDRICKX I., KIM S. N., KOZAREVA Z., NAKOV P., Ó SÉAGHDHA D., PADÓ S., PENNACCHIOTTI M., ROMANO L. & SZPAKOWICZ S. (2010). SemEval-2010 task 8 : Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, p. 33–38, Uppsala, Sweden : Association for Computational Linguistics.

HOFFMANN J., BORGEAUD S., MENSCH A., BUCHATSKAYA E., CAI T., RUTHERFORD E., CASAS D. D. L., HENDRICKS L. A., WELBL J., CLARK A., HENNIGAN T., NOLAND E., MILLICAN K., DRIESSCHE G. V. D., DAMOC B., GUY A., OSINDERO S., SIMONYAN K., ELSEN E., RAE J. W., VINYALS O. & SIFRE L. (2022). *Training Compute-Optimal Large Language Models*. Rapport interne. arXiv :2203.15556 [cs] type : article.

HOLTZMAN A., WEST P., SHWARTZ V., CHOI Y. & ZETTLEMOYER L. (2022). *Surface Form Competition : Why the Highest Probability Answer Isn't Always Right*. Rapport interne. arXiv :2104.08315 [cs] type : article, DOI : 10.48550/arXiv.2104.08315.

HSIEH C.-Y., LI C.-L., YEH C.-K., NAKHOST H., FUJII Y., RATNER A., KRISHNA R., LEE C.-Y. & PFISTER T. (2023). *Distilling Step-by-Step ! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes*. Rapport interne. arXiv :2305.02301 [cs] type : article.

KAPLAN J., MCCANDLISH S., HENIGHAN T., BROWN T. B., CHESS B., CHILD R., GRAY S., RADFORD A., WU J. & AMODEI D. (2020). *Scaling Laws for Neural Language Models*. Rapport interne. arXiv :2001.08361 [cs, stat] type : article, DOI : 10.48550/arXiv.2001.08361.

KRALLINGER M., RABAL O., AKHONDI S., PÉREZ M., SANTAMARÍA J., RODRÍGUEZ G. P., TSATSARONIS G., INTXAURRONDO A., LÓPEZ J. A. B., NANDAL U., BUEL E. V., CHANDRASEKHAR A., RODENBURG M., LÆGREID A., DOORNENBAL M. A., OYARZÁBAL J., LOURENÇO A. & VALENCIA A. (2017). Overview of the BioCreative VI chemical-protein interaction Track.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). BART : Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7871–7880, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.703.

LIU H., ZHANG X., FAN L., FU X., LI Q., WU X.-M. & LAM A. Y. (2019a). Reconstructing Capsule Networks for Zero-shot Intent Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 4799–4809, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1486.

LIU P., YUAN W., FU J., JIANG Z., HAYASHI H. & NEUBIG G. (2021). *Pre-train, Prompt, and Predict : A Systematic Survey of Prompting Methods in Natural Language Processing*. Rapport interne. arXiv :2107.13586 [cs] type : article.

LIU T. & LOW B. K. H. (2023). *Goat : Fine-tuned LLaMA Outperforms GPT-4 on Arithmetic Tasks*. Rapport interne. arXiv :2305.14201 [cs] type : article.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019b). Roberta : A robustly optimized bert pretraining approach.

LONGPRE S., HOU L., VU T., WEBSON A., CHUNG H. W., TAY Y., ZHOU D., LE Q. V., ZOPH B., WEI J. & ROBERTS A. (2023). The flan collection : Designing data and methods for effective instruction tuning.

LOSHCHILOV I. & HUTTER F. (2019). *Decoupled Weight Decay Regularization*. Rapport interne. arXiv :1711.05101 [cs, math] version : 3 type : article.

LU J., ZHU D., HAN W., ZHAO R., MAC NAMEE B. & TAN F. (2023). What makes pre-trained language models better zero-shot learners ? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2288–2303, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.128.

LUDAN J. M., MENG Y., NGUYEN T., SHAH S., LYU Q., APIDIANAKI M. & CALLISON-BURCH C. (2023). *Explanation-based Finetuning Makes Models More Robust to Spurious Cues*. Rapport interne. arXiv :2305.04990 [cs] version : 2 type : article.

MAAS A. L., DALY R. E., PHAM P. T., HUANG D., NG A. Y. & POTTS C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, p. 142–150, USA : Association for Computational Linguistics.

MALO P., SINHA A., KORHONEN P., WALLENIUS J. & TAKALA P. (2014). Good debt or bad debt : Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, **65**.

MENG Y., ZHANG Y., HUANG J., XIONG C., JI H., ZHANG C. & HAN J. (2020). Text Classification Using Label Names Only : A Language Model Self-Training Approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 9006–9017, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.724.

MOLLAS I., CHRYSOPOULOU Z., KARLOS S. & TSOUMAKAS G. (2022). ETHOS : an Online Hate Speech Detection Dataset. *Complex & Intelligent Systems*, **8**(6), 4663–4678. arXiv :2006.08328 [cs, stat], DOI : 10.1007/s40747-021-00608-2.

MOSQUERA A. (2022). Tackling Data Drift with Adversarial Validation : An Application for German Text Complexity Estimation. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, p. 39–44, Potsdam, Germany : Association for Computational Linguistics.

MUENNIGHOFF N., WANG T., SUTAWIKA L., ROBERTS A., BIDERMAN S., SCAO T. L., BARI M. S., SHEN S., YONG Z.-X., SCHOELKOPF H., TANG X., RADEV D., AJI A. F., ALMUBARAK K., ALBANIE S., ALYAFEAI Z., WEBSON A., RAFF E. & RAFFEL C. (2023). *Crosslingual Generalization through Multitask Finetuning*. Rapport interne. arXiv :2211.01786 [cs] type : article, DOI : 10.48550/arXiv.2211.01786.

NITYASYA M. N., WIBOWO H. A., PRASOJO R. E. & AJI A. F. (2021). *Costs to Consider in Adopting NLP for Your Business*. Rapport interne. arXiv :2012.08958 [cs] type : article.

OPENAI (2023). *GPT-4 Technical Report*. Rapport interne. arXiv :2303.08774 [cs] type : article, DOI : 10.48550/arXiv.2303.08774.

OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C. L., MISHKIN P., ZHANG C., AGARWAL S., SLAMA K., RAY A., SCHULMAN J., HILTON J., KELTON F., MILLER L., SIMENS M., ASKELL A., WELINDER P., CHRISTIANO P., LEIKE J. & LOWE R. (2022). *Training language models to follow instructions with human feedback*. Rapport interne. arXiv :2203.02155 [cs] type : article.

PENEDO G., MALARTIC Q., HESSLOW D., COJOCARU R., CAPPELLI A., ALOBEIDLI H., PANNIER B., ALMAZROUEI E. & LAUNAY J. (2023). *The RefinedWeb Dataset for Falcon LLM : Outperforming Curated Corpora with Web Data, and Web Data Only*. Rapport interne. arXiv :2306.01116 [cs] type : article, DOI : 10.48550/arXiv.2306.01116.

PENG B., ALCAIDE E., ANTHONY Q., ALBALAK A., ARCADINHO S., CAO H., CHENG X., CHUNG M., GRELLA M., GV K. K., HE X., HOU H., KAZIENKO P., KOCON J., KONG J., KOPTYRA B., LAU H., MANTRI K. S. I., MOM F., SAITO A., TANG X., WANG B., WIND J. S., WOZNIAK S., ZHANG R., ZHANG Z., ZHAO Q., ZHOU P., ZHU J. & ZHU R.-J. (2023). Rwkv : Reinventing rnns for the transformer era.

PILÁN I. & VOLODINA E. (2018). Investigating the importance of linguistic complexity features across different datasets related to language learning. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, p. 49–58, Santa Fe, New-Mexico : Association for Computational Linguistics.

QIN C., ZHANG A., ZHANG Z., CHEN J., YASUNAGA M. & YANG D. (2023). *Is ChatGPT a General-Purpose Natural Language Processing Task Solver ?* Rapport interne. arXiv :2302.06476 [cs] type : article.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D. & SUTSKEVER I. (2019). Language models are unsupervised multitask learners.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Rapport interne. arXiv :1910.10683 [cs, stat] type : article.

SANH V., WEBSON A., RAFFEL C., BACH S. H., SUTAWIKA L., ALYAFEAI Z., CHAFFIN A., STIEGLER A., SCAO T. L., RAJA A., DEY M., BARI M. S., XU C., THAKKER U., SHARMA S. S., SZCZECHLA E., KIM T., CHHABLANI G., NAYAK N., DATTA D., CHANG J., JIANG M. T.-J., WANG H., MANICA M., SHEN S., YONG Z. X., PANDEY H., BAWDEN R., WANG T., NEERAJ T., ROZEN J., SHARMA A., SANTILLI A., FEVRY T., FRIES J. A., TEEHAN R., BERS T., BIDERMAN S., GAO L., WOLF T. & RUSH A. M. (2022). *Multitask Prompted Training Enables Zero-Shot Task Generalization*. Rapport interne. arXiv :2110.08207 [cs] type : article.

SCHICK T. & SCHÜTZE H. (2021). *Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference*. Rapport interne. arXiv :2001.07676 [cs] type : article.

SMITH R., FRIES J. A., HANCOCK B. & BACH S. H. (2022). *Language Models in the Loop : Incorporating Prompting into Weak Supervision*. Rapport interne. arXiv :2205.02318 [cs] type : article.

SOCHER R., PERELYGIN A., WU J., CHUANG J., MANNING C. D., NG A. & POTTS C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 1631–1642, Seattle, Washington, USA : Association for Computational Linguistics.

SUN Y., DONG L., HUANG S., MA S., XIA Y., XUE J., WANG J. & WEI F. (2023). Retentive network : A successor to transformer for large language models.

TAY Y., DEHGHANI M., TRAN V. Q., GARCIA X., WEI J., WANG X., CHUNG H. W., SHAKERI S., BAHRI D., SCHUSTER T., ZHENG H. S., ZHOU D., HOULSBY N. & METZLER D. (2023). *UL2 : Unifying Language Learning Paradigms*. Rapport interne. arXiv :2205.05131 [cs] type : article.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F., RODRIGUEZ A., JOULIN A., GRAVE E. & LAMPLE G. (2023). Llama : Open and efficient foundation language models.

WANG T., ROBERTS A., HESSLOW D., SCAO T. L., CHUNG H. W., BELTAGY I., LAUNAY J. & RAFFEL C. (2022). *What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization ?* Rapport interne. arXiv :2204.05832 [cs, stat] type : article.

WANG Y., YU Z., ZENG Z., YANG L., WANG C., CHEN H., JIANG C., XIE R., WANG J., XIE X., YE W., ZHANG S. & ZHANG Y. (2023). *PandaLM : An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization*. Rapport interne. arXiv :2306.05087 [cs] type : article.

WEI J., BOSMA M., ZHAO V., GUU K., YU A. W., LESTER B., DU N., DAI A. M. & LE Q. V. (2022). Finetuned Language Models are Zero-Shot Learners.

WU M., WAHEED A., ZHANG C., ABDUL-MAGEED M. & AJI A. F. (2023). Lamini-lm : A diverse herd of distilled models from large-scale instructions. *CoRR*, **abs/2304.14402**.

XIA C., ZHANG C., YAN X., CHANG Y. & YU P. S. (2018). *Zero-shot User Intent Detection via Capsule Neural Networks*. Rapport interne. arXiv :1809.00385 [cs] type : article, DOI : 10.48550/arXiv.1809.00385.

YEH H.-S., LAVERGNE T. & ZWEIGENBAUM P. (2022). Decorate the Examples : A Simple Method of Prompt Design for Biomedical Relation Extraction. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 3780–3787, Marseille, France : European Language Resources Association.

ZHANG J., YU Y., LI Y., WANG Y., YANG Y., YANG M. & RATNER A. J. (2021). WRENCH : A Comprehensive Benchmark for Weak Supervision. *ArXiv*.

ZHANG S., DONG L., LI X., ZHANG S., SUN X., WANG S., LI J., HU R., ZHANG T., WU F. & WANG G. (2023). *Instruction Tuning for Large Language Models : A Survey*. Rapport interne. arXiv :2308.10792 [cs] type : article.

ZHAO W. X., ZHOU K., LI J., TANG T., WANG X., HOU Y., MIN Y., ZHANG B., ZHANG J., DONG Z., DU Y., YANG C., CHEN Y., CHEN Z., JIANG J., REN R., LI Y., TANG X., LIU Z., LIU P., NIE J.-Y. & WEN J.-R. (2023a). *A Survey of Large Language Models*. Rapport interne. arXiv :2303.18223 [cs] type : article.

ZHAO X., OUYANG S., YU Z., WU M. & LI L. (2023b). Pre-trained language models can be fully zero-shot learners. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 15590–15606, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.869.

A Models

Model	Number of Parameters	Instruction-Tunned
bigscience/bloom (**?**)	560M, 1B1, 1B7, 3B, 7B1	No
bigscience/bloomz (Muennighoff *et al.*, 2023)	560M, 1B1, 1B7, 3B, 7B1	Yes
tiiuae/falcon	7B, 40B	Yes/No
tiiuae/falcon-rw	7B, 40B	No
MBZUAI/LaMini-Cerebras (Wu *et al.*, 2023)	111M, 256M, 590M, 1.3B	Yes
MBZUAI/LaMini-GPT (Wu *et al.*, 2023)	124M, 774M, 1.5B	Yes
mosaicml/mpt	7B 30b	Yes/No
databricks/dolly-v2	3b, 7B, 12b	Yes
EleutherAI/pythia (Biderman *et al.*, 2023)	70M, 160M, 410M, 1B, 1.4B, 2.8, 6.9B, 12B	No
openlm-research/open_llama	3B 7B 13B	No
openlm-research/open_llama_v2	3B 7B	No
pankajmathur/orca_dolly	3B	Yes
pankajmathur/orca_alpaca	3B	Yes
pankajmathur/orca_mini	7B, 3B, 13B	Yes
pankajmathur/orca_mini_v2	7B, 13B	Yes
pankajmathur/orca_mini_v3	7B, 13B	Yes

TABLE 5 – Decoder Only Models

Model	Number of Parameters	Instruction-Tunned
MBZUAI/LaMini-Flan-T5 (Wu *et al.*, 2023)	77M, 248M, 783M	Yes
T5 vanilla (Raffel *et al.*, 2020)	77M, 248M, 770M, 3B, 11B	No
bigscience/mt0 (Muennighoff *et al.*, 2023)	300M, 582, 1.2B, 3.8B, 13B	Yes
Bart (Lewis *et al.*, 2020)	255M, 561M	No

TABLE 6 – Encoder-Decoder Only Models

B Datasets

Datasets	Tasks	#Classes	#Test Examples	Balance ratios
AGNews	Topic Classification	4	12000	0.897
BBCNews	Topic Classification	5	2000	0.742
CDR bio	Relation Classification	2	4673	0.478
Chemprot	Chemical Relation Classification	10	1607	0.004
ETHOS	Sentiment Classification	2	998	0.766
financial_phrasebank	Topic Classification	3	2264	0.218
IMDB	Sentiment Classification	2	2500	1.000
SemEval	Relation Classification	9	600	0.042
SMS	Spam Classification	2	500	0.155
Spouse	Relation Classification	2	2701	0.088
SST2	Sentiment Classification	2	1821	0.997
SST5	Sentiment Classification	5	2210	0.441
TREC	Question Classification	6	500	0.065
Yelp	Sentiment Classification	2	3800	0.915
Youtube	Spam Classification	2	250	0.894

TABLE 7 – Descriptions des jeux de données

C Prompts

TABLE 8 – Prompt used

dataset	prompts
sms	`Is the following message spam? Answer by yes or no.\n"TEXT"`
youtube	`Is the following comment spam? Answer by yes or no.\n"TEXT"`
spouse	`Context: "TEXT"\n\nAre ENTITY2 and ENTITY1 married? Answer by yes or`
cdr	`Context: "TEXT"\n\nDoes ENTITY1 induce ENTITY2 ? Answer by yes or no.`
chemprot	`Context: "TEXT"\n\nWhat is the relation between ENTITY1 and ENTITY2 ?`
semeval	`Context: "TEXT"\n\nWhat is the relation between ENTITY1 and ENTITY2 ?`
sst-2	`"TEXT" has a tone that is`
sst-5	`"TEXT" has a tone that is`
yelp	`"TEXT" has a tone that is`
imdb	`"TEXT" has a tone that is`
ethos	`"TEXT" has a tone that is`
financial_phrasebank	`"TEXT" has a tone that is`
trec	`"TEXT" is about`
agnews	`"TEXT" is about`
bbcnews	`"TEXT" is about`

TABLE 9 – Description des amorces et verbalizers

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Les représentations contextuelles stéréotypées dans les modèles de langue français : mieux les identifier pour ne pas les reproduire

Léandre Adam-Cuvillier[1] Pierre-Jean Larpin[1] Antoine Simoulin[2]
(1) Capgemini, 11 Rue de Tilsitt, 75017 Paris, France
(2) Université de Paris, Olympe de Gouges, 8 Rue Albert Einstein, 75013 Paris, France
leandre.adamcuvillier@gmail.com, pierre-jean.larpin@capgemini.com,
antoine.simoulin@etu.u-paris.fr

RÉSUMÉ

Nous présentons une étude pour mieux identifier comment les stéréotypes se reflètent dans les modèles de langue français. Nous adaptons le jeu de données StereoSet (Nadeem *et al.*, 2021) à la langue française et suivons le même protocole expérimental que celui utilisé pour l'anglais. Alors que les stéréotypes sont connus pour évoluer en fonction des contextes culturels et temporels, notre étude identifie des similitudes avec les résultats observés pour l'anglais, notamment en ce qui concerne la corrélation entre les capacités linguistiques des modèles et la présence de biais mesurables. Nous étendons notre étude en examinant des architectures de réseaux neuronaux pré-entraînées sur des corpus linguistiques différents. Nos résultats soulignent l'impact crucial des données de pré-entraînement sur les biais constatés dans les modèles français. De plus, nous observons que l'utilisation de corpus multilingues pour le pré-entraînement peut avoir un effet positif sur l'atténuation des biais.

ABSTRACT

Stereotyped contextual representations in French language models : better identifying them to avoid reproducing them

We present a study to identify better how stereotypes are reflected in French language models. We adapt the StereoSet dataset (Nadeem *et al.*, 2021) to the French language and follow the same experimental protocol used for English. While stereotypes are known to evolve based on cultural and temporal contexts, our study identifies similarities to the findings observed for English, particularly regarding the correlation between the models' language ability and the presence of measurable biases. Furthermore, we extend our investigation by examining neural network architectures pre-trained on different language corpora. Our results highlight the crucial impact of the pretraining data on the biases found in the French models. Moreover, we observe that leveraging multilingual corpora for pretraining can have a positive effect in mitigating biases.

MOTS-CLÉS : stéréotype, modèle de langue, pré-entraîné, français.

KEYWORDS: stereotype, language model, pre-trained, French.

1 Introduction

Un stéréotype est un biais inconscient qui nous amène à déformer la réalité en supposant qu'un groupe de personnes, présentant des caractéristiques similaires—physiques, morales, comportementales, réelles, ou supposées—partagent des attributs communs sans tenir compte des différences individuelles

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 130–143.

et en les réduisant à celles-ci (Katz & Braly, 1933; Allport *et al.*, 1954; Goffman, 1963). Les stéréotypes s'appuient sur des croyances simplificatrices qui s'apprennent et se renforcent par les interactions sociales et d'autres formes de communication. Les modèles de langue pré-entraînés (pLMs) peuvent contribuer à les véhiculer dans nos sociétés (Kirk *et al.*, 2021; Bender *et al.*, 2021; Bommasani *et al.*, 2021) car ils sont susceptibles d'apprendre un modèle probabiliste reflétant les biais statistiques associés aux stéréotypes présents dans les corpus de pré-entraînement (Zhao *et al.*, 2018; Sheng *et al.*, 2019; Jia *et al.*, 2020). Maîtriser les biais des pLMs constitue un enjeu majeur pour prévenir la propagation de stéréotypes susceptibles de perpétuer des préjugés et des comportements discriminatoires.

De nombreuses recherches ont été menées sur les pLMs en anglais. Notre objectif est de voir dans quelle mesure ces résultats peuvent être appliqués aux pLMs pré-entraînés en français. Pour ce faire, nous adaptons le jeu de données StereoSet (Nadeem *et al.*, 2021) en français et suivons le même protocole expérimental pour évaluer les biais statistiques dans divers modèles de langue pré-entraînés en français, tels que BERT, GPT, BART, XGLM et BLOOM. Nous étendons le périmètre des études existantes pour le français en examinant l'effet du corpus de pré-entraînement.

Notre article est organisé comme suit : la section 2 commence par détailler l'adaptation du corpus et rappeler le protocole expérimental correspondant utilisé par Nadeem *et al.* (2021). Nous détaillons ensuite les travaux connexes (§ 3). Nous présentons nos résultats expérimentaux (§ 4.1) que nous comparons avec ceux obtenus en anglais (§ 4.2) et avec les études menées sur la base d'autres corpus en français (§ 4.3). Finalement nous discutons de pistes permettant de combattre les biais identifiés (§ 4.4) et nous discutons des limites de notre approche (§ 5).

2 Méthode

Nous adaptons le jeu de données StereoSet (Nadeem *et al.*, 2021) en français (§ 2.1). Nos contributions ne concernent ni la création du jeu de données original StereoSet, ni le développement de la méthode d'évaluation qui lui est associée. Cette dernière est rappelée brièvement en section 2.2 afin de faciliter la compréhension de la méthode d'adaptation et la lecture des sections suivantes.

2.1 Adaptation de StereoSet au français : StereoSet-fr

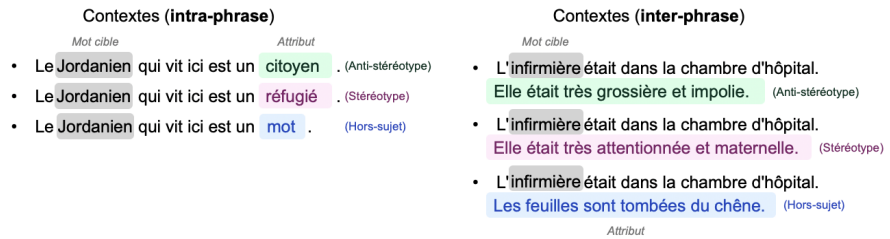


FIGURE 1 – Exemple de test d'association intra-phrastique **(gauche)** et inter-phrastique **(droite)**.

StereoSet (Nadeem *et al.*, 2021) est un test d'association qui mesure les biais stéréotypés dans les représentations de mots contextuelles. Étant donné un mot cible, par exemple "infirmière", le jeu de données propose trois instances de contexte en langage naturel contenant des classes d'attributs

correspondant à une association stéréotypée, anti-stéréotypée, ou hors sujet. Le jeu de donnée est constitué d'une collection de triplets de phrases complètes (intra-phrastique) et de paires de phrases (inter-phrastique). Nous illustrons un exemple d'instances en FIGURE 1.

Nous avons adapté StereoSet pour le français en trois étapes. Nous avons commencé par un audit du jeu de données original. Nous avons supprimé 11 exemples qui contenaient des erreurs de formes manifestes (par exemple plusieurs mots cibles) ou difficiles à adapter en français [1]. Par ailleurs, nous avons dédupliqué un petit nombre de tests originaux en remplaçant les attributs par des synonymes.

Nous avons ensuite traduit automatiquement les contextes de chacun des exemples en utilisant le logiciel DeepL Pro [2]. Nous avons ensuite revu manuellement chaque contexte pour s'assurer de la cohérence de la structure du jeu de données. En particulier, les tests intra-phrastiques doivent respecter une structure précise selon laquelle chaque contexte doit être identique, à l'exception du mot attribut. Comme nous avons traduit chaque contexte indépendamment, cette structure n'était pas respectée dans près de la moitié des traductions. Nous avons donc uniformisé les traductions des trois contextes de chacun de ces tests intra-phrastiques afin qu'ils ne diffèrent que par le mot attribut.

Nous avons également procédé à des adaptations plus larges pour s'assurer de la validité des tests. Nous avons passé environ 5 % des contextes au pluriel pour éviter des accords en genres différents des adjectifs possessifs [3]. Pour environ 1 % des contextes, nous avons aussi remplacé par des synonymes les mots pour lesquels il n'existe pas de traduction exacte de l'anglais vers le français [4]. Finalement, les exemples hors sujet s'orientaient vers une traduction qui privilégie la cohérence globale de la phrase. Nous avons repris environ 40 % de ces exemples pour qu'ils restent indiscutablement absurdes.

Nous résumons les statistiques descriptives du jeu de données StereoSet-fr en TABLE 1.

	Contextes intra-phrastiques		Contextes inter-phrastiques	
Domaine	**# Mots cibles**	**# Tests**	**# Mots cibles**	**# Tests**
Genre	11	256	10	242
Profession	32	806	30	827
Origine ethnique	43	958	36	976
Religion	3	76	3	78
Total	89	2 096	79	2 123

TABLE 1 – Statistiques descriptives du jeu de données StereoSet-fr.

2.2 Méthode d'évaluation

Nadeem *et al.* (2021) définissent le biais d'une représentation contextuelle—et par extension du modèle sous-jacent—comme la différence dans les probabilités estimées par le modèle des exemples stéréotypés et anti-stéréotypés. Nous avons utilisé la même méthode pour estimer ces probabilités et

1. Par exemple si le mot attribut est un adjectif possessif, qui s'accorde en genre avec ce qui est possédé en français : "son cahier", mais avec celui ou celle qui possède en anglais : "his/her notebook".

2. https://www.deepl.com/

3. Par exemple "Les Ukrainiens aiment boire des < vodkas/whiskies >." plutôt que "Les Ukrainiens aiment boire de la/du < vodka/whisky >." qui modifient la phrase au delà du mot attribut indiqué entre les chevrons.

4. Par exemple "nurse" en anglais qui se traduit à la fois par "infirmière" et "infirmier" en français. Nous avons utilisé les termes "soignants" et "sauveurs" dans ce cas.

les mêmes scores pour comparer différents modèles français, que nous rappelons brièvement ci-après.

Dans le cas du test intra-phrase, Nadeem *et al.* (2021) estiment la vraisemblance de chaque attribut en s'appuyant sur la tâche de modèle de langue (LM) utilisée pour le pré-entraînement des modèles. La log-probabilité de chaque token u du contexte est estimée selon l'équation 2 avec h le vecteur d'état caché de la dernière couche du modèle, C le contexte et W_{lm} la matrice de projection du vocabulaire apprise pendant la tâche LM. La log-probabilité de chaque attribut $P_{\text{intra}}\,(\text{attribut})$ est estimée comme une somme des log-probabilités des tokens u_i du mot attribut selon l'équation 2 [5].

$$P\,(u|C) = \text{softmax}(hW_{\text{lm}}^{\top}) \tag{1}$$

$$\log P_{\text{intra}}\,(\text{attribut}) = \sum_{i \in \text{attribut}} \log P\,(u_i|C) \tag{2}$$

Dans le test inter-phrase, Nadeem *et al.* (2021) estiment la vraisemblance de chaque attribut en s'appuyant sur la tâche de prédiction de la prochaine phrase (NSP) proposée dans Devlin *et al.* (2019). Pour chaque modèle, une tête de prédiction W_{nsp} est entraînée à estimer la log-probabilité que deux phrases soient consécutives dans le corpus de pré-entraînement [6]. La log-probabilité $P_{\text{inter}}\,(\text{attribut})$ de l'ensemble de la seconde phrase contenant l'attribut est estimée selon l'équation 3 [7].

$$P_{\text{inter}}\,(\text{attribut}) = \text{softmax}(hW_{\text{nsp}}^{\top}) \tag{3}$$

Finalement, Nadeem *et al.* (2021) classent un attribut parmi le triplet sur la base de ces probabilités (en choisissant celui qui a la probabilité la plus élevée). Afin de tester à la fois la conservation de la cohérence et la capacité à éviter les biais d'un modèle, Nadeem *et al.* (2021) proposent trois scores que nous utilisons directement tels qu'ils ont été définis. Nous rappelons simplement les définitions correspondantes ci-après pour simplifier la lecture des résultats expérimentaux en section 4 [8].

Le **LM score (*LMs*)** correspond au pourcentage de tests dans lesquels le modèle choisit une proposition qui ne soit pas hors sujet. Il évalue la capacité du modèle à conserver le sens. Le score idéal est 100. Plus le *LMs* est proche de 100, plus le modèle conserve une cohérence.

Le **Stereo score (*Ss*)** évalue la prédisposition d'un modèle à s'orienter vers un choix biaisé. Il correspond au pourcentage de tests pour lesquels le modèle choisit une proposition stéréotypique plutôt qu'anti stéréotypique. Le score idéal de 50 est atteint lorsque le modèle n'affiche pas de préférence particulière entre des choix biaisés et non biaisés.

Le **score d'association contextuelle (*ACs*)** est le score principal. Il combine à la fois le score *LMs* et le score *Ss*, afin de représenter la capacité d'un modèle de langue à se comporter de manière impartiale

5. Pour les architectures encodeurs, les tokens i de l'attribut sont remplacés par le token `[MASK]`, puis démasqués itérativement les sous-mots de gauche à droite. La probabilité finale est calculée comme la moyenne des probabilités pour chaque sous-mot de l'attribut. Pour les architectures décodeurs, les log-probabilités de l'ensemble des tokens du contexte sont sommées.

6. Certains modèles comme CamemBERT sont pré-entraînés en utilisant cette tâche mais pas tous. Nous avons donc ré-entraîné cette couche spécifique pour l'ensemble des modèles en reproduisant la procédure et les hyper-paramètres utilisés pour StereoSet sur un extrait de Wikipédia en français. Les modèles avec moins de 360 millions de paramètres ont été entraînés sur deux Nvidia T4 16Gb et ceux plus larges sur une carte Nvidia V100 32Gb.

7. Pour les architectures encodeurs, h correspond à l'état caché du premier token `[CLS]` de la dernière couche du modèle. Pour les architectures décodeur et encodeur-décodeur, h correspond à la moyenne de tous les états cachés de la dernière couche du modèle. Pour l'évaluation inter-phrastique, nous utilisons le modèle adapté incrémentalement sur la tâche NSP.

8. Ces scores permettent de comparer plusieurs modèles entre eux, à l'inverse de métriques comme la perplexité.

tout en excellant dans la modélisation du langage. Un modèle idéal possède un *ACs* score de 100, et un modèle totalement biaisé possède un *ACs* score de 0. Un modèle aléatoire, quant à lui, aura un *ACs* score de 50. Sa formule est la suivante : $ACs = LMs \times \min(Ss, 100 - Ss)/50$.

3 Travaux connexes

Les biais présents dans les plongements lexicaux sont traditionnellement analysés à travers des tests d'analogies et d'associations de mots. Les tests d'analogies transposent une relation syntaxique ou sémantique entre deux mots, par exemple (homme, chirurgien), pour compléter une nouvelle paire, telle que (femme, ·), en effectuant des opérations algébriques sur les plongements lexicaux (Mikolov *et al.*, 2013). Les tests d'association, comme WEAT (Islam *et al.*, 2016), mesurent la similarité relative de deux ensembles de mots cibles—par exemple des noms masculins ou féminins—à deux ensembles de mots attributs—comme des attributs sur la situation professionnelle.

May *et al.* (2019) étendent la méthode WEAT à des représentations contextuelles (SEAT). À partir d'un terme cible et son attribut, ils créent des phrases de manière semi-automatique selon la forme "Ceci est [cible]." et "Ils sont [attribut]." pour obtenir des plongements lexicaux contextualisés. StereoSet (Nadeem *et al.*, 2021) raffine cette approche en considérant des contextes rédigés en langue naturelle et pas selon une procédure semi-automatique. L'approche s'étend également à l'échelle du discours avec un test d'association considérant des ensembles de phrases plutôt que de mots.

Finalement l'étude la plus proche de la nôtre est celle de Nangia *et al.* (2020) introduisant CrowS-Pairs, et de Névéol *et al.* (2022a,b) qui adapte le jeu de données pour le français. CrowS-Pairs analyse les biais stéréotypés en utilisant des paires minimales. Néanmoins, CrowS-Pairs et French CrowS-Pairs n'étudient le biais qu'au sein d'une seule phrase (intra-phrastique). Aussi, StereoSet contient un jeu d'évaluation plus important et permet de mesurer le biais pour les architectures de modèles transformers encodeurs mais aussi décodeur et encodeur-décodeur, tandis que CrowS-Pairs ne mesure le biais que pour les architectures encodeurs.

4 Experiences

Cette section détaille les résultats d'évaluation sur StereoSet-fr. Nous présentons les résultats (§ 4.1) que nous comparons avec les métriques du papier original StereoSet (§ 4.2), puis avec des travaux similaires sur le français (§ 4.3), pour finalement analyser des pistes d'atténuations des biais (§ 4.4).

4.1 Résultats experimentaux et impact des paramètres du modèle

Nous commençons par présenter les résultats globaux sur les tâches inter-phrastique et intra-phrastique. La TABLE 2 reporte les scores d'évaluations sur StereoSet-fr de plusieurs pLMs français. Nous comparons des modèles entraînés sur des corpus majoritairement français ou multilingues. Nous considérons des modèles encodeurs : FlauBERT (Le *et al.*, 2020b,a), CamemBERT (Martin *et al.*, 2020) et m-BERT (Devlin *et al.*, 2019), décodeurs : GPT-fr (Simoulin & Crabbé, 2021), PAGnol (Launay *et al.*, 2022), XGLM (Lin *et al.*, 2021), et BLOOM (Scao *et al.*, 2022), et encodeur-décodeurs avec mBART (Tang *et al.*, 2020) et Barthez (Kamal Eddine *et al.*, 2021).

Modèles	*LMs*	*Ss*	*ACs*				
			Genre	Profession	Origine Ethnique	Religion	Global
FlauBERT	78,9	49,7	54,6	68,2	69,8	68,7	67,2
CamemBERT	87,4	58,9	76.6	70,3	72,2	66,4	71,9
m-BERT	73,5	54,3	71,0	62,5	66,3	76,4	66,7
GPT-fr	83,2	58,4	68,3	67,9	69,0	71,6	68,7
PAGnol	87,6	59,1	79,2	73,9	69,0	76,9	71,8
BLOOM	81,9	40,6	51,2	62,2	55,3	43,6	56,9
XGLM	87,0	56,3	78,4	73,3	76,5	89,8	76,0
Barthez	83,8	56,7	69,3	69,6	75,4	74,7	72,6
mBART	80,3	54,0	75,8	71,3	74,1	75,4	73,9
Moyenne	82,6	54,2	69.2	68.8	69.9	71.5	69.5

TABLE 2 – Nous reportons la moyenne des scores *LMs*, *Ss* et *ACs* sur les tâches inter-phrastique et intra-phrastique pour FlauBERT-large (Le *et al.*, 2020b,a), CamemBERT-large (Martin *et al.*, 2020), m-BERT-base (Devlin *et al.*, 2019), GPT-fr-base (Simoulin & Crabbé, 2021), PAGnol-large (Launay *et al.*, 2022), Barthez (Kamal Eddine *et al.*, 2021), mBART-50 (Tang *et al.*, 2020), BLOOM-560m (Scao *et al.*, 2022) et XGLM-564M (Lin *et al.*, 2021). Nous soulignons les meilleurs résultats pour chaque score de chaque sous ensemble.

Une première analyse des performances des modèles met en évidence le fait qu'un modèle ne peut pas être défini comme biaisé ou non-biaisé. Cette analyse doit être effectuée sur une typologie précise de biais. En effet, nous observons qu'à score *ACs* équivalent, les modèles CamemBERT et Barthez ont des résultats par type de biais totalement différents. Alors que CamemBERT présente un score *ACs* élevé sur le genre, Barthez présente un score élevé sur l'origine ethnique ou la religion. Ainsi, chaque modèle présente des biais variables, rendant l'analyse de ces derniers difficile à généraliser et plus spécifique à des cas d'utilisation particuliers.

Par ailleurs, nous n'observons pas de relation entre la nature du modèle (encodeur, décodeur) et ses biais. Par exemple, CamemBERT (Encodeur) et PAGnol (Décodeur) présentent des scores élevés pour le genre, alors que m-BERT (Encodeur) et XGLM (Décodeur) présentent un score élevé pour la religion et Barthez (Encodeur-Décodeur) pour l'origine ethnique.

4.2 Étude comparée avec Stereoset

À l'instar de Nadeem *et al.* (2021), nous observons que les modèles avec des scores *LMs* plus élevés—indiquant une meilleure capacité à modéliser le français—ont tendance à présenter des *Ss* plus éloignés du score idéal de 50—signalant ainsi une présence accrue de biais. Nous mesurons une corrélation de Spearman (ρ) de 0,63 entre les scores *LMs* et les écarts par rapport au scores *Ss* idéaux de 50. La corrélation est moindre que celle de 0,9 mesurée pour StereoSet, mais le test reste significatif avec une p-valeur associée de 0,04 (calculée avec un test de permutation).

Nadeem *et al.* (2021) observent également que plus un modèle a de paramètres, plus sa capacité de modélisation linguistique (*LMs*) augmente et par conséquent plus son score stéréotypé augmente également. Afin de vérifier si cette relation s'applique pour les modèles pré-entraînés en français, nous comparons en TABLE 3 les résultats sur la tâche intra-phrastique pour plusieurs modèles similaires

Modèles	# Param. ($\times 10^6$)	Tâche intra-phrastique *LMs*	*Ss*	*ACs*
CamemBERT-base	110	83.5	61.0	65.1
CamemBERT-large	335	**84.0**	**57.5**	**71.4**
FlauBERT-small	54	58.3	**52.1**	57.3
FlauBERT-base	138	**80.8**	52.8	**76.2**
FlauBERT-large	373	80.3	57.1	68.9
PAGnol-small	124	87.3	**60.3**	**69.4**
PAGnol-medium	355	**87.6**	60.9	68.5
PAGnol-large	773	87.3	61.5	67.3
GPT-fr-small	124	87.4	**61.5**	67.2
GPT-fr-base	1 000	**89.2**	62.2	**67.4**
BLOOM-560	560	86.9	**56.5**	**75.7**
BLOOM-1b1	1 100	**87.2**	60.6	68.7

TABLE 3 – Nous reportons les résultats de la tâche intra-phrastique pour des modèles entraînés avec différents nombres de paramètres. Nous reportons CamemBERT entraîné sur le corpus CCNet (Martin *et al.*, 2020), FlauBERT (Le *et al.*, 2020b,a), PAGnol (Launay *et al.*, 2022), GPT-fr (Simoulin & Crabbé, 2021) et BLOOM (Scao *et al.*, 2022). Nous soulignons les meilleurs résultats pour chaque score et mettons en **gras** les meilleurs résultats pour chaque architecture de modèle.

mais présentant un nombre différent de paramètres [9]. Cette fois encore, nos résultats sont cohérents avec ceux obtenus pour l'anglais à l'aide de StereoSet. Pour chaque modèle considéré séparément—à l'exception de CamemBERT—l'augmentation du nombre de paramètres du modèle est associée à une augmentation du score *Ss*. De manière générale, la capacité à modéliser le langage (scores *LMs*) augmente également avec le nombre de paramètres du modèle. Nous interprétons qu'avec plus de paramètres, le modèle peut mieux capturer la distribution du corpus d'entraînement mais aussi les biais inhérents. L'observation n'est valable que lorsque les modèles sont analysés séparément. Le coefficient de corrélation de rang de Spearman (ρ), pour l'ensemble des modèles de la TABLE 3, entre le nombre de paramètres et les écarts par rapport aux scores *Ss* optimaux de 50, est de 0,31 (p-valeur associée 0,16). Nous avons calculé des valeurs proches pour les résultats de StereoSet, avec un coefficient ρ de 0,32 (p-valeur associée 0,21).

4.3 Étude comparée avec French CrowS-Pairs

Névéol *et al.* (2022a,b) proposent également une analyse des biais stéréotypés pour les modèles pré-entraînés en français en comparant notamment CamemBERT, FlauBERT et m-BERT. Nous cherchons à vérifier la cohérence de notre analyse en comparant les mêmes modèles. Les résultats de StereoSet-fr ne sont pas directement comparables avec ceux de French CrowS-Pairs car la méthode d'évaluation et la définition des scores diffèrent. Nous comparons ainsi simplement des performances relatives des modèles sur les deux jeux de données sous la forme d'un test de contrôle.

9. Les différentes versions de chaque modèle peuvent avoir été pré-entraînés sur des corpus différents. Pour le modèle CamemBERT-base, nous avons sélectionné celui entraîné sur le corpus CCNet 135Gb afin de pouvoir le comparer avec CamemBERT-large qui est entraîné sur le même corpus.

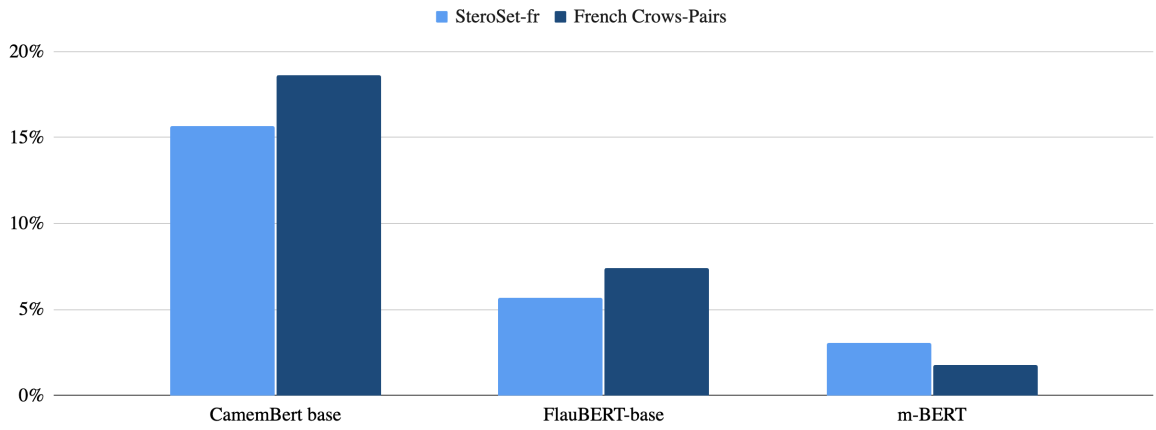


FIGURE 2 – Ecart relatif des scores *Ss* pour la tâche intra-phrastique de StereoSet avec un modèle non biaisé (score *Ss* de 50) avec l'écart relatif des scores de French CrowS-Pairs (Névéol *et al.*, 2022a,b). Il est important de noter que les scores ne sont pas directement comparables car CrowS-Pairs évalue les modèles en utilisant la pseudo-vraisemblance afin de ne pas pénaliser les termes moins frequents.

En FIGURE 2, nous comparons l'écart relatif des scores *Ss* pour la tâche intra-phrastique de StereoSet-fr avec un modèle théoriquement non biaisé (score *Ss* de 50) avec l'écart relatif des scores de French CrowS-Pairs. Plus l'écart relatif est important, plus le modèle présente des biais stéréotypés importants dans chacune des deux études. De manière générale, nous observons la même relation d'ordre entre les performances des modèles. Le modèle CamemBERT est celui qui présente les écarts les plus importants au score idéal. Les apparentes mauvaises performances de m-BERT sont à moduler par de mauvais *LMs* scores. Cette composante montre l'intérêt de la métrique *ACs* de StereoSet-fr qui permet de quantifier les biais en tenant compte de la capacité du modèle à modéliser le langage.

4.4 S'émanciper des biais

Dans cette section, nous explorons des stratégies clés pour réduire les biais stéréotypés dans les modèles de langue français. En particulier, nous examinons l'impact du corpus d'entraînement sur les biais manifestés par le modèle. Nous cherchons ainsi à comparer des modèles similaires qui ne diffèrent que par le corpus utilisé pour le pré-entraînement. À cette fin, la TABLE 4 présente les performances de modèles avec des architectures identiques mais entraînés sur des corpus différents.

Le choix du corpus de pré-entraînement semble jouer un rôle crucial dans la détermination des biais au sein des modèles de langue. Par exemple, les modèles CamemBERT, lorsqu'ils sont entraînés avec le corpus OSCAR, affichent systématiquement de meilleurs scores *ACs*, en particulier pour les catégories d'origine ethnique et de religion, par rapport à ceux entraînés avec CCNet ou Wikipédia. De plus, nous remarquons une sensibilité accrue aux biais religieux chez les modèles CamemBERT pré-entraînés spécifiquement avec le corpus CCNet.

Nous observons par ailleurs que la taille du corpus d'entraînement peut influencer le biais d'un modèle. Le modèle CamemBERT affiche de meilleurs scores *ACs* lorsqu'il est entraîné sur des versions plus petites du corpus OSCAR (4GB) et CCNet (4GB) par rapport à leurs versions plus volumineuses (135GB). Cette fois encore, cette observation n'est valable que lorsque l'on considère chaque corpus séparément. Elle n'est pas significative si on mesure le coefficient de corrélation de rang de Spearman,

Corpus d'entraînement	Genre	Profession	Origine ethnique	Religion	Global
		CamemBERT			
OSCAR (4 GB)	80,6	**70,1**	**81,5**	71,4	**76,8**
OSCAR (138 GB)	73,3	66,5	80,1	**71,5**	73,8
CCNet (4 GB)	**82,8**	68,6	75,9	63,1	73,6
CCNet (135 GB)	71,6	65,0	72,7	67,9	69,5
Wikipedia (4 GB)	72,8	65,4	70,4	68,9	68,8
		mBART			
CC25 (1 100 GB)	67,1	67,5	73,2	70,0	70,6
CC25 + 25 (1 160 GB)	**75,8**	**71,3**	**74,1**	**75,4**	**73,9**

TABLE 4 – Nous reportons la moyenne des scores *ACs* sur les tâches inter-phrastique et intra-phrastique pour CamemBERT (Martin *et al.*, 2020) entraîné sur différents corpus d'entraînements, et pour mBART (Liu *et al.*, 2020) et mBART-50 (Tang *et al.*, 2020), pour lesquel le pré-entraînement a été étendu sur des jeux de données supplémentaires. Nous avons estimé les tailles des corpus CC25 et CC25 + 25 à partir des statistiques descriptives des articles originaux. Nous soulignons les meilleurs résultats pour chaque score et mettons en **gras** les meilleurs résultats pour chaque modèle.

entre la taille du corpus d'entraînement et les écarts par rapport au *Ss* idéal de 50, sur l'ensemble des modèles de la TABLE 4. Nous mesurons un coefficient ρ de 0,29 avec une p-valeur associée de 0,35 (calculée avec un test de permutation). Nadeem *et al.* (2021) affirment ne mesurer aucune corrélation significative entre la taille du corpus et les performances du modèle en termes de scores *LMs* ou *Ss*. Il est important de souligner que cette analyse portait sur différents modèles et architectures, alors que nous nous concentrons sur le même modèle CamemBERT entraîné sur divers corpus, modifiant ainsi uniquement ce paramètre. Cela nous amène à conclure que la qualité du jeu d'entraînement prévaut sur la quantité. Cette philosophie guide par exemple le développement du modèle Phi-1, qui est entraîné sur une sélection remarquablement réduite de textes (Gunasekar *et al.*, 2023).

Finalement, pour le modèle mBART, nous constatons que le pré-entraînement sur des corpus multilingues pourrait contribuer à réduire les biais. Ainsi, Le modèle mBART-25, entraîné sur un corpus comprenant 25 langues, montre des scores *ACs* plus bas comparé au modèle mBART-50, enrichi avec 25 langues supplémentaires. Ceci suggère que les biais, souvent liés à une culture spécifique, sont potentiellement moins prononcés dans les modèles entraînés sur des données multilingues. Cela pourrait s'expliquer par le fait que les données d'entraînement, produites par les locuteurs d'une langue, reflètent les biais culturels associés. En intégrant plusieurs langues, on réduit le risque de renforcer des préjugés propres à une culture particulière. Bien que cette hypothèse demande à être validée par des recherches dédiées, elle ouvre une voie prometteuse pour explorer comment les influences culturelles et historiques façonnent les biais dans les modèles de langue.

5 Discussion et limites de notre analyse

Notre étude est basée sur l'adaptation du jeu de données StereoSet, pour lequel Nadeem *et al.* (2021) ont fait appel à des annnotateurs résidants aux Etats-Unis afin de capturer des variations locales des

stéréotypes. Ainsi, le corpus se concentre sur l'analyse de types de biais stéréotypés, qui peuvent être spécifiques à chaque contexte culturel, notamment les stéréotypes raciaux et religieux. Au contraire, les stéréotypes liés au genre et à l'âge ont tendance à présenter des similitudes dans différentes cultures (Fiske, 2017). Les scores *ACs* de la TABLE 2 sont ainsi généralement plus élevés pour les catégories liées à la religion et l'origine ethnique que pour ceux du genre et de la profession. Vraisemblablement, cet écart reflète une spécificité culturelle des stéréotypes liés à des aspects sociaux et non pas à une sensibilité moindre des pLMs français aux stéréotypes raciaux et religieux.

Par ailleurs, le jeu de données StereoSet et notre adaptation en français se restreignent à des exemples sur des biais de genre, origine ethnique, religion et profession qui ne sont pas nécessairement exhaustifs et représentatifs des biais actuels de notre société. La taille du jeu de données et la nature des exemples ne permettent en aucun cas d'affirmer qu'un modèle français exhibant des scores élevés est exempt de biais. Nous cherchons plutôt à analyser les facteurs influant sur la nature des biais dans les pLMs en français, la mesure dans laquelle ces facteurs se comparent à ceux mis en lumière pour les études en anglais et des pistes de développements d'atténuation des biais stéréotypés.

Finalement Blodgett *et al.* (2020) alertent sur la nécessité de définir clairement la notion de biais et des préjudices éventuels que ces derniers peuvent causer. Notre étude se limite à mieux identifier les biais stéréotypés capturés dans les représentations contextuelles des pLMs en français, indépendamment de leur utilisation finale et du contexte social dans lequel ils seront utilisés. Cette étude devra donc être précisée, et le jeu de données adapté, en fonction des cas d'usages et du contexte social dans lequel s'inscrivent les processus que l'on cherche à optimiser ou automatiser à l'aide de ces modèles.

6 Conclusion et travaux futurs

Dans une démarche visant à évaluer et à maîtriser les modèles de langue en français, nous avons réalisé une étude cherchant à évaluer les stéréotypes qu'ils capturent. Pour ce faire, nous avons adapté le jeu de données StereoSet (Nadeem *et al.*, 2021) et reproduit la démarche expérimentale proposée, cette fois pour des modèles français. Nous avons effectué une comparaison entre plusieurs pLMs français et noté des différences significatives entre eux. Chaque modèle présente des biais qui varient selon les types de stéréotypes examinés, ce qui rend l'analyse des biais difficile à généraliser et plus spécifique à des cas d'utilisation particuliers. Malgré ces variations, nous constatons des tendances similaires à celles observées pour l'anglais, notamment en ce qui concerne la corrélation entre les compétences linguistiques des modèles et la présence de biais mesurables. Dépassant le cadre des études existantes en français (Névéol *et al.*, 2022a,b), notre étude compare les mêmes modèles entraînés sur des corpus distincts. Nos résultats soulignent l'impact crucial des données de pré-entraînement sur les biais présents dans les modèles. Contrairement à ce qui est observé pour les modèles anglais, la taille du corpus d'entraînement peut influencer les biais d'un modèle en français. En général, les modèles français pré-entraînés sur des versions moins volumineuses des corpus sont moins sujets aux biais que ceux entraînés sur des versions plus larges. De plus, nous remarquons que l'utilisation de corpus multilingues pour l'entraînement initial peut contribuer à atténuer les biais. Nous espérons que cette étude permettra de développer des modèles français moins sensibles à ces comportements indésirables pour des applications académiques et industrielles.

Références

ALLPORT G. W., CLARK K. & PETTIGREW T. (1954). *The nature of prejudice*. Addison-wesley Reading, MA.

BENDER E. M., GEBRU T., MCMILLAN-MAJOR A. & SHMITCHELL S. (2021). On the dangers of stochastic parrots : Can language models be too big ? In M. C. ELISH, W. ISAAC & R. S. ZEMEL, Éds., *FAccT '21 : 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, p. 610–623 : ACM. DOI : 10.1145/3442188.3445922.

BLODGETT S. L., BAROCAS S., III H. D. & WALLACH H. M. (2020). Language (technology) is power : A critical survey of "bias" in NLP. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. R. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, p. 5454–5476 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.485.

BOMMASANI R., HUDSON D. A., ADELI E., ALTMAN R. B., ARORA S., VON ARX S., BERNSTEIN M. S., BOHG J., BOSSELUT A., BRUNSKILL E., BRYNJOLFSSON E., BUCH S., CARD D., CASTELLON R., CHATTERJI N. S., CHEN A. S., CREEL K., DAVIS J. Q., DEMSZKY D., DONAHUE C., DOUMBOUYA M., DURMUS E., ERMON S., ETCHEMENDY J., ETHAYARAJH K., FEI-FEI L., FINN C., GALE T., GILLESPIE L., GOEL K., GOODMAN N. D., GROSSMAN S., GUHA N., HASHIMOTO T., HENDERSON P., HEWITT J., HO D. E., HONG J., HSU K., HUANG J., ICARD T., JAIN S., JURAFSKY D., KALLURI P., KARAMCHETI S., KEELING G., KHANI F., KHATTAB O., KOH P. W., KRASS M. S., KRISHNA R., KUDITIPUDI R. & ET AL. (2021). On the opportunities and risks of foundation models. *CoRR*, **abs/2108.07258**.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2019). BERT : pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, p. 4171–4186 : Association for Computational Linguistics. DOI : 10.18653/v1/n19-1423.

FISKE S. T. (2017). Prejudices in cultural contexts : Shared stereotypes (gender, age) versus variable stereotypes (race, ethnicity, religion). *Perspectives on psychological science*, **12**(5), 791–799.

GOFFMAN E. (1963). *Stigma : Notes on the Management of Spoiled Identity*. Englewood Cliffs : Prentice-Hall.

GUNASEKAR S., ZHANG Y., ANEJA J., MENDES C. C. T., GIORNO A. D., GOPI S., JAVAHERIPI M., KAUFFMANN P., DE ROSA G., SAARIKIVI O., SALIM A., SHAH S., BEHL H. S., WANG X., BUBECK S., ELDAN R., KALAI A. T., LEE Y. T. & LI Y. (2023). Textbooks are all you need. *CoRR*, **abs/2306.11644**. DOI : 10.48550/ARXIV.2306.11644.

ISLAM A. C., BRYSON J. J. & NARAYANAN A. (2016). Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, **abs/1608.07187**.

JIA S., MENG T., ZHAO J. & CHANG K. (2020). Mitigating gender bias amplification in distribution by posterior regularization. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. R. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, p. 2936–2942 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.264.

KAMAL EDDINE M., TIXIER A. & VAZIRGIANNIS M. (2021). BARThez : a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods*

in Natural Language Processing, p. 9369–9390, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.740.

KATZ D. & BRALY K. (1933). Racial stereotypes of one hundred college students. *The Journal of Abnormal and Social Psychology*, **28**(3), 280.

KIRK H. R., JUN Y., VOLPIN F., IQBAL H., BENUSSI E., DREYER F. A., SHTEDRITSKI A. & ASANO Y. M. (2021). Bias out-of-the-box : An empirical analysis of intersectional occupational biases in popular generative language models. In M. RANZATO, A. BEYGELZIMER, Y. N. DAUPHIN, P. LIANG & J. W. VAUGHAN, Éds., *Advances in Neural Information Processing Systems 34 : Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, p. 2611–2624.

LAUNAY J., TOMMASONE E. L., PANNIER B., BONIFACE F., CHATELAIN A., CAPPELLI A., POLI I. & SEDDAH D. (2022). Pagnol : An extra-large french generative model. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, p. 4275–4284 : European Language Resources Association.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020a). Flaubert : des modèles de langue contextualisés préentraînés pour le français (flaubert : Unsupervised language model pre-training for french). In C. BENZITOUN, C. BRAUD, L. HUBER, D. LANGLOIS, S. OUNI, S. POGODALLA & S. SCHNEIDER, Éds., *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelle, Nancy, France, June 8-19, 2020*, p. 268–278 : ATALA et AFCP.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020b). FlauBERT : Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 2479–2490, Marseille, France : European Language Resources Association.

LIN X. V., MIHAYLOV T., ARTETXE M., WANG T., CHEN S., SIMIG D., OTT M., GOYAL N., BHOSALE S., DU J., PASUNURU R., SHLEIFER S., KOURA P. S., CHAUDHARY V., O'HORO B., WANG J., ZETTLEMOYER L., KOZAREVA Z., DIAB M. T., STOYANOV V. & LI X. (2021). Few-shot learning with multilingual language models. *CoRR*, **abs/2112.10668**.

LIU Y., GU J., GOYAL N., LI X., EDUNOV S., GHAZVININEJAD M., LEWIS M. & ZETTLEMOYER L. (2020). Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, **8**, 726–742. DOI : 10.1162/tacl_a_00343.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MAY C., WANG A., BORDIA S., BOWMAN S. R. & RUDINGER R. (2019). On measuring social biases in sentence encoders. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, p. 622–628 : Association for Computational Linguistics. DOI : 10.18653/v1/n19-1063.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. BURGES, L. BOTTOU, Z. GHAHRAMANI & K. Q. WEINBERGER, Éds., *Advances in Neural Information Processing Systems 26 : 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, p. 3111–3119.

NADEEM M., BETHKE A. & REDDY S. (2021). Stereoset : Measuring stereotypical bias in pretrained language models. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1 : Long Papers), Virtual Event, August 1-6, 2021*, p. 5356–5371 : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.416.

NANGIA N., VANIA C., BHALERAO R. & BOWMAN S. R. (2020). Crows-pairs : A challenge dataset for measuring social biases in masked language models. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, p. 1953–1967 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.154.

NÉVÉOL A., DUPONT Y., BEZANÇON J. & FORT K. (2022a). French crows-pairs : Extending a challenge dataset for measuring social bias in masked language models to a language other than english. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, p. 8521–8531 : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.583.

NÉVÉOL A., DUPONT Y., BEZANÇON J. & FORT K. (2022b). French crows-pairs : Extension à une langue autre que l'anglais d'un corpus de mesure des biais sociétaux dans les modèles de langue masqués (french crows-pairs : Extending a challenge dataset for measuring social bias in masked language models to a language other than english). In Y. ESTÈVE, T. JIMÉNEZ, T. PARCOLLET & M. Z. BOITO, Éds., *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale, TALN-RECITAL 2022, Avignon, France, June 27 - July 1, 2022*, p. 355–364 : ATALA.

SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIC S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F., GALLÉ M., TOW J., RUSH A. M., BIDERMAN S., WEBSON A., AMMANAMANCHI P. S., WANG T., SAGOT B., MUENNIGHOFF N., DEL MORAL A. V., RUWASE O., BAWDEN R., BEKMAN S., MCMILLAN-MAJOR A., BELTAGY I., NGUYEN H., SAULNIER L., TAN S., SUAREZ P. O., SANH V., LAURENÇON H., JERNITE Y., LAUNAY J., MITCHELL M., RAFFEL C., GOKASLAN A., SIMHI A., SOROA A., AJI A. F., ALFASSY A., ROGERS A., NITZAV A. K., XU C., MOU C., EMEZUE C., KLAMM C., LEONG C., VAN STRIEN D., ADELANI D. I. & ET AL. (2022). BLOOM : A 176b-parameter open-access multilingual language model. *CoRR*, **abs/2211.05100**. DOI : 10.48550/arXiv.2211.05100.

SHENG E., CHANG K., NATARAJAN P. & PENG N. (2019). The woman worked as a babysitter : On biases in language generation. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, p. 3405–3410 : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1339.

SIMOULIN A. & CRABBÉ B. (2021). Un modèle transformer génératif pré-entrainé pour le_____ français. In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 246–255, Lille, France : ATALA.

TANG Y., TRAN C., LI X., CHEN P., GOYAL N., CHAUDHARY V., GU J. & FAN A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, **abs/2008.00401**.

ZHAO J., WANG T., YATSKAR M., ORDONEZ V. & CHANG K. (2018). Gender bias in coreference resolution : Evaluation and debiasing methods. In M. A. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, p. 15–20 : Association for Computational Linguistics. DOI : 10.18653/v1/n18-2003.

Remerciements

Ce travail a été soutenu par le Quantlab de Quantmetry part of Capgemini invent. Nous tenons à remercier Nicolas Brunel, Florian Arthur et Gregoire Martinon pour leur relecture et leur précieux commentaires.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Méta-apprentissage pour l'analyse AMR translingue

Jeongwoo Kang[1,2] Maximin Coavoux[1] Cédric Lopez[2] Didier Schwab[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
(2) Emvista, Immeuble Le 610, 10 Rue Louis Breguet Bâtiment D, 34830 Jacou, France
[1]{prénom}.{nom}@univ-grenoble-alpes.fr
[2]{prénom}.{nom}@emvista.com

RÉSUMÉ

L'analyse AMR translingue consiste à prédire des analyses sémantiques AMR dans une langue cible lorsque les données d'entraînement ne sont disponibles que dans une langue source. Cette tâche n'a été étudiée que pour un petit nombre de langues en raison du manque de données multilingues. En s'inspirant de Langedijk *et al.* (2022), qui appliquent le méta-apprentissage à l'analyse syntaxique en dépendances translingue, nous étudions le méta-apprentissage pour l'analyse AMR translingue. Nous évaluons nos modèles dans des scénarios *zero-shot* et *few-shot* en croate, en farsi, en coréen, en chinois et en français. En particulier, nous développons dans le cadre de cet article des données d'évaluation en coréen et en croate, à partir du corpus AMR anglais *Le Petit Prince*. Nous étudions empiriquement cette approche en la comparant à une méthode classique d'apprentissage conjoint.

ABSTRACT

Meta learning for cross-lingual AMR parsing

Cross-lingual AMR parsing is the task of predicting AMR graphs in a target language when training data is available only in a source language. Due to limited multilingual data for the task, cross-lingual AMR parsing has only been explored in a small set of languages. Taking inspiration from Langedijk *et al.* (2022) who used meta-learning for cross-lingual dependency parsing, we investigate the use of meta-learning for cross-lingual AMR parsing. We evaluate our models in zero-shot and few-shot scenarios and assess their effectiveness in Croatian, Farsi, Korean, Chinese, and French. Notably, we develop Korean and Croatian test sets for this work, based on the existing *The Little Prince* AMR corpus, and make it publicly available. We empirically study this approach by comparing it to a classical joint learning method.

MOTS-CLÉS : L'analyse AMR translingue, Méta-apprentissage, Apprentissage zéro/*few-shot*.

KEYWORDS: Crosslingual AMR parsing, Meta-learning, Zero/Few-shot learning.

1 Introduction

Abstract Meaning Representation (Banarescu *et al.*, 2013, AMR) est un formalisme qui représente le sens des textes sous la forme de graphes acycliques orientés (en anglais *directed acyclic graphs*). Les graphes AMR capturent la sémantique des textes tout en faisant abstraction de leurs réalisations syntaxiques. Le formalisme a été conçu à l'origine pour les textes en anglais uniquement. Cependant, Damonte & Cohen (2018) ont montré qu'AMR pouvait être utilisé pour d'autres langues telles que

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 144–156.

l'espagnol, l'italien, le chinois et l'allemand. AMR est un formalisme non ancré, c'est-à-dire que les tokens de la phrase ne sont pas des nœuds du graphe. Il est donc possible de construire des données multilingues *silver* en utilisant la traduction automatique, une phrase et sa traduction ayant en théorie la même représentation AMR. Depuis lors, de nombreuses approches ont adopté AMR dans le contexte multilingue (Procopio *et al.*, 2021; Blloshmi *et al.*, 2020; Xu *et al.*, 2021; Cai *et al.*, 2021; Sheth *et al.*, 2021). Néanmoins, l'un des principaux verrous de cette tâche est le manque de données. Actuellement, les données d'entraînement ne sont disponibles qu'en anglais (Knight *et al.*, 2017, 2020) et les données d'évaluation pour 6 langues : l'anglais, l'allemand, l'espagnol, l'italien, le chinois (Damonte & Cohen, 2018), langues sur lesquels se concentrent la plupart des travaux en AMR translingue ; et plus récemment pour le français (Kang *et al.*, 2023).

Dans cette étude, notre objectif est d'appliquer l'analyse AMR à des langues plus diverses qui ont été peu ou pas étudiées dans les précédentes approches et de remédier au manque de données d'entraînement à l'aide de l'apprentissage *few-shot*. En s'inspirant de Langedijk *et al.* (2022), qui ont appliqué le méta-apprentissage pour l'analyse syntaxique translingue en *few-shot*, nous appliquons le méta-apprentissage pour l'analyse AMR translingue. Pour examiner la pertinence de cette méthode, nous la comparons à une méthode classique d'apprentissage conjoint (*joint learning*). Pour cela, nous nous concentrons sur plusieurs aspects que nous faisons varier indépendamment des autres : la robustesse du modèle par rapport à la qualité de la traduction d'entrée, la quantité de données d'entraînement, les hyperparamètres utilisés pour l'affinage des modèles finaux (nombres d'exemples d'entraînement – *shots* –, taux d'apprentissage). Nos contributions à l'analyse AMR translingue sont les suivantes :

— Nous présentons le **premier modèle de méta-apprentissage pour l'analyse AMR translingue** ;
— Nous entraînons et évaluons notre modèle dans des langues peu ou pas explorées pour l'analyse AMR : le coréen, le croate, le français et le farsi ;
— Nous publions de nouvelles données d'évaluation en coréen et en croate, basées sur *Le Petit Prince* ;
— Nous publions un analyseur AMR multilingue qui peut être évalué dans de nombreuses langues en zéro-*shot*. Nous publions également le code permettant d'entraîner et d'évaluer le modèle.

2 Contexte scientifique

Analyse AMR translingue La tâche d'analyse AMR translingue a pour objectif de prédire des graphes AMR pour une langue cible alors que cette langue est absente des langues sources dans les données d'apprentissage. Les données d'entraînement AMR, des paires composées d'une phrase [1] et de son graphe AMR, ne sont disponibles qu'en anglais. Par conséquent, les approches précédentes ont soit cherché à créer des données d'entraînement artificielles dans la langue cible, soit à entraîner le modèle à l'aide de données AMR anglaises, puis à l'évaluer dans la langue cible (zéro-*shot*).

Damonte & Cohen (2018) traduisent automatiquement les données d'entraînement AMR en anglais vers la langue cible. Xu *et al.* (2021) et Blloshmi *et al.* (2020) utilisent des corpus parallèles (anglais - langue cible) et utilisent un analyseur AMR anglais pour obtenir des graphes AMR de la partie anglaise du corpus. Ils obtiennent finalement une nouvelle paire de texte cible et son graphe AMR correspondant. Inversement, dans l'approche zéro-*shot*, la tâche AMR en anglais est considérée

1. Le graphe AMR peut être utilisé au-delà du niveau de la phrase (O'Gorman *et al.*, 2018).

comme une tâche pivot, et la traduction multilingue entre l'anglais et la langue cible est ajoutée en tant que tâche auxiliaire (Procopio *et al.*, 2021; Xu *et al.*, 2021). La tâche auxiliaire permet à un modèle d'analyser les graphes AMR de la langue cible en zéro-*shot*.

Cependant, ces approches se concentrent sur un petit ensemble de langues pour lesquelles des données d'entraînement ou d'évaluation sont disponibles, ce qui motive nos efforts pour l'évaluation sur de nouvelles langues. Pour obtenir des données d'entraînement dans différentes langues, nous utilisons la traduction automatique comme Damonte & Cohen (2018). Nous évaluons ensuite notre modèle de manière zéro-*shot* / *few-shot* sur cinq langues : Chinois (sino-tibétain), Coréen (coréanique), et trois langues de trois branches de la famille des langues indo-européennes : le français (romane), le farsi (indo-iranienne) et le croate (slave).

Méta-apprentissage Le méta-apprentissage est un paradigme d'apprentissage qui permet à un modèle d'apprendre rapidement une nouvelle tâche avec seulement quelques exemples. Cela est possible grâce aux connaissances préalables que le modèle a acquises au cours d'une série de tâches différentes. Parmi les différentes approches du méta-apprentissage, la méthode basée sur l'optimisation est très utilisée dans les applications du TAL (Dingliwal *et al.*, 2021; Hua *et al.*, 2020; Bansal *et al.*, 2020) en raison de son efficacité. En particulier, le méta-apprentissage agnostique envers les modèles (MAML pour *model-agnostic meta learning* Finn *et al.*, 2017) est très utilisé (Nooralahzadeh *et al.*, 2020; Gu *et al.*, 2018; Singh *et al.*, 2022; Langedijk *et al.*, 2022).

L'idée derrière MAML est de trouver de bons paramètres initiaux θ qui peuvent être ajustés à de nouvelles tâches avec seulement quelques étapes d'optimisation et quelques exemples d'entraînement. Pour cela, MAML procède en *simulant l'entraînement et l'évaluation avec peu d'exemples* sur des tâches d'entraînement. En outre, le modèle est entraîné avec différentes tâches afin qu'il puisse apprendre à s'adapter rapidement à toutes les tâches similaires [2]. Dans les applications translingues, chaque tâche correspond à une langue différente, ce qui est l'objet de notre étude (cf section 3).

L'approche la plus proche de la nôtre est celle de Langedijk *et al.* (2022), qui adoptent MAML pour l'analyse syntaxique en dépendances translingue. Ils appliquent MAML pour apprendre de bons paramètres initiaux à partir de langues sources, puis s'évaluent sur un ensemble disjoint de langues cibles. Dans nos travaux, nous nous concentrons plutôt sur une tâche d'analyse *sémantique*. De plus, ils disposent de données d'entraînement multilingues, alors que nos données n'existent qu'en anglais. AMR est un formalisme non ancré donc une phrase anglaise et sa traduction devraient avoir le même graphe AMR. Par conséquent, nous générons nos données multilingues par la traduction automatique des données anglaises. Une autre approche similaire à la notre est celle de Sherborne & Lapata (2023) qui ont appliqué le méta-apprentissage à l'analyse sémantique SQL translingue. Bien qu'utile pour représenter (et exécuter) des requêtes de base de données exprimées en langage naturel, SQL n'est pas un formalisme sémantique polyvalent comme AMR. À notre connaissance, notre travail est donc le premier à utiliser MAML dans le cadre de l'analyse AMR translingue.

3 Meta X-AMR

Analyse AMR seq2seq Trois approches d'analyse AMR sont trés utilisées : l'analyse AMR par transitions (Damonte *et al.*, 2017), l'analyse AMR par graphe (*graph-based*, Zhang *et al.*, 2019; Cai

2. Tâches cibles avec une distribution similaire à celle des tâches sources.

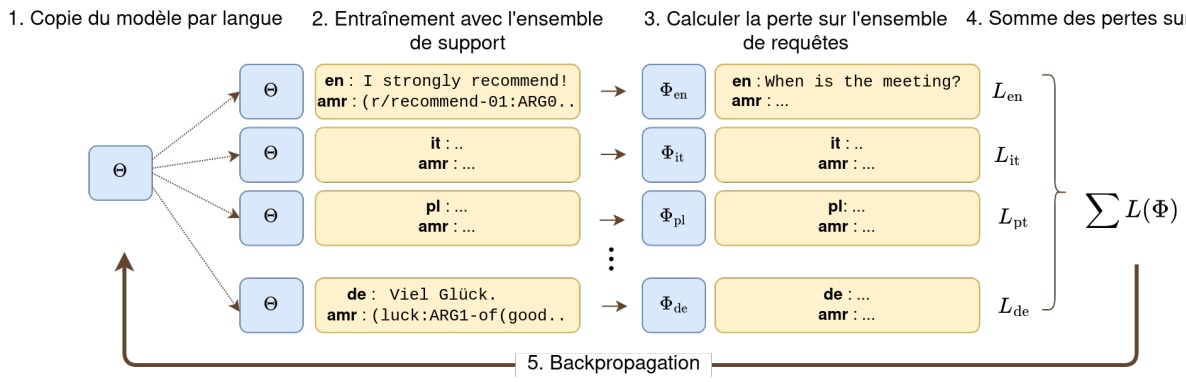


FIGURE 1 – Une étape d'entraînement MAML pour l'analyse AMR translingue.

& Lam, 2019), et l'analyse AMR seq2seq (Bevilacqua *et al.*, 2021). Nous utilisons cette dernière qui considère l'analyse AMR comme la génération d'un graphe AMR à partir de textes d'entrée à l'aide d'un modèle seq2seq. Dans cette approche, les graphes AMR doivent d'abord être linéarisés, c'est-à-dire représentés sous la forme d'une simple chaîne de caractères. Nous adoptons l'algorithme de parcours en profondeur pour la linéarisation comme Bevilacqua *et al.* (2021).

Nous utilisons le modèle mBart (Tang *et al.*, 2020) [3] pour entraîner notre analyseur AMR multilingue, comme Procopio *et al.* (2021). Le modèle mBart est un *transformer* (Vaswani *et al.*, 2017) pré-entraîné qui se compose de plusieurs couches d'encodeurs et de décodeurs. Comme le résultat de ce modèle est un graphe linéarisé, nous restructurons le graphe AMR par des étapes de post-traitement pour l'évaluation. Nous utilisons le code de van Noord & Bos (2017) [4] pour la linéarisation et la délinéarisation. Nous renvoyons les lecteurices à van Noord & Bos (2017) pour une description complète du processus.

MAML pour l'analyse AMR translingue Nous utilisons MAML (Finn *et al.*, 2017) pour entraîner notre analyseur AMR. L'objectif est d'entraîner un modèle qui s'adapte rapidement aux langues cibles avec aucun ou avec quelques exemples. La procédure d'entraînement est décrite ci-dessous et illustrée dans la Figure 1.

Étape 1 - Boucle interne : À chaque étape d'entraînement, le modèle initial (Θ) est copié une fois par langue i. Pour chaque langue i, $2 \times K$ exemples sont échantillonnés aléatoirement à partir de D_i^{train} et divisés en ensemble de support (*support set*) et ensemble de requête (*query set*) : K exemples pour chacun. En utilisant l'ensemble de support, le modèle est temporairement mis à jour avec une descente de gradient stochastique avec un taux d'apprentissage α (Eq. 1). Cette étape est répétée pendant P étapes d'adaptation afin d'obtenir Φ_i :

$$\Phi_i \leftarrow \Theta - \alpha \bigtriangledown_{\Theta} \mathcal{L}(\Theta_i). \tag{1}$$

Ensuite, la perte est calculée pour évaluer le modèle temporaire Φ_i sur l'ensemble de requêtes. La perte $\mathcal{L}_i(\Phi_i)$ est conservée pour l'étape suivante. L'ensemble de l'étape est appelé "boucle interne" et elle est répétée sur l'ensemble des tâches, c'est-à-dire pour le nombre de langues d'apprentissage. Ce nombre est désigné par I.

3. Nous utilisons le modèle `facebook/mbart-large-50` via la plateforme de *transformers* (Wolf *et al.*, 2020).
4. https://github.com/RikVN/AMR

Étape 2 - Boucle externe : La somme de $\mathcal{L}_i(\Phi_i)$ est calculée sur l'ensemble des langues d'apprentissage afin de mettre à jour le modèle initial Θ par descente de gradient stochastique avec un taux d'apprentissage β. L'ensemble de cette étape est appelé "boucle externe" [5] :

$$\Theta \leftarrow \Theta - \beta \sum_i \bigtriangledown_{\Phi_i} \mathcal{L}_i(\Phi_i). \tag{2}$$

Étape 3 : Répéter les étapes 1 et 2 jusqu'à ce que le nombre total d'étapes d'entraînement soit atteint.

Étape 4 : Une fois l'entraînement terminé, nous évaluons le modèle sur les langues cibles en zéro-*shot* ou *few-shot*. C'est-à-dire que le modèle est évalué sur de nouvelles langues cibles qui n'ont pas été vues pendant l'entraînement, sans affinage ou avec un affinage utilisant quelques exemples.

4 Expérimentations

4.1 Données

Notre méthode est similaire à celle de Langedijk *et al.* (2022) dans l'application du méta-apprentissage pour une tâche d'analyse translingue en *few-shot*. Pourtant, les données d'entraînement d'AMR ne sont disponibles qu'en anglais, alors qu'ils disposent de données d'entraînement multilingues pour l'analyse syntaxique en dépendance. Pour créer des données d'entraînement d'AMR multilingues, nous appliquons la traduction automatique comme dans les approches précédentes (Damonte & Cohen, 2018; Xu *et al.*, 2021; Blloshmi *et al.*, 2020). Nous utilisons DeepL [6] pour la traduction automatique et traduisons les données d'entraînement AMR anglaises (LDC2020T02 Knight *et al.*, 2020) en 13 langues : allemand, italien, roumain, finnois, russe, turc, japonais, tchèque, néerlandais, polonais, suédois, estonien et indonésien. Les 13 langues ont été choisies pour leur diversité linguistique et couvrent 5 familles de langues : indo-européen (germanique, roman, slave), ouralien, turcique, japonique et austronésien. Nous utilisons un total de 14 langues, dont l'anglais, pour nos données d'entraînement.

Pour sauvegarder le meilleur modèle pendant l'entraînement, nous évaluons notre modèle en k-*shot*. Pour cela, nous avons besoin des données de validation ainsi que de k exemples d'affinage dans la même langue. Pour les données de validation, nous utilisons les données de test en **espagnol** du corpus AMR 2.0 (Damonte & Cohen, 2020). Pour les données d'affinage, nous traduisons k exemples aléatoires issus des données de validation anglaises vers l'espagnol. Une fois l'entraînement terminée, nous évaluons notre modèle sur les langues cibles : le français, le chinois, le farsi, le coréen, et le croate. Pour les trois premières langues, nous utilisons le corpus AMR du Petit Prince annoté dans chaque langue, respectivement à partir de Kang *et al.* (2023), `https://amr.isi.edu/` et Takhshid *et al.* (2022) [7]. Pour le croate et le coréen, nous créons nos données de test en alignant manuellement le corpus du Petit Prince dans chaque langue sur les graphes AMR anglais correspondants. Nous mettons

5. Notez que dans l'équation 2, nous utilisons $\bigtriangledown_{\Phi_i} \mathcal{L}_i(\Phi_i)$ au lieu de $\bigtriangledown_{\theta} \mathcal{L}_i(\Phi_i)$ parce que nous appliquons MAML du premier ordre (*first-order* MAML) pour éviter le calcul très coûteux de la dérivée seconde.

6. `https://www.deepl.com`

7. L'ensemble de données original en farsi consiste en des graphes AMR dont les nœuds sont en farsi. Étant donné que nous utilisons des graphes AMR avec des nœuds en anglais, nous n'utilisons que les textes d'entrée du corpus et les graphes du corpus AMR en anglais.

ces données de test à la disposition du public [8].

4.2 Méta-entraînement et évaluation

Nous adoptons le modèle `mbart-large-50` (Tang *et al.*, 2020) de la bibliothèque `transformers` (Wolf *et al.*, 2020) pour entraîner notre analyseur AMR multilingue. Pour mettre en œuvre MAML, nous utilisons la bibliothèque `learn2learn` (Arnold *et al.*, 2020). Nous entraînons notre modèle pendant 30 000 étapes et l'évaluons toutes les 500 étapes avec les données de validation en espagnol. Nous mettons fin à l'entraînement si le score SMATCH (Cai & Knight, 2013) de validation ne s'améliore pas pendant plus de 7 500 étapes (interruption précoce de l'entraînement, *early stopping*). Pour la validation et le test, nous utilisons l'évaluation en k-shot, où le modèle est affiné avec k exemples avant d'être évalué sur l'ensemble de test/validation. Le nombre de cycles d'affinage, appelé étape d'adaptation, est noté P. Sauf indication contraire, nous fixons $P = 0$ et $k = 0$ (évaluation en 0-*shot*). MAML nécessite deux taux d'apprentissage, un pour la boucle interne (α) et un pour la boucle externe (β). Nous avons effectué une recherche par quadrillage (*grid search*) pour identifier un ensemble optimal de taux d'apprentissage et avons utilisé $\alpha = 1 \times 10^{-5}$, $\beta = 3 \times 10^{-5}$ tout au long des expériences. Pour β, nous utilisons un taux d'apprentissage qui croît linéairement jusqu'à β sur les 1 500 premières étapes. Sauf indication contraire, nous appliquons 1×10^{-5} pour affiner le modèle pendant la validation et le test. À chaque étape d'itération pendant l'entraînement, $2 \times K$ sont échantillonnés pour former un lot (*batch*) à partir des ensemble de requêtes et de support. Par conséquent, la taille de lot N est égale à $2 \times K \times I$, où I représente le nombre de langues d'apprentissage. Par défaut, nous attribuons $K = 8$ et $I = 14$, sauf indication contraire. Nous présentons les scores d'évaluation en utilisant SMATCH (Cai & Knight, 2013), une métrique d'évaluation pour les graphes AMR.

4.3 Modèle de base avec apprentissage conjoint

Nous entraînons un modèle de base avec une méthode d'apprentissage conjoint où plusieurs tâches sont apprises simultanément afin d'améliorer les performances globales du modèle. Le modèle mBart est utilisé comme décrit à la section 4.2. Pour l'entraînement, nous utilisons les données d'entraînement AMR en 14 langues listées à la section 4.1. Nous concaténons ces données multilingues et à chaque étape d'itération, nous sélectionnons aléatoirement N exemples d'entraînement à partir de ces données pour calculer la perte et optimiser le modèle en conséquence. Le modèle est évalué en 0-*shot* ou k-*shot* selon le cadre de l'expérience (les détails sont décrits dans chaque paragraphe de la section 5). Il convient de noter que notre objectif est de réaliser une étude comparative avec l'approche du méta-apprentissage. Par conséquent, sauf mention contraire, nous appliquons les mêmes hyperparamètres et la même méthode de test/évaluation pour les deux approches (par exemple, la taille des lots, la taille de k-*shot*). Cependant, alors que le méta-apprentissage nécessite deux taux d'apprentissage pour une boucle interne et une boucle externe, le modèle de base ne nécessite qu'un seul taux d'apprentissage pendant l'entraînement. Nous utilisons un taux d'apprentissage uniforme pour l'apprentissage 3×10^{-5} avec une croissance linéaire pendant les 1 500 premières étapes d'entraînement.

8. https://github.com/Emvista/Meta-XAMR-2024

	fr	zh	ko	fa	hr	avg
base_DeepL	56.3	45.6	42.1	46.3	**51.4**	48.3
base_mBart	56.2	44.5	41.2	46.1	51.3	47.8
MAML_DeepL	**56.5**	**46.1**	**42.2**	**46.7**	50.8	**48.4**
MAML_mBart	55.6	45.1	40.8	46.1	48.9	47.3

TABLE 1 – SMATCH en fonction de la source de traduction.

5 Questions de recherche et discussions

Nous examinons les points forts et les points faibles de notre méthode en répondant aux questions de recherche ci-dessous. Pour l'évaluation, nous faisons varier systématiquement cinq hyperparamètres individuellement tout en gardant les autres paramètres fixes et évaluons leur influence sur la performance du modèle. Nous évaluons chaque modèle en le comparant à son modèle de base adverse et apportons des éléments de réponses à 6 questions. Les questions Q1 et Q2 portent sur la manière dont les deux modèles réagissent à des facteurs spécifiques pendant la phase d'apprentissage, tandis que les questions Q3 à Q5 concernent les phases d'affinage fin et d'évaluation. Les discussions sur les questions conduisent à une discussion finale Q6 sur la question : le méta-apprentissage est-il l'approche optimale pour l'analyse AMR translingue ?

Q1 : Quelle est la robustesse du modèle en ce qui concerne la qualité de la traduction ? Pour évaluer l'impact de la traduction sur notre méthode, nous utilisons un autre modèle de traduction pour traduire nos données d'apprentissage. Plus précisément, nous utilisons les modèles de traduction mBart, provenant du hub Huggingface [9], pour traduire nos données d'entraînement en 13 langues. Ensuite, nous utilisons ces données traduites pour entraîner les modèles MAML et les modèles de base. Enfin, nous comparons les résultats de l'évaluation de ces modèles avec ceux entraînés à l'aide du système de traduction automatique DeepL.

Resultats Pour les modèles MAML et les modèles de base, l'utilisation d'un modèle de traduction open-source mBart entraîne une baisse des performances (voir tableau 1). Dans les deux cas, le score SMATCH du coréen diminue le plus lors de l'utilisation du modèle de traduction mBart. Le modèle MAML est plus affecté par ce changement. En moyenne, le modèle de base perd 0,9 %, tandis que le modèle MAML perd 2,3 %. Ce résultat montre que le modèle de méta-apprentissage est plus sensible à la qualité des textes traduits que le modèle de base.

Q2 : Le modèle apprend-il efficacement dans des environnements avec moins de ressources ? Nous évaluons la robustesse de notre méthode dans les environnements avec moins de ressources où seule une petite partie des données d'apprentissage est disponible. À cette fin, nous échantillonnons au hasard 1 000 exemples pour chaque langue (les mêmes exemples pour toutes les langues) et utilisons uniquement ces données échantillonnées comme données d'entraînement.

Résultats Le tableau 2 illustre les scores SMATCH obtenus par les modèles MAML et les modèles de base dans différentes conditions d'entraînement : en utilisant les données complètes (base_full, MAML_full) ou en utilisant seulement 1 000 exemples (base_1000, MAML_1000). Sans surprise, les

9. https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

	fr	zh	ko	fa	hr	avg
base_full	56.3	45.6	42.1	46.3	**51.4**	48.3
base_1000	41.4	35.1	33.3	36.9	38.5	37.0
MAML_full	**56.5**	**46.1**	**42.2**	**46.7**	50.8	**48.4**
MAML_1000	38.9	33.9	32.8	36.1	35.0	35.3

TABLE 2 – SMATCH en fonction de la taille des données d'apprentissage.

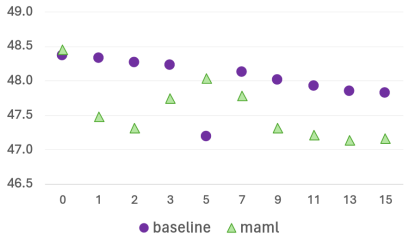


FIGURE 2 – Scores moyens de SMATCH sur les langues cibles en fonction des étapes d'adaptation.

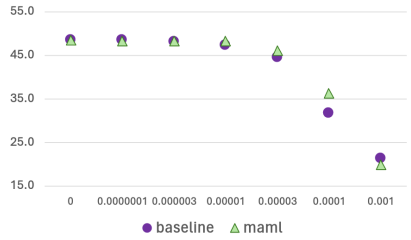


FIGURE 3 – Scores moyens de SMATCH sur les langues cibles en fonction du taux d'apprentissage de l'affinage fin.

performances des deux modèles ont considérablement diminué lorsqu'ils ont été entraînés sur un petit ensemble de données. Plus précisément, le modèle MAML connaît une baisse plus importante du score SMATCH, de 27 %, par rapport au modèle de base, qui a enregistré une baisse de 23,3 %. Cette différence suggère que le modèle MAML est plus sensible à la dégradation des performances dans les scénarios à faibles ressources.

Q3 : De combien d'étapes d'adaptation le modèle a-t-il besoin pour apprendre efficacement une nouvelle tâche ? Nous affinons nos modèles avec 32 exemples sur les langues cibles, puis nous les évaluons en langues cibles, pour l'évaluation *k-shot*. Étant donné que les données d'affinage ne sont pas disponibles pour les langues cibles, nous utilisons DeepL pour traduire les données de validation anglaises afin d'obtenir des données. Le modèle est ajusté de manière itérative avec les données de l'affinage et le nombre d'étapes d'adaptation est le nombre d'époques lors de l'affinage finale *k-shot*. Pour évaluer l'influence des étapes d'adaptation sur les performances du modèle, nous augmentons le nombre d'itérations et évaluons le modèle en conséquence. Nous échantillonnons 32 exemples de manière aléatoire à trois reprises et utilisons le score moyen des trois processus d'évaluation. Le taux d'apprentissage de l'affinage est fixé à 1×10^{-5} dans toutes les expériences.

Résultats La figure 2 [10] représente visuellement les scores moyens des tests SMATCH dans les langues cibles. Lorsque l'étape d'adaptation est égale à 0, le modèle est évalué en mode zéro-*shot*. De manière surprenante, les résultats indiquent que les modèles MAML et celui de base sont moins efficaces après l'adaptation. Nous émettons l'hypothèse que le modèle pré-entraîné mBart possède déjà une connaissance suffisante de nos langues cibles, et qu'affiner le modèle avec seulement quelques exemples dans chaque langue peut nuire à la capacité du modèle. Cela peut également être attribué à la différence de domaine entre les données de d'affinage et les données de test. Le premier comprend du contenu provenant de domaines généraux tels que des actualités, des forums en ligne, et des journaux, tandis que le données de test se composent du roman écrit dans les années 1940, *Le Petit*

10. Les données numériques seront également présentées dans l'annexe en version finale

k_size	baseline	MAML
0	48.3	48.4
32	48.2	47.3
64	48.2	47.7
128	**48.5**	**48.5**

TABLE 3 – Scores SMATCH du modèle de base et du modèle MAML en fonction de k

	fr	zh	ko	fa	hr	avg
baseline	56.4	45.6	42.1	46.3	**51.4**	48.4
MAML	**56.5**	**46.1**	**42.2**	**46.7**	50.8	**48.5**

TABLE 4 – Scores SMATCH du modèle de base et du modèle MAML (évaluation à zéro-*shot*).

Prince. Par conséquent, le changement de domaine entre les 2 jeux de données peut avoir contribué à l'incapacité du modèle à s'adapter efficacement au domaine de test. Une autre hypothèse est la taille restreinte des données d'affinage, qui peut avoir entravé la performance du modèle, ou un taux d'apprentissage inadéquat conduisant aux résultats d'affinage indésirables. Nous approfondissons les hypothèses sur le taux d'apprentissage et la taille de k dans les questions suivantes.

Q4 : Taux d'apprentissage élevé ou faible pour l'affinage ? Pour examiner les performances du modèle en fonction des différents taux d'apprentissage, nous affinons notre modèle avec différents taux d'apprentissage. Nous appliquons les mêmes paramètres que dans Q4, tels que l'échantillonnage des données trois fois avec une taille k égale à 32.

Résultats La figure 3 [11] présente une représentation visuelle des scores SMATCH moyens du test dans les différentes langues cibles. Le modèle de base et le modèle MAML présentent un comportement similaire, à savoir qu'un taux d'apprentissage plus faible permet d'obtenir de meilleurs résultats. Lorsque le taux d'apprentissage est égal à 0, c'est-à-dire lorsque le modèle n'est pas affiné, les deux modèles affichent les meilleures performances. Cela correspond aux résultats de Q3, mais on peut se demander pourquoi l'affinage dans les langues cibles ne conduit pas à un gain de performance. Cela peut être dû à la petite taille de k et dans la question suivante, nous évoquerons les résultats avec une plus grande taille de k.

Q5 : k-*shot*, quel est l'effet de la taille de k ? Pour répondre à cette question, nous utilisons différentes tailles de données d'affinage $k = 0, 32, 64, 128$. Comme pour Q3 et Q4, les données d'affinage sont échantillonnées trois fois et nous utilisons le score moyen. Nous appliquons le taux d'apprentissage 1×10^{-5} pour affiner les modèles.

Résultats Le tableau 3 montre que pour des valeurs de $32 \leq k \leq 128$, plus k est grand, plus le score est élevé. Cependant, à l'exception des modèles avec 128 exemples d'affinage, la plupart des modèles ne présentent pas d'amélioration par rapport à l'évaluation 0-*shot*. Il semble paradoxal qu'un modèle affiné soit moins performant qu'un modèle non affiné. Le modèle MAML est particulièrement affecté par l'étape d'affinage et présente une baisse de performance plus importante que le modèle de base. La baisse la plus importante est observée entre le modèle à 0 *shot* et le modèle à 32 *shot*, avec une différence de 2,3 %, alors que le modèle de base à 32 *shot* ne se dégrade que de 0,2 % par rapport au modèle à 0 *shot*. Par conséquent, cela nous amène à revoir les hypothèses discutées dans Q3 concernant la connaissance préalable du modèle mBart dans nos langues cibles et le changement de domaine entre les données d'affinage et celles du test.

11. Les données numériques seront également présentées dans l'annexe en version finale

Q6 : Comment l'analyse AMR translingue doit-elle être mise en place ? Le tableau 4 résume les scores SMATCH les plus élevés obtenus par les modèles de base et MAML lors de l'évaluation zéro-*shot*. La différence de performance entre ces modèles est marginale et varie en fonction de la langue cible. Par conséquent, il est difficile de tirer une conclusion définitive quant à la supériorité de l'une des méthodes. Cependant, notre examen nous a permis de constater que les modèles MAML présentent une plus grande sensibilité aux changements dans les types d'entrée et les tailles des données d'apprentissage. Notamment, leurs performances se détériorent de manière significative dans les scénarios avec peu de ressources ou lors de l'utilisation de différents modèles de traduction pour les entrées. En outre, des incohérences apparaissent lors de l'affinage du modèle avec différentes étapes d'adaptation, ce qui complique l'interprétation des résultats et rend difficile l'identification des perspectives d'amélioration. À l'inverse, nos observations indiquent qu'une approche simple d'apprentissage conjoint permet d'obtenir des performances comparables à celles du modèle MAML. Cela montre que la méthode d'apprentissage conjoint reste un point de départ solide pour l'analyse AMR translingue. Par conséquent, MAML n'apparaît pas comme la solution optimale pour cette tâche, compte tenu de ses performances instables.

Limites Notre modèle n'est pas plus performant qu'un simple modèle monolingue entraîné avec les données AMR dans la langue cible traduite par un système de traduction automatique. Cependant, notre approche peut être utilisée pour des langues peu dotées en ressources. En outre, nous n'avons pas appliqué de recherche par quadrillage (*grid-search*) pour trouver les meilleurs taux d'apprentissage pour les modèles de base et avons utilisé le même taux d'apprentissage que Procopio *et al.* (2021), qui a également utilisé mBart pour l'analyse AMR translingue en seq2seq. Cela a pu affecter les résultats en faveur du méta-apprentissage. Néanmoins, cela n'affecte pas notre conclusion de l'étude empirique qui révèle la faiblesse de l'approche du méta-apprentissage pour l'analyse AMR translingue.

6 Conclusion

Cette étude examine l'efficacité du méta-apprentissage par rapport à l'apprentissage conjoint dans l'analyse AMR translingue. Nous évaluons nos modèles dans des langues peu ou pas étudiées pour l'analyse AMR, notamment le français, le chinois, le coréen, le farsi et le croate. Pour faciliter l'évaluation, nous développons de nouveaux jeux de tests pour le coréen et le croate et publions les données pour promouvoir la diversité des langues d'évaluation pour l'analyse AMR translingue. Nous explorons différents contextes afin d'effectuer une analyse approfondie du méta-apprentissage par rapport à l'apprentissage conjoint. Nos résultats suggèrent que la méthode d'apprentissage conjoint est une approche de base robuste, tandis que le méta-apprentissage semble être une approche non optimale pour l'analyse AMR translingue en raison de ses performances peu robustes aux variations de configurations expérimentales.

Remerciements

Nous remercions les relecteurices anonymes pour leurs retours. Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2023-AD011012853R1 attribuée par GENCI.

Références

ARNOLD S. M. R., MAHAJAN P., DATTA D., BUNNER I. & ZARKIAS K. S. (2020). learn2learn : A library for Meta-Learning research. *arXiv*, **abs/2008.12284**.

BANARESCU L., BONIAL C., CAI S., GEORGESCU M., GRIFFITT K., HERMJAKOB U., KNIGHT K., KOEHN P., PALMER M. & SCHNEIDER N. (2013). Abstract Meaning Representation for sembanking. In A. PAREJA-LORA, M. LIAKATA & S. DIPPER, Éds., *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, p. 178–186, Sofia, Bulgaria : Association for Computational Linguistics.

BANSAL T., JHA R. & MCCALLUM A. (2020). Learning to few-shot learn across diverse natural language classification tasks. In D. SCOTT, N. BEL & C. ZONG, Éds., *Proceedings of the 28th International Conference on Computational Linguistics*, p. 5108–5123, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.448.

BEVILACQUA M., BLLOSHMI R. & NAVIGLI R. (2021). One spring to rule them both : Symmetric amr semantic parsing and generation without a complex pipeline. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**(14), 12564–12573. DOI : 10.1609/aaai.v35i14.17489.

BLLOSHMI R., TRIPODI R. & NAVIGLI R. (2020). XL-AMR : Enabling cross-lingual AMR parsing with transfer learning techniques. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 2487–2500, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.195.

CAI D. & LAM W. (2019). Core semantic first : A top-down approach for AMR parsing. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3799–3809, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1393.

CAI D., LI X., HO J. C.-S., BING L. & LAM W. (2021). Multilingual AMR parsing with noisy knowledge distillation. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2021*, p. 2778–2789, Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-emnlp.237.

CAI S. & KNIGHT K. (2013). Smatch : an evaluation metric for semantic feature structures. In H. SCHUETZE, P. FUNG & M. POESIO, Éds., *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 748–752, Sofia, Bulgaria : Association for Computational Linguistics.

DAMONTE M. & COHEN S. (2020). Abstract meaning representation 2.0 - four translations ldc2020t07.

DAMONTE M. & COHEN S. B. (2018). Cross-lingual Abstract Meaning Representation parsing. In M. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 1146–1155, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1104.

DAMONTE M., COHEN S. B. & SATTA G. (2017). An incremental parser for Abstract Meaning Representation. In M. LAPATA, P. BLUNSOM & A. KOLLER, Éds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, p. 536–546, Valencia, Spain : Association for Computational Linguistics.

DINGLIWAL S., GAO S., AGARWAL S., LIN C.-W., CHUNG T. & HAKKANI-TUR D. (2021). Few shot dialogue state tracking using meta-learning. In P. MERLO, J. TIEDEMANN & R. TSARFATY, Éds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 1730–1739, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.148.

FINN C., ABBEEL P. & LEVINE S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks.

GU J., WANG Y., CHEN Y., LI V. O. K. & CHO K. (2018). Meta-learning for low-resource neural machine translation. In E. RILOFF, D. CHIANG, J. HOCKENMAIER & J. TSUJII, Éds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 3622–3631, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1398.

HUA Y., LI Y.-F., HAFFARI G., QI G. & WU T. (2020). Few-shot complex knowledge base question answering via meta reinforcement learning. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 5827–5837, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.469.

KANG J., COAVOUX M., SCHWAB D. & LOPEZ C. (2023). Analyse sémantique AMR pour le français par transfert translingue. In C. SERVAN & A. VILNAT, Éds., *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : travaux de recherche originaux – articles courts*, p. 55–62, Paris, France : ATALA.

KNIGHT K., BADARAU B., BARANESCU L., BONIAL C., BARDOCZ M., GRIFFITT K., HERMJAKOB U., MARCU D., PALMER M., O'GORMAN T. & SCHNEIDER N. (2017). Abstract meaning representation (amr) annotation release 2.0 - linguistic data consortium.

KNIGHT K., BADARAU B., BARANESCU L., BONIAL C., BARDOCZ M., GRIFFITT K., HERMJAKOB U., MARCU D., PALMER M., O'GORMAN T. & SCHNEIDER N. (2020). Abstract meaning representation (amr) annotation release 3.0 - linguistic data consortium.

LANGEDIJK A., DANKERS V., LIPPE P., BOS S., CARDENAS GUEVARA B., YANNAKOUDAKIS H. & SHUTOVA E. (2022). Meta-learning for fast cross-lingual adaptation in dependency parsing. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 8503–8520, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.582.

NOORALAHZADEH F., BEKOULIS G., BJERVA J. & AUGENSTEIN I. (2020). Zero-shot cross-lingual transfer with meta learning. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 4547–4562, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.368.

O'GORMAN T., REGAN M., GRIFFITT K., HERMJAKOB U., KNIGHT K. & PALMER M. (2018). AMR beyond the sentence : the multi-sentence AMR corpus. In E. M. BENDER, L. DERCZYNSKI & P. ISABELLE, Éds., *Proceedings of the 27th International Conference on Computational Linguistics*, p. 3693–3702, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

PROCOPIO L., TRIPODI R. & NAVIGLI R. (2021). SGL : Speaking the graph languages of semantic parsing via multilingual translation. In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 325–337, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.30.

SHERBORNE T. & LAPATA M. (2023). Meta-learning a cross-lingual manifold for semantic parsing. *Transactions of the Association for Computational Linguistics*, **11**, 49–67. DOI : 10.1162/tacl_a_00533.

SHETH J., LEE Y.-S., FERNANDEZ ASTUDILLO R., NASEEM T., FLORIAN R., ROUKOS S. & WARD T. (2021). Bootstrapping multilingual AMR with contextual word alignments. In P. MERLO, J. TIEDEMANN & R. TSARFATY, Éds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 394–404, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.30.

SINGH S., WANG R. & HOU F. (2022). Improved meta learning for low resource speech recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 4798–4802. DOI : 10.1109/ICASSP43922.2022.9746899.

TAKHSHID R., SHOJAEI R., AZIN Z. & BAHRANI M. (2022). Persian abstract meaning representation. *arXiv*, **abs/2205.07712**.

TANG Y., TRAN C., LI X., CHEN P.-J., GOYAL N., CHAUDHARY V., GU J. & FAN A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv*, **abs/2008.00401**.

VAN NOORD R. & BOS J. (2017). Neural semantic parsing by character-based translation : Experiments with abstract meaning representations. *Computational Linguistics in the Netherlands Journal*, **7**, 93–108.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., SCAO T. L., GUGGER S., DRAME M., LHOEST Q. & RUSH A. M. (2020). Huggingface's transformers : State-of-the-art natural language processing.

XU D., LI J., ZHU M., ZHANG M. & ZHOU G. (2021). XLPT-AMR : Cross-lingual pre-training via multi-task learning for zero-shot AMR parsing and text generation. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 896–907, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.73.

ZHANG S., MA X., DUH K. & VAN DURME B. (2019). AMR parsing as sequence-to-graph transduction. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 80–94, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1009.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Recherche de relation à partir d'un seul exemple fondée sur un modèle *N-way K-shot* : une histoire de distracteurs

Hugo Thomas Guillaume Gravier Pascale Sébillot
Univ Rennes, INSA Rennes, CNRS, Inria, IRISA - UMR 6074
hugo.thomas@irisa.fr, guig@irisa.fr, pascale.sebillot@irisa.fr

RÉSUMÉ

La recherche de relation à partir d'un exemple consiste à trouver dans un corpus toutes les occurrences d'un type de relation liant deux entités dans une phrase, nommé type cible et caractérisé à l'aide d'un seul exemple. Nous empruntons le scénario d'entraînement et évaluation *N-way K-shot* à la tâche de classification de relations rares qui prédit le type de relation liant deux entités à partir de peu d'exemples d'entraînement, et l'adaptons à la recherche de relation avec un exemple. Lors de l'évaluation, un modèle entraîné pour la classification de relations en *N-way K-shot* est utilisé, dans lequel $K = 1$ pour le type cible, qui constitue l'une des N classes (du *N-way*), les *N-1* classes restantes étant des distracteurs modélisant la classe de rejet. Les résultats sur FewRel et TACREV démontrent l'efficacité de notre approche malgré la difficulté de la tâche. L'étude de l'évolution des performances en fonction du nombre de distracteurs et des stratégies de leur choix met en avant une bonne configuration globale, à savoir un nombre élevé de distracteurs à une distance intermédiaire du type de relation cible dans l'espace latent appris par le modèle. Le diagnostic *a posteriori* de notre méthode révèle l'existence de configurations optimales pour chaque type cible que nos analyses actuelles échouent à caractériser, ouvrant la voie à de futurs travaux.

ABSTRACT

One-shot relation retrieval based on an N-way K-shot model: a matter of distractors

One-shot relation retrieval consists in searching inside a corpus for all occurrences of a relation type linking two entities in an utterance, called the target type and characterized via a single example. In this paper, we perform this task by diverting the training and testing N-way K-shot scenario commonly used for the few-shot relation classification task, i.e., the prediction of the type of relation linking two entities with few training examples. At test time, our approach uses a model trained for N-way K-shot relation classification, in which $K = 1$ for the target type, which constitutes one of the N ways (i.e., N classes), the *N-1* others being distractors modeling the rejection class. Results on TACREV and FewRel demonstrate the effectiveness of our approach on the challenging task of one-shot relation retrieval. Investigating the influence of the number of distractors used and of their choice exposes an efficient overall configuration, i.e., a high number of distractors at an intermediate distance from the target relation embedding in the latent space learned by the model. Post-hoc analysis of our framework reveals target-type-dependent optimal configurations which our current experiments fail to characterize, paving the way for future work.

MOTS-CLÉS : extraction de relations, apprentissage frugal, apprentissage *N-way K-shot*.

KEYWORDS: relation extraction, few-shot learning, N-way K-shot learning.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 157–168.

1 Introduction

L'extraction de relations consiste à détecter, dans un texte, les relations liant deux entités (entités nommées ou pronoms leur faisant référence). Conceptuellement, cette tâche se divise en deux sous-tâches : la détection de relations, qui consiste à prédire s'il existe une relation entre deux entités présentes au sein d'une phrase, et la classification de relations, qui suppose l'existence d'une relation et vise à déterminer son type parmi un ensemble prédéfini. En pratique, ces deux sous-tâches sont souvent effectuées conjointement à l'aide d'un modèle qui choisit entre différents types de relations ou une classe de rejet. L'extraction de relations a fait l'objet de nombreux travaux, exploitant notamment des approches neuronales (Zhang *et al.*, 2017; Detroja *et al.*, 2023). Ces avancées restent cependant limitées par la quantité de données annotées nécessaires, conduisant à des performances moindres sur les types de relations rares, c.-à-d. pour lesquels peu d'exemples annotés sont disponibles. Dans cet article, nous nous penchons sur la tâche de recherche de relation avec un exemple. Cette tâche particulière vise à retrouver, dans un grand corpus, toutes les occurrences d'un type de relation cible identifié à l'aide d'un seul exemple. On suppose donc que cet exemple unique suffit à caractériser le type de relation, qui ne possède ni nom ni définition explicites et ne correspond à aucune classe connue d'un modèle. La recherche de relation avec un exemple trouve des cas d'applications concrets par exemple dans un contexte d'analyse de médias, en sociologie ou dans le cadre du journalisme d'enquête, où un utilisateur repère une relation d'intérêt (relation cible) lors de sa recherche et souhaite retrouver d'autres occurrences de ce type de relation dans un corpus.

Dans ce contexte, il n'est pas envisageable de ré-entraîner un nouveau modèle à chaque changement de type de relation cible, pour des questions de temps de calcul et d'absence de données d'entraînement en quantité suffisante. Pour traiter notre problématique, nous nous inspirons des méthodes de classification de relations rares s'appuyant sur des techniques d'apprentissage par transfert (Han *et al.*, 2018) qui permettent d'apprendre à représenter et classifier correctement des types de relations rares grâce à un apprentissage réalisé sur des types de relations fréquents. Ces méthodes sont souvent entraînées et évaluées selon le scénario *N-way K-shot*, dans lequel un *batch* d'entraînement est composé de N classes – des types de relations dans notre cas – tirées au hasard parmi des classes connues, chacune d'entre elles étant représentée par un petit nombre K de phrases supports issues des données d'entraînement. L'objectif consiste à classifier des énoncés – appelés requêtes – parmi l'une des N classes. Ce paradigme d'apprentissage adapté à un faible nombre d'exemples met en avant l'apprentissage de prototypes, c.-à-d. de vecteurs représentant chacun des N types de relations et construits à l'aide des K exemples disponibles. En pratique, les N prototypes d'un *batch* sont utilisés pour classifier chaque requête parmi les N types de relations en évaluant la distance – éventuellement apprise lors de l'entraînement – entre ces prototypes et le plongement de la requête appelé vecteur requête. Les modèles réalisant cette procédure sont nommés *modèles d'apprentissage de prototypes*, dont une typologie des architectures est définie par Dopierre *et al.* (2021).

Si le paradigme *N-way K-shot* est particulièrement pertinent pour la classification de relations rares (Han *et al.*, 2018), il ne répond pas directement à notre problématique de recherche de relation à partir d'un seul exemple et doit être fortement adapté pour ce faire. Dans cet article, nous proposons RaReMUD (*Rare Relation Mining Using Distractors*), un modèle dérivé de ce paradigme pour traiter cette tâche. Le modèle est d'abord entraîné pour la classification de relations dans le scénario *N-way K-shot* sur des types de relations fréquents afin d'apprendre à construire des prototypes de relations fiables et, le cas échéant, une métrique de comparaison des plongements (prototypes et vecteurs requêtes). Il est ensuite évalué sur des types de relations rares, toujours dans le scénario *N-way K-shot* : une des N classes au sein d'un *batch* correspond au type de relation cible représenté grâce à un seul

exemple support, les (prototypes des) *N-1* autres classes étant des *distracteurs* [1] modélisant une classe de rejet, c.-à-d. l'absence du type de relation cible. Cette phase d'évaluation correspond à la recherche d'occurrences du type de relation cible dans un corpus, les requêtes d'un *batch* étant à associer soit au type de relation cible, soit à un des distracteurs et par conséquent à rejeter. Nous évaluons plusieurs déclinaisons de notre méthode RaReMUD, incluant différents modèles d'apprentissage de prototypes en *N-way K-shot*, sur FewRel et TACREV, démontrant son efficacité sur la tâche de recherche de relation à l'aide d'un seul exemple et établissant une référence pour cette tâche. Nous mesurons l'influence des hyperparamètres de RaReMUD sur ses performances afin de déterminer une configuration globale idéale ; nos expériences portent en particulier sur l'impact du nombre de distracteurs utilisés lors de l'évaluation, de la stratégie de leur choix, et du nombre d'exemples pour le type de relation cible. Enfin, le diagnostic *a posteriori* de notre méthode révèle l'existence de configurations optimales pour chaque type cible que nos analyses actuelles échouent à caractériser, mettant en évidence une marge de gain de performances pour RaReMUD.

2 Travaux connexes

À notre connaissance, la recherche de relation à partir d'un seul exemple (ou de peu d'exemples) n'a pas été explorée à ce jour. Cette tâche se différencie de la classification de relations rares par deux aspects principaux : d'une part, c'est une tâche de détection et non de classification ; d'autre part, elle s'effectue en ensemble ouvert, le type de relation cible n'appartenant pas à un ensemble prédéfini. Si la recherche peut être redéfinie en une tâche de classification binaire (détection *vs.* rejet), la question-clé de la façon de modéliser l'absence de la relation cible demeure.

Malgré ces différences, les travaux les plus proches de notre problématique sont ceux traitant de la classification de relations rares qui visent la prédiction de types de relations parmi un ensemble prédéfini à l'aide de peu d'exemples d'entraînement. Les plus récents s'appuient sur l'apprentissage profond et tirent parti de modèles comme BERT ou GPT, entraînés et évalués dans le scénario *N-way K-shot* (Han *et al.*, 2021; Qu *et al.*, 2020; Sainz *et al.*, 2021; Gao *et al.*, 2020; Brody *et al.*, 2021; Chen *et al.*, 2023; Li *et al.*, 2024). Dans ce contexte, plusieurs solutions ont été proposées pour obtenir des modèles capables de rejeter des énoncés ne correspondant à aucun des N types de relations présents dans un *batch* : Tan *et al.* (2019) effectuent ce rejet par un apprentissage contrastif ; Gao *et al.* (2019) introduisent la notion de vecteur de rejet en ajoutant un vecteur en plus des N prototypes de relations d'un *batch*. Pendant l'entraînement, ce vecteur est appris directement sans être adossé à des exemples supports, contrairement aux prototypes de relations construits chacun à l'aide de K phrases supports. Cette idée a été étendue par Sabo *et al.* (2021) qui proposent d'apprendre plusieurs vecteurs de rejet pour mieux modéliser cette classe de rejet. RaReMUD emprunte cette idée de vecteurs de rejet multiples pour modéliser l'absence du type de relation cible, mais fonde son modèle de rejet sur des prototypes de relations réelles – appelés distracteurs – contruits à l'aide des phrases supports existantes et non sur un apprentissage de paramètres du modèle de classification comme Sabo *et al.* (2021). Ce choix résulte du fait que, dans un contexte d'ensemble ouvert, la classe de rejet dépend du type de relation cible et ne peut être apprise de manière figée.

1. Par souci de simplification, nous employons, dans l'article, indifféremment le terme *distracteur* pour faire référence au type de relation concerné ou à son prototype.

3 Tâche et méthodologie

Dans cette partie, nous introduisons le formalisme de la tâche et rappelons les principes des modèles d'apprentissage de prototypes, avant de détailler notre méthode et nos choix d'implémentation.

3.1 Définition de la tâche

En recherche de relation, les exemples de relations sont des quadruplets de la forme (s, t, r, q), où la phrase s est support du type de relation $r \in \mathcal{R}$ entre les entités tête et queue $(t, q) \in \mathcal{E}^2$. Les entités considérées sont des entités nommées ou des pronoms leur faisant référence. Par exemple, la phrase $s =$ « La capitale de la Bolivie est Sucre. » est support du type de relation $r =$ « capitale de » en considérant les entités $t =$ « Sucre » et $q =$ « Bolivie ». On suppose qu'un couple d'entités dans une phrase ne peut être support que d'un type de relation mais qu'il peut y avoir autant de relations dans une même phrase qu'il y a de paires d'entités dans celle-ci. En recherche de relation avec un exemple, un seul quadruplet est disponible pour caractériser le type de relation cible r, la tâche consistant à rechercher tous les quadruplets de ce même type de relation r dans un corpus.

3.2 Apprentissage de prototypes

En apprentissage de prototypes, dans chaque *batch*, un modèle apprend à construire un prototype par type de relation (soit *N* prototypes) en s'appuyant sur la structure des *batches* du scénario *N-way K-shot*. Ces derniers sont composés d'un ensemble support noté $\mathcal{S}$ et d'un ensemble requête noté $\mathcal{Q}$. L'ensemble support $\mathcal{S} = \{s_k^i; i \in [1, N], k \in [1, K]\}$ contient *N* types distincts de relations, chacun avec *K* quadruplets supports. L'ensemble requête $\mathcal{Q} = \{s_j; j \in [1, L]\}$ est composé de L phrases et leurs paires d'entités associées, auxquelles on devra attribuer un des *N* types de relations de l'ensemble support. Les prototypes de relations sont construits grâce aux plongements de leurs exemples supports, appelés vecteurs supports, et chaque requête $s_j \in \mathcal{Q}$ est associée à son plongement, appelé vecteur requête. À l'aide de ces vecteurs, le modèle prend une décision de classification fondée sur la distance entre les vecteurs requêtes et les prototypes. Conceptuellement, les approches d'apprentissage de prototypes requièrent les trois composants suivants, dont les paramètres sont ajustés en utilisant une fonction de coût pour la classification fondée sur l'entropie croisée :

- l'encodage de relations qui représente une phrase et sa paire d'entités par un vecteur de relation : vecteur requête pour toute requête de $\mathcal{Q}$, et vecteur support (servant à la création d'un prototype) pour chaque support de $\mathcal{S}$;
- la construction des prototypes, qui associe *K* vecteurs issus de phrases supports d'un type de relation de $\mathcal{S}$ à un vecteur unique nommé prototype ;
- la comparaison des vecteurs requêtes et des prototypes fondée sur une mesure de distance.

L'apprentissage *N-way K-shot* simule une classification de relations rares au niveau de chaque *batch* mais, globalement, le modèle peut être entraîné sur une grande quantité de données annotées, perdant ainsi la notion de rareté des données. En pratique, deux ensembles de types de relations disjoints sont constitués, $\mathcal{T}_{\text{fr}}$ et $\mathcal{T}_{\text{ra}}$, et les exemples du jeu de données sont séparés selon ces deux ensembles, formant $\mathcal{D}_{\text{fr}}$ et $\mathcal{D}_{\text{ra}}$. Les composants du modèle sont entraînés sur le corpus $\mathcal{D}_{\text{fr}}$, et évalués sur le corpus $\mathcal{D}_{\text{ra}}$ afin de mesurer les performances du modèle sur des types de relations rares en pratique au sein de chaque *batch*.

3.3 RaReMUD, une approche *N-way K-shot* pour la recherche de relation avec un exemple

Notre approche RaReMUD repose sur un modèle d'apprentissage de prototypes entraîné pour la classification de relations en *N-way K-shot* et l'exploite pour la recherche de relation avec un exemple. Comme décrit dans l'introduction, la phase d'apprentissage considère une tâche classique de classification de relations et permet d'apprendre de manière optimale les composants du système listés ci-dessus. Il est ensuite exploité dans un contexte de recherche de relation à partir d'un exemple, ce dernier étant utilisé comme l'une des classes dans une approche *N-way K-shot* et complété par $N-1$ distracteurs modélisant la classe de rejet. Chaque distracteur est construit à l'aide de K exemples supports d'un type de relation fréquent tirés aléatoirement, et est figé durant la recherche ; le choix de fonder les distracteurs sur des types de relations de $\mathcal{T}_{\text{fr}}$ est justifié par la nécessité d'avoir plusieurs exemples annotés pour construire des prototypes fiables. Le prototype du type de relation cible, quant à lui, correspond au plongement du seul exemple annoté disponible. Lors de la recherche – correspondant à la phase d'évaluation –, RaReMUD détecte la présence ou l'absence du type de relation cible dans chaque requête, en comparant le vecteur de cette requête au prototype du type de relation cible et aux *N-1* distracteurs.

Décrire une configuration complète de l'approche RaReMUD nécessite de définir (a) les éléments d'architecture du modèle d'apprentissage de prototypes utilisés et (b) le nombre de distracteurs et la stratégie pour les choisir parmi les types de relations de $\mathcal{T}_{\text{fr}}$. Nous décrivons les quatre architectures étudiées, avant de détailler les choix possibles de distracteurs.

3.3.1 Architectures de modèles d'apprentissage de prototypes

Dans toutes les expériences réalisées, l'encodage d'une phrase et de sa paire d'entités en un vecteur de relation se fait selon Soares *et al.* (2019) : les entités sont délimitées par des balises (<E1> et </E1> pour l'entité tête, <E2> et </E2> pour l'entité queue) ; les *tokens* constituant la phrase sont encodés à l'aide d'un modèle transformeur, et le plongement ou vecteur de relation est obtenu en concaténant les vecteurs des *tokens* <E1> et <E2>. Le reste de l'architecture suit une des variantes de modèles d'apprentissage de prototypes décrites dans Dopierre *et al.* (2021), que nous présentons succinctement. Dans ProtoNet, les N prototypes sont obtenus en effectuant la moyenne des plongements correspondant aux K exemples supports de chaque classe, et la distance cosinus est utilisée pour comparer les vecteurs requêtes aux prototypes. Le modèle ProtoNet++ étend cette idée en améliorant les prototypes grâce à la prise en compte d'exemples non annotés : les vecteurs des exemples non annotés sont ajoutés aux prototypes avec une pondération fondée sur leur similarité cosinus à ces mêmes prototypes. L'approche MatchingNet s'affranchit de la construction des prototypes et compare directement un vecteur requête à chacun des K vecteurs supports d'un type de relation : la distance de la requête à ce type de relation est obtenue en moyennant les distances à ses K vecteurs supports. Enfin, RelationNet ajoute à ProtoNet un apprentissage de métrique s'appuyant sur un modèle neuronal de tenseurs (Socher *et al.*, 2013) en remplacement de la distance cosinus.

3.3.2 Choix des distracteurs

La clé de RaReMUD réside dans le choix, pour un type cible donné, des *N-1* types de relations de $\mathcal{T}_{\text{fr}}$ employés comme distracteurs. Une première approche, qui constituera notre référence, consiste

simplement à prendre tous les types de relations de $\mathcal{T}_{\text{fr}}$ comme distracteurs. Cette approche semble cependant peu optimale, notamment en termes de temps de calcul. Nous comparons dans la suite différentes stratégies pour choisir un nombre plus restreint de distracteurs adaptés à un type de relation cible défini par une phrase support et son prototype associé :

— un tirage *aléatoire* des types de relations : cette stratégie constitue une référence faible à titre de comparaison pour les stratégies suivantes ;
— les types de relations dont les prototypes sont *les plus proches* du prototype cible dans l'espace latent construit par le modèle : ces distracteurs permettent *a priori* une détection fine du type de relation cible en rapprochant les prototypes de la classe de rejet de celui de la classe cible ;
— à l'opposé, les distracteurs *les plus éloignés*, c.-à-d. les *N-1* classes les plus lointaines du type de relation cible dans l'espace latent, représentant un choix moins risqué en assurant de rejeter les exemples éloignés du type de relation cible ;
— des distracteurs *intermédiaires*, situés entre les plus proches et les plus éloignés dans l'espace latent : ce choix constitue potentiellement un juste milieu entre la finesse des distracteurs les plus proches et la sécurité des distracteurs les plus lointains ;
— des distracteurs *mixtes*, mélange à proportions égales des trois choix précédents de distracteurs, en espérant tirer parti du meilleur des trois.

4 Expériences

Les expériences décrites ci-après visent à évaluer RaReMUD et à étudier l'influence de ses hyperparamètres sur ses performances en recherche de relation à partir d'un exemple. Nous décrivons tout d'abord les jeux de données utilisés, puis fournissons des détails d'implémentation des modèles et procédures expérimentales. Nous analysons ensuite, à travers les expériences menées, la dépendance des performances de RaReMUD au choix des distracteurs, plus précisément au nombre de distracteurs employés et aux stratégies pour les choisir parmi $\mathcal{T}_{\text{fr}}$. Nous observons les liens entre le nombre d'exemples annotés pour le type de relation cible et les scores obtenus par RaReMUD, avant de mener une analyse *a posteriori* de notre approche étudiant ses performances dans sa configuration optimale.

4.1 Jeux de données

Les expériences s'appuient sur deux des jeux de données les plus populaires et récents pour la classification de relations : d'une part, TACREV (Alt *et al.*, 2020), une version revisitée de TACRED, qui contient une variété de types de relations, y compris un pour l'absence de relation (ignoré lors de nos expériences) ; d'autre part, FewRel 1.0 (Han *et al.*, 2018), spécifiquement conçu pour la classification de relations rares. Ces jeux de données subissent les prétraitements suivants : les exemples contenant des entités du jeu de données d'évaluation sont supprimés des données d'entraînement, et les exemples (erronés) dans lesquels une entité est seulement constituée d'un adjectif possessif (*his*, *her*, *their*...) sont retirés. 75 % des types de relations forment l'ensemble $\mathcal{T}_{\text{fr}}$ (soit 60 types pour FewRel et 29 pour TACREV), les 25% restants formant $\mathcal{T}_{\text{ra}}$ (20 types pour FewRel et 10 pour TACREV). Au final, FewRel contient 47 428 exemples et TACREV 21 773, séparés en trois parties : les exemples supports de types de relations appartenant à $\mathcal{T}_{\text{fr}}$ sont répartis en un jeu d'entraînement (70 % de ces exemples, soit 27 146 pour FewRel et 13 012 pour TACREV) et un jeu de validation (soit 6 296 pour FewRel et 5 436 pour TACREV), et les exemples supports de types appartenant à $\mathcal{T}_{\text{ra}}$ constituent le jeu d'évaluation (13 986 exemples pour FewRel et 3 325 pour TACREV). Pour chaque type de $\mathcal{T}_{\text{ra}}$,

trois exemples supports sont sélectionés afin de construire indépendamment trois prototypes. Ces prototypes et les distracteurs associés sont chacun leur tour comparés au corpus constitué de tous les autres exemples du jeu d'évaluation.

4.2 Détails des implémentations

Tous les modèles sont entraînés en *N-way K-shot* pour la classification de relations puis évalués, dans le cadre RaReMUD, pour la recherche de relation à partir d'un (ou des quelques) exemple(s). L'encodage des relations est réalisé par le modèle RoBERTa [2] (Liu *et al.*, 2019) adapté de manière efficace par LoRA (Hu *et al.*, 2021). L'entraînement est interrompu après trois époques sans amélioration de la mesure F1 sur le jeu de validation. Nous avons expérimentalement fixé pour l'entraînement $N = K = 5$, compromis efficace entre N et K en tenant en compte de la mémoire disponible [3]. Pour la variante ProtoNet++, le jeu de données *New York Times*, disponible comme jeu de validation de FewRel 2.0, est utilisé comme données non annotées pour augmenter les prototypes lors de l'apprentissage du modèle, du fait de sa disponibilité et de sa relative proximité du domaine des données d'entraînement. En pratique, la phase d'évaluation correspond à un scénario *N-way K-shot* avec $K = 1$ pour le type cible et $K = 5$ pour les distracteurs, traitant l'ensemble des énoncés requêtes du jeu d'évaluation avec $N = R + 1$.

Pour la stratégie de choix aléatoire des distracteurs, nous effectuons 5 tirages indépendants par expérience avec une graine aléatoire fixe par souci de reproductibilité. La mesure F1, nommée F-mesure par la suite, est la métrique d'évaluation retenue pour juger de la capacité de RaReMUD à retrouver les occurrences du type de relation cible parmi les requêtes, compromis discutable mais efficace entre précision et rappel. Les expériences présentées dans cet article ont été réalisées sur le banc d'essai Grid'5000, soutenu par un groupement d'intérêt scientifique hébergé par l'Inria et comprenant le CNRS, RENATER et plusieurs universités ainsi que d'autres organisations (voir `https://www.grid5000.fr`).

4.3 Influence du nombre de distracteurs

Nous étudions tout d'abord l'influence du nombre de distracteurs, noté R, utilisés lors de la phase d'évaluation. La figure 1 illustre l'évolution de la F-mesure de notre méthode en fonction de R sur FewRel pour les 4 modèles étudiés avec un tirage aléatoire des distracteurs. On constate tout d'abord une très forte variabilité de la F-mesure induite par les différences de scores obtenus selon les types de relations cibles, et par la diversité de prototype du type cible créé selon l'unique exemple annoté servant à le construire. Nous avons observé que cette variabilité est principalement due aux types de relations cibles, la diversité des prototypes ayant une influence plus modérée. Par exemple, avec ProtoNet sur FewRel et 20 distracteurs, les types "owned by" et "country", très variables, obtiennent de très faibles scores (resp. 3 % et 4 %) à l'inverse de "crosses" et "league", très spécifiques ($\sim$ 95 %). Par ailleurs, la F-mesure médiane pour chaque valeur de R est, de manière générale, relativement basse. Ces deux points sont à mettre en regard de la difficulté de la tâche de recherche de relation à partir d'un seul exemple. De manière plus fondamentale, il apparaît que la référence ($R = 60$) obtient une meilleure F-mesure qu'un faible nombre de distracteurs mais n'est pas optimale : les valeurs élevées de R inférieures à 60 semblent conduire à de meilleures performances, ce que nous

2. Le modèle `roberta-base` de Huggingface est utilisé.
3. Une implémentation de RaReMUD est disponible à `https://gitlab.inria.fr/huthomas/raremud`.

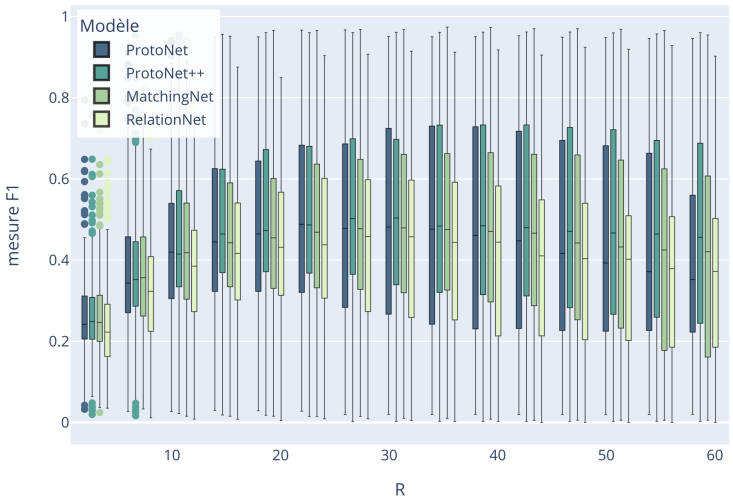


FIGURE 1 – F-mesure sur FewRel en fonction du nombre de distracteurs pour les différents modèles

TABLE 1 – F-mesure (en %) sur FewRel et TACREV pour 5 stratégies de choix des distracteurs.

	FewRel 1.0				TACREV			
Stratégie	ProtoNet	ProtoNet++	MatchingNet	RelationNet	ProtoNet	ProtoNet++	MatchingNet	RelationNet
plus proches	38.6±25.7	44.4±27.3	39.0±25.4	34.9±20.3	34.2±33.2	27.9±32.0	27.1±29.6	23.6±19.6
plus éloignés	42.6±20.6	41.2±20.2	41.4±21.7	40.2±21.3	35.4±32.6	34.9±32.1	35.0±32.8	30.8±27.8
intermédiaires	46.7±22.3	46.2±22.3	44.9±22.8	43.7±21.8	39.3±35.0	37.8±34.1	37.5±35.0	32.8±28.6
mixtes	36.9±24.8	43.1±26.1	37.9±24.5	35.0±20.5	33.5±32.2	27.2±30.5	28.7±30.3	25.3±20.7
aléatoires	43.7±25.0	46.8±25.4	44.0±24.5	39.2±21.3	37.2±34.3	34.0±34.6	32.9±33.4	28.5±24.1

avons confirmé par un test statistique de comparaison des moyennes de Student au risque de 5 %. Les résultats sur TACREV mènent aux mêmes conclusions. En résumé, RaReMUD obtient donc ses meilleures performances lorsqu'un grand nombre de distracteurs sont choisis parmi tous ceux disponibles, sans que ce nombre soit statistiquement discernable du fait de la grande variabilité des résultats.

4.4 Influence de la stratégie de choix des distracteurs

Nous comparons ensuite les stratégies de choix des distracteurs. Afin de s'affranchir du choix du nombre de distracteurs pour chaque stratégie, nous moyennons les F-mesures obtenues pour une stratégie de choix en faisant varier R. Le tableau 1 rend compte de cette évaluation pour les 4 modèles en fonction de la stratégie utilisée. Malgré la variabilité des résultats (dont les raisons ont été mentionnées en section 4.3), il apparaît que le choix naïf de tirer des distracteurs aléatoires est sous-optimal. La meilleure stratégie moyenne commune à tous les modèles et types de relations cibles semble être celle des distracteurs intermédiaires, constat confirmé par un test de comparaison des moyennes de Student au risque de 5%. Toutefois, certains types de relations cibles obtiennent marginalement de meilleures performances avec des distracteurs aléatoires – le type de relation cible *"country of citizenship"* de FewRel obtenant p. ex. une F-mesure de 51.54±14.48% avec les distracteurs aléatoires contre 41.15±10.40% avec les intermédiaires pour le modèle ProtoNet – ce qui indique que trouver la stratégie optimale pour un type cible donné reste une question ouverte.

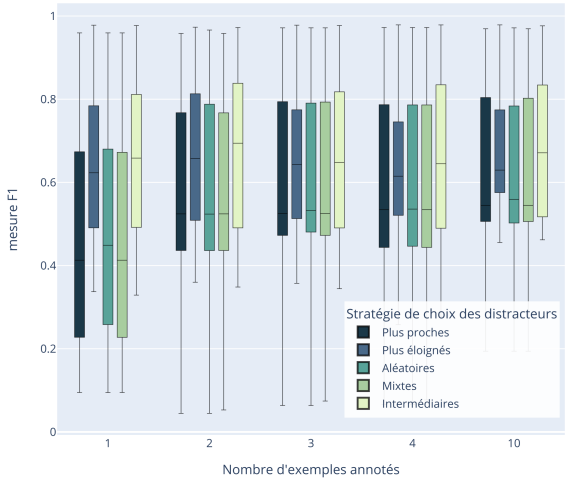


FIGURE 2 – F-mesure de ProtoNet sur FewRel en fonction du nombre d'exemples fournis

4.5 Influence du nombre d'exemples annotés pour la construction de prototype

Les expériences précédentes étant réalisées dans le cas extrême de recherche de relation avec un seul exemple, nous vérifions l'intérêt, par exemple dans un cadre applicatif, d'annoter plus d'un seul exemple pour un type de relation cible. Sur la figure 2, la F-mesure du modèle ProtoNet – l'utilisation des autres modèles menant aux mêmes conclusions – est mise en relation avec le nombre d'exemples fournis pour le type cible sur le jeu de données FewRel. Il apparaît que l'annotation de deux exemples apporte un gain de F-mesure visible par rapport à celle d'un unique exemple. Ce gain s'estompe toutefois rapidement lorsque des exemples supplémentaires sont annotés. Ces conclusions sont communes aux différentes stratégies de choix de distracteurs et au jeu de données TACREV, ce qui les renforce. Dans un cadre applicatif, l'annotation de deux exemples peut donc être recommandée comme un bon compromis entre gain de performance et temps d'annotation.

4.6 Analyse *a posteriori* du choix des distracteurs optimaux

La meilleure configuration globale trouvée jusqu'ici, qui consiste à choisir un nombre élevé de distracteurs à une distance intermédiaire du type de cible, n'est pas nécessairement optimale pour un type de relation cible donné. À des fins de diagnostic de notre approche, nous effectuons l'analyse *a posteriori* suivante : les performances sont évaluées avec toutes les combinaisons possibles de moins de 6 distracteurs – la combinatoire augmentant rapidement avec le nombre de distracteurs – pour chaque type cible afin de déterminer les performances optimales absolues de RaReMUD avec des distracteurs optimaux par type de relation cible. La figure 3 représente la F-mesure sur FewRel de ProtoNet pour plusieurs types de relations cibles avec les distracteurs intermédiaires en quantité R idéale révélée par les expériences de la section 4.3 (R=55 pour FewRel) en bleu foncé et avec les 6 distracteurs optimaux déterminés *a posteriori* pour chaque type cible (cette quantité étant choisie empiriquement comme optimale parmi les valeurs disponibles) en vert clair ; L'importante marge de progression entre les deux configurations révèle le potentiel considérable de notre approche sur la tâche pourtant difficile de recherche de relation avec un exemple. Par ailleurs, certains types cibles obtiennent des scores relativement faibles dans les deux configurations, indiquant vraisemblablement

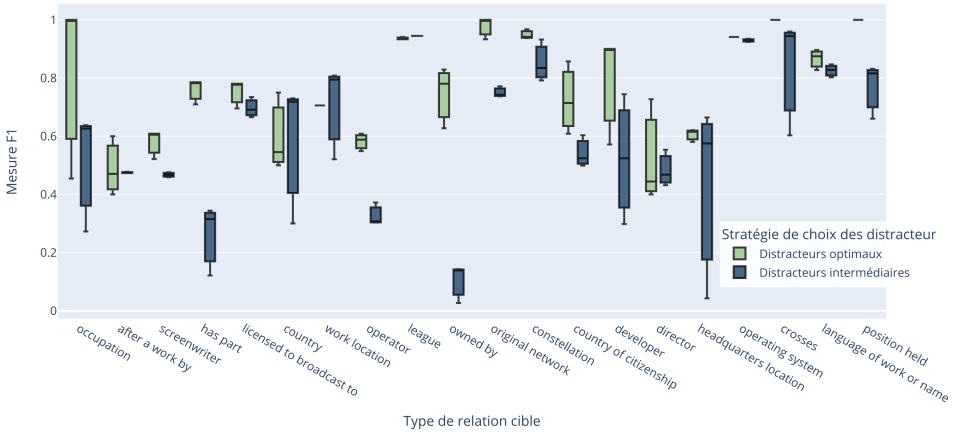


FIGURE 3 – Comparaison des F-mesures avec des distracteurs optimaux et intermédiaires sur les types de relations rares de FewRel pour le modèle ProtoNet

leur difficulté inhérente ou l'absence de distracteurs disponibles adaptés à ces types de relation. Ces deux constats ouvrent la voie à de futurs travaux.

5 Conclusion

Dans cet article, nous avons proposé et étudié RaReMUD (*Rare Relation Mining Using Distractors*), une approche adaptant le scénario de classification de relations rares *N-way K-shot* à la recherche de relation à partir d'un seul exemple. Nos résultats sur FewRel et TACREV, en fondant RaReMUD sur plusieurs architectures de modèles d'apprentissage de prototypes, démontrent l'efficacité de notre approche sur cette tâche exigeante. Il ressort aussi de nos expériences qu'un choix des distracteurs est nécessaire à l'optimisation de notre méthode, et reste une décision complexe pouvant dépendre du type de relation cible. Nous montrons qu'une stratégie fiable de choix consiste à conserver un nombre de distracteurs élevé, inférieur au nombre maximal, mêlant des types de relations de l'ensemble d'entraînement ni trop proches ni trop éloignés du type de relation cible dans l'espace des prototypes. L'étude de l'influence sur les performances du nombre d'exemples du type de relation cible disponibles souligne que le simple ajout d'un second exemple contribue à augmenter fortement la F-mesure des modèles, encourageant de potentiels travaux exploitant RaReMUD dans un cadre applicatif à annoter au moins deux exemples. Nos expériences révèlent enfin que la stratégie fiable de sélection des distracteurs que nous avons mise en évidence est sous-optimale puisqu'une évaluation exhaustive des combinaisons possibles de distracteurs peut conduire à un choix de distracteurs plus performant pour un type de relation cible donné ; la façon de faire ce choix *a priori* en se fondant sur le seul exemple support du type de la relation cible reste toutefois une question qu'il convient d'explorer. Des premières expériences sur des indicateurs simples – fréquence de la relation, densité autour du prototype – n'ont pas permis de corréler ces indicateurs au choix optimal des distracteurs. Enfin, à des fins pratiques et de contrôle expérimental, nos expériences ont, jusqu'à présent, été effectuées dans le paradigme de laboratoire *N-way K-shot* et gagneraient à se rapprocher de conditions d'utilisation réelles. Ceci requiert des approches de fouille plus efficaces que la comparaison exhaustive d'exemples avec N prototypes et des jeux de données dédiés à la tâche de recherche de relations rares, TACREV et FewRel étant conçus pour la classification de relations.

Références

ALT C., GABRYSZAK A. & HENNIG L. (2020). TACRED revisited : A thorough evaluation of the TACRED relation extraction task. In *Annual Meeting of the Association for Computational Linguistics*, p. 1558–1569.

BRODY S., WU S. & BENTON A. (2021). Towards realistic few-shot relation extraction. In *Conference on Empirical Methods in Natural Language Processing*, p. 5338–5345.

CHEN X., WU H. & SHI X. (2023). Consistent prototype learning for few-shot continual relation extraction. In *Annual Meeting of the Association for Computational Linguistics*, p. 7409–7422.

DETROJA K., BHENSDADIA C. & BHATT B. S. (2023). A survey on relation extraction. *Intelligent Systems with Applications*, **19**.

DOPIERRE T., GRAVIER C. & LOGERAIS W. (2021). A neural few-shot text classification reality check. In *Conference of the European Chapter of the Association for Computational Linguistics*, p. 935–943.

GAO T., HAN X., XIE R., LIU Z., LIN F., LIN L. & SUN M. (2020). Neural snowball for few-shot relation learning. In *Conference on Artificial Intelligence*, volume 34, p. 7772–7779.

GAO T., HAN X., ZHU H., LIU Z., LI P., SUN M. & ZHOU J. (2019). FewRel 2.0 : Towards more challenging few-shot relation classification. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, p. 6251–6256.

HAN J., CHENG B. & LU W. (2021). Exploring task difficulty for few-shot relation extraction. In *Conference on Empirical Methods in Natural Language Processing*, p. 2605–2616.

HAN X., ZHU H., YU P., WANG Z., YAO Y., LIU Z. & SUN M. (2018). FewRel : A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Conference on Empirical Methods in Natural Language Processing*, p. 4803–4809.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S. & CHEN W. (2021). Lora : Low-rank adaptation of large language models. *CoRR*, **abs/2106.09685**.

LI R., ZHONG J., HU W., DAI Q., WANG C., WANG W. & LI X. (2024). Adaptive class augmented prototype network for few-shot relation extraction. *Neural Networks*, **169**, 134–142.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized BERT pretraining approach. *CoRR*, **abs/1907.11692**.

QU M., GAO T., XHONNEUX L.-P. & TANG J. (2020). Few-shot relation extraction via Bayesian meta-learning on relation graphs. In *International Conference on Machine Learning*, p. 7867–7876.

SABO O., ELAZAR Y., GOLDBERG Y. & DAGAN I. (2021). Revisiting few-shot relation classification : Evaluation data and classification schemes. *Transactions of the Association for Computational Linguistics*, **9**, 691–706.

SAINZ O., DE LACALLE O. L., LABAKA G., BARRENA A. & AGIRRE E. (2021). Label verbalization and entailment for effective zero and few-shot relation extraction. In *Conference on Empirical Methods in Natural Language Processing*, p. 1199–1212.

SOARES L. B., FITZGERALD N., LING J. & KWIATKOWSKI T. (2019). Matching the blanks : Distributional similarity for relation learning. In *Annual Meeting of the Association for Computational Linguistics*, p. 2895–2905.

SOCHER R., CHEN D., MANNING C. D. & NG A. Y. (2013). Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*, volume 26, p. 926–934.

TAN M., YU Y., WANG H., WANG D., POTDAR S., CHANG S. & YU M. (2019). Out-of-domain detection for low-resource text classification tasks. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, p. 3566–3572.

ZHANG Q., CHEN M. & LIU L. (2017). A review on entity relation extraction. In *International Conference on Mechanical, Control and Computer Engineering*, p. 178–183.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Reconnaissance d'entités cliniques en *few-shot* en trois langues

Marco Naguib[1] Aurélie Névéol[1] Xavier Tannier[2]
(1) Université Paris-Saclay, CNRS, LISN, 91405 Orsay cedex, France
(2) Sorbonne Université, Inserm, Université Sorbonne Paris Nord, LIMICS, 75006 Paris, France
marco.naguib@lisn.upsaclay.fr, aurelie.neveol@lisn.upsaclay.fr,
xavier.tannier@sorbonne-universite.fr

RÉSUMÉ

Les grands modèles de langues deviennent la solution de choix pour de nombreuses tâches de traitement du langage naturel, y compris dans des domaines spécialisés où leurs capacités *few-shot* devraient permettre d'obtenir des performances élevées dans des environnements à faibles ressources. Cependant, notre évaluation de 10 modèles causaux (auto-régressifs) et 16 modèles masqués montre que, bien que les modèles causaux utilisant des prompts puissent rivaliser en termes de reconnaissance d'entités nommées (REN) en dehors du domaine clinique, ils sont dépassés dans le domaine clinique par des *taggers* biLSTM-CRF plus légers reposant sur des modèles masqués. De plus, les modèles masqués ont un bien moindre impact environnemental que les modèles causaux. Ces résultats, cohérents dans les trois langues étudiées, suggèrent que les modèles à apprentissage *few-shot* ne sont pas encore adaptés à la production de REN dans le domaine clinique, mais pourraient être utilisés pour accélérer la création de données annotées de qualité.

ABSTRACT

Few-shot learning for clinical entity recognition in three languages.

Large language models have become the preferred solution for many natural language processing tasks, including specialized domains where their few-shot capabilities should deliver high performance in low-resource environments. However, our evaluation of 10 auto-regressive models and 16 masked models shows that while prompt-based auto-regressive models can compete in named entity recognition (NER) outside the clinical domain, they are outperformed within the clinical domain by lighter biLSTM-CRF taggers based on masked models. Additionally, masked models have a much lower environmental impact than auto-regressive models. These consistent results across the three languages studied suggest that few-shot learning models are not yet suited for NER production in the clinical domain but could be used to expedite the creation of quality annotated data.

MOTS-CLÉS : Apprentissage en *few-shot* ; modèles de langues ; reconnaissance d'entités nommées.

KEYWORDS: *few-shot* learning ; large language models ; named entity recognition.

1 Introduction

Les documents cliniques représentent d'importantes sources d'informations (Demner-Fushman *et al.*, 2009), souvent présentées sous forme de texte non structuré (Escudié *et al.*, 2017). L'extraction efficace des informations de ces documents vers une forme plus structurée peut améliorer la recherche

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 169–197.

clinique, la surveillance de la santé publique et l'aide à la décision clinique automatique (Wang *et al.*, 2018).

La reconnaissance d'entités nommées (REN) constitue une étape cruciale de cette extraction d'informations. Elle consiste à identifier et à typer des mentions d'intérêt dans un texte. Dans le cadre de l'extraction d'informations cliniques, il s'agit notamment des entités cliniques telles que les maladies ou les médicaments. L'extraction de ces entités peut grandement faciliter la normalisation des concepts (Cho *et al.*, 2017; Wajsbürt *et al.*, 2021; Sung *et al.*, 2022) et l'interprétation du profilage et du phénotypage des patients (Gérardin *et al.*, 2022). Alors que la REN dans le domaine général (identification des entités telles que les personnes et les lieux) a été largement étudiée dans la communauté du traitement automatique des langues (TAL), la REN clinique est souvent considérée comme plus complexe : les entités cliniques sont souvent exprimées en jargon ou en termes ambigus, et les textes cliniques présentent une structure grammaticale non standard (Luo *et al.*, 2020; Leaman *et al.*, 2015).

Les modèles de langues sont progressivement devenus l'approche principale pour aborder la REN (Li *et al.*, 2022; Wang *et al.*, 2022). Des travaux antérieurs se sont concentrés sur la REN générale (Devlin *et al.*, 2019) ainsi que sur la REN clinique (Gérardin *et al.*, 2022; Sun *et al.*, 2021). Ces travaux peuvent être principalement divisés en deux approches, selon le type de modèles de langues utilisés.

La première approche consiste à utiliser des **modèles de langues masqués (MLM)** pré-entraînés. Ce type de modèles est d'abord pré-entraîné pour prédire des mots masqués sélectionnés au hasard dans de grands corpus de textes à l'aide d'une représentation vectorielle dense de chaque token (mot, par exemple) dans le texte (Devlin *et al.*, 2019; Peters *et al.*, 2018). Pour utiliser ces modèles pour la REN, on apprend généralement une projection linéaire à transformer les représentations vectorielles des mots en étiquettes désignant les entités nommées dans la phrase. En parallèle, on ajuste (*fine-tune)* également les paramètres du modèle de langues pour la tâche de REN. Cette approche a fait l'objet d'une grande attention de la part de la communauté, et est devenue la solution de référence pour la construction de systèmes de REN robustes.

Toutefois, cette approche rencontre deux principaux obstacles dans le contexte de la REN clinique. Tout d'abord, en raison de la nature sensible des documents cliniques, les corpus publics sont rares, soumis à des licences restrictives et peu disponibles dans des langues autres que l'anglais. Cela contraint la communauté à utiliser des solutions construites sur des MLM pré-entraînés principalement sur des corpus de domaine général, ce qui peut entraîner des problèmes de changement de domaine (*domain shift*). Deuxièmement, pour que l'entraînement soit efficace, de grands corpus de textes annotés dans le domaine d'intérêt sont nécessaires (Jia *et al.*, 2019; Liu *et al.*, 2021). Or, les campagnes d'annotation de REN clinique sont très coûteuses en temps et en ressources, nécessitant un haut niveau d'expertise du domaine pour être menées à bien (Luo *et al.*, 2020; Névéol *et al.*, 2014; Doğan *et al.*, 2014; Báez *et al.*, 2020). De plus, en raison de la diversité des cas cliniques, les données annotées pour une application biomédicale ne sont pas nécessairement transférables à une autre. D'où la nécessité de développer des approches de REN clinique efficaces en termes de données, également connues sous le nom de REN en *few-shot* (en peu d'exemples).

La deuxième approche, plus récente, consiste à utiliser des **modèles de langues causaux (CLM)** pré-entraînés. Ces modèles, considérablement plus grands, sont pré-entraînés sur des corpus (souvent plus importants) en tant que modèles génératifs et auto-régressifs. En d'autres termes, le modèle reçoit en entrée une série de tokens ou *prompt* et estime la série de tokens suivante la plus probable. Pour exploiter ces modèles de langues dans des tâches telles que la REN, il est possible de formuler la tâche

en langage naturel dans un *prompt*. Le *prompt* est conçu de manière à ce que la continuité du texte implique la résolution de la tâche. Le modèle de langue est ensuite utilisé pour prédire cette continuité. Ce processus est souvent appelé « *in-context learning* » (ICL) (Brown *et al.*, 2020). Éventuellement, il est possible de créer un *prompt* comprenant quelques exemples résolus de la tâche pour d'autres instances (dans ce cas, des instances annotées en entités nommées, spécifiques à la tâche), avant la nouvelle instance de test (Lee *et al.*, 2022). Le modèle produit ainsi une estimation de l'étiquetage en entités nommées le plus probable pour l'instance de test. Alors que les MLM ont été étudiés pour la REN en *few-shot* (Du *et al.*, 2021), les CLM semblent plus naturellement adaptés à ce contexte. L'apprentissage ICL a en fait démontré un succès particulier avec les CLM dans l'apprentissage en *few-shot*, montrant des résultats prometteurs dans un large éventail de tâches de TAL (Shin *et al.*, 2022; Wei *et al.*, 2022; Srivastava *et al.*, 2023).

Cependant, la supériorité des CLM sur les MLM pour la REN en *few-shot* est discutable. De nombreux efforts étudiant l'apprentissage en « *few-shot* » avec des CLM choisissent les *prompts* en fonction de leurs performances sur de grands jeux de données de validation (Brown *et al.*, 2020; Tam *et al.*, 2021; Radford *et al.*, 2021; Qin & Eisner, 2021). Cela pose un problème car il a été démontré que l'ICL dépendait fortement de la structure du *prompt* : un petit changement dans la formulation de la tâche, les exemples présentés, l'ordre des exemples ou le format d'étiquetage peut affecter la performance. Par conséquent, faire ces choix en supposant l'existence d'un grand jeu de données de validation annotées conduit à des résultats qui s'avèrent trop optimistes (Perez *et al.*, 2021) et impossibles à trouver dans un cadre réel de quelques exemples annotées. Deuxièmement, la plupart de ces études se sont principalement concentrées sur la langue anglaise et sur des modèles basés sur GPT (Wang *et al.*, 2023b; Ashok & Lipton, 2023; Hu *et al.*, 2023b; Jimenez Gutierrez *et al.*, 2022). Cela peut conduire à des *prompts* trop adaptés à cette langue et à ce modèle de langue. Il est donc nécessaire de mener une étude systématique, indépendante du modèle, sur l'élaboration des *prompts* dans le contexte clinique et pour des langues autres que l'anglais. Les contributions de ce travail sont les suivantes :

1. Nous présentons et comparons les techniques de *prompting* appliquées à la REN les plus récentes lorsqu'elles sont appliquées à la REN clinique dans trois langues : l'anglais, le français et l'espagnol. À notre connaissance, il s'agit du premier travail axé sur les *prompts* de REN pour les langues autres que l'anglais, et du premier travail comparant les *prompts* pour la REN clinique.
2. Nous accordons une attention particulière aux *prompts* de balisage, une technique proposée récemment (Wang *et al.*, 2023b), et nous mesurons les améliorations qu'elle apporte.
3. Nous offrons une comparaison juste avec les MLM les plus performants dans un cadre *few-shot*, à travers les langues, les modèles et les structures de *prompts*, lorsqu'ils sont appliqués à la REN clinique.
4. Nous menons des expériences facilement reproductibles, en utilisant des méthodes faciles à mettre en œuvre, exclusivement sur des ensembles de données et des modèles de langues publiquement accessibles.

2 Etat de l'art

Reconnaissance d'entité nommées en *few-shot* avec les MLM pré-entraînés La méthode classique pour utiliser les MLM dans la REN est de les utiliser comme encodeurs. Habituellement, une couche d'étiquetage REN est entraînée à partir de zéro pour projeter l'encodage du texte dans l'éti-

quetage REN de ses tokens (Devlin *et al.*, 2019). D'autres approches adaptent les MLM au contexte *few-shot*. L'apprentissage de métrique (Fritzler *et al.*, 2019; Yang & Katiyar, 2020; Huang *et al.*, 2021a) propose d'entraîner les systèmes à apprendre une métrique sur l'espace de sortie, permettant de classer de nouvelles instances en fonction de leur distance par rapport aux instances étiquetées. L'encodage des types d'entités (Aly *et al.*, 2021; Ma *et al.*, 2022a; Hou *et al.*, 2020) exploite les noms ou descriptions des types d'entités pour mieux les étiqueter.

Reconnaissance d'entité nommées en *few-shot* avec les CLM pré-entraînés Récemment, la construction de *prompts* a suscité l'intérêt de la communauté (Brown *et al.*, 2020; Liu *et al.*, 2023). Les travaux connexes se sont concentrés sur l'étude de la formulation du *prompt* (Wei *et al.*, 2022; Ashok & Lipton, 2023; Vilar *et al.*, 2023; Wang *et al.*, 2023b), également connue sous le nom de « *prompt engineering* », d'autres travaux ont proposé une optimisation continue du *prompt* (Ma *et al.*, 2022b; Layegh *et al.*, 2023; Hu *et al.*, 2023a).

Il n'existe pas de méthode standard, largement adoptée, pour construire les *prompts* de REN (Liu *et al.*, 2023). Trois familles de *prompts* principales émergent : Le ***prompting* contraint** tente de mieux formuler la tâche REN en contraignant la génération à remplir des patrons spécifiques créés à la main, généralement adaptés aux MLMs (Cui *et al.*, 2021; Shen *et al.*, 2023; Ye *et al.*, 2023; Schick & Schütze, 2021). Les ***prompts* de listage** consistent simplement à faire prédire au modèle de langues les entités sous forme de liste (Ashok & Lipton, 2023). Les ***prompts* de balisage** ont été étudiés plus récemment par (Wang *et al.*, 2023b). Ils font en sorte que le modèle de langues entoure les mentions d'entités avec des balises spéciales.

Reconnaissance d'entité nommées clinique en *few-shot* Peu d'études se sont concentrées sur la REN en *few-shot* clinique basée sur les CLM. Dans (Hu *et al.*, 2023b), GPT-3 et ChatGPT sont évalués sur la tâche i2b2/VA 2010 (Uzuner *et al.*, 2011) dans un contexte *few-shot*. Dans (Jimenez Gutierrez *et al.*, 2022), GPT-3 est évalué sur un ensemble de tâches d'extraction d'informations biomédicales, y compris le NCBI-Disease (Doğan *et al.*, 2014). Une autre approche intéressante consiste à affiner en partie (Liao *et al.*, 2023) un CLM de domaine général sur des textes cliniques (Han *et al.*, 2023; Toma *et al.*, 2023), et à *prompter* le CLM qui en résulte. Les MLMs ont également été explorés pour la REN en *few-shot* dans le domaine biomédical (Ge *et al.*, 2023). L'apprentissage de métrique (Yang & Katiyar, 2020) et l'encodage des types (Aly *et al.*, 2021; Ma *et al.*, 2022a) ont été étudiés, ainsi que d'autres approches telles que l'apprentissage actif (Kormilitzin *et al.*, 2021), le pré-apprentissage supervisé (Huang *et al.*, 2021b) et le *in-context learning* basé sur les MLMs (Lee *et al.*, 2022).

3 Expérimentation

Corpus utilisés Afin d'évaluer les modèles, nous utilisons 14 corpus annotés en entités nommées, accessibles publiquement. Pour chaque langue, nous avons choisi deux corpus hors domaine et deux ou trois corpus dans le domaine clinique, visant des ressources comparables (même genre, mêmes types, mêmes schémas d'annotation) entre les langues, dans la mesure du possible. Nous utilisons les sous-ensembles officiels d'entraînement, de validation et de test de chaque corpus, quand ceux-ci sont disponibles.

WikiNER (Nothman *et al.*, 2013) est un corpus multilingue annoté en entités nommées, extrait de

Wikipédia fin 2010 dans neuf langues, annotant automatiquement les hyperliens vers personnes, lieux ou organisations. Nous utilisons les versions anglaise, française et espagnole.

CoNLL-2002 (Tjong Kim Sang, 2002) et **CoNLL-2003** (Tjong Kim Sang & De Meulder, 2003) sont des corpus multilingues annotés manuellement en entités nommées de types personnes, lieux et organisations, publiés pour les tâches partagées CoNLL. Nous utilisons les données espagnoles de 2002 (une collection d'articles tirés de la presse espagnole) et les données anglaises de 2003 (articles de presse de Reuters).

Quaero French Press (Grouin *et al.*, 2011) est un corpus annoté manuellement d'émissions radiophoniques francophones, avec des annotations pour 5 types d'entités : les personnes, lieux, organisations, fonctions et installations.

E3C (Magnini *et al.*, 2021) est un corpus multilingue européen de textes cliniques collectés à partir de multiples sources telles que PubMed [1] et SciELO [2]. Nous utilisons les versions anglaise, française et espagnole de ce corpus, annotées sémantiquement en types d'entités, acteurs, parties du corps, événements, RMLs (mesures et résultats de tests) et entités cliniques.

La tâche partagée **n2c2-2019** (Luo *et al.*, 2020) se concentre sur la normalisation des concepts médicaux à partir du corpus MCN (Luo *et al.*, 2019), composé de compte-rendus d'hospitalisation tirés de deux établissements hospitaliers dans le Massachusetts. Nous utilisons les identifiants uniques de concept (CUI) pour associer les mentions aux groupes sémantiques UMLS. (Lindberg *et al.*, 1993; McCray *et al.*, 2001).

Le corpus **NCBI-Disease** (Doğan *et al.*, 2014) corpus rassemble des résumés PubMed où les mentions de maladies sont annotées en quatre types selon leur syntaxe : maladies spécifiques, classes de maladies, mentions composites et modificateurs.

QuaeroFrenchMed (Névéol *et al.*, 2014) se compose de deux parties : **EMEA**, une collection de notices patient concernant des médicaments commercialisés en Europe, et **MEDLINE**, de titres d'articles scientifiques indexés dans MEDLINE. Ces deux parties sont annotées en 10 types d'entités nommées, correspondant aux groupes sémantiques UMLS.

Le corpus **The Chilean Waiting List** (Báez *et al.*, 2020) contient ordonnances anonymisées pour des consultations à partir de la liste d'attente dans les hôpitaux publics chiliens, annotées manuellement avec 10 types d'entités : abréviations, parties du corps, résultats cliniques, procédure de diagnostic, maladies, membres de la famille, résultats de laboratoire ou de test, procédures de laboratoire, médicaments, procédures, signes ou symptômes et procédures thérapeutiques. Notons que ces types peuvent être redondants (par exemple, toutes les procédures de diagnostic sont également annotées en tant que procédures).

Configuration de l'apprentissage en *few-shot* Pour simuler le contexte de *few-shot*, nous fournissons aux modèles seulement quelques exemples annotés, représentant l'ensemble des exemples autorisés pour l'apprentissage, le *prompting* et la validation. Dans cette étude, nous choisissons de nous concentrer principalement sur $k = 100$ phrases, ce qui correspond à une à deux heures d'annotation dans le domaine clinique (Névéol *et al.*, 2014; Campillos *et al.*, 2018). Nous utilisons une graine aléatoire fixe p pour choisir k exemples parmi tous ceux disponibles dans le corpus réel. Dans la section 5.2, nous discutons de l'effet du choix de k et du choix de p.

1. http://pubmed.ncbi.nlm.nih.gov/
2. https://scielo.org/

En outre, nous testons les modèles les plus performants avec l'entièreté des annotations à disposition pour une comparaison à la *skyline*.

Modèles de langues Nous évaluons 10 modèles causaux et 16 modèles masqués sur les tâches de REN précisées ci-dessus. Ces modèles sont listés dans le tableau de résultats 1 et décrits en plus de détail dans l'annexe 1. Alors que le français et l'espagnol sont couverts par de certains des modèles causaux, nous pouvons observer que l'anglais est omniprésent. La plupart des modèles de type BERT sont monolingues, à l'exception des modèles multilingues mBERT et XLM-RoBERTa.

REN avec des modèles de langues masqués Comme mentionné dans la section 2, les modèles de langues masqués ont été adaptés à l'apprentissage en *few-shot* dans des architectures adaptées aux contextes *few-shot*. Cependant, dans ce travail, nous souhaitons comparer la nouvelle approche CLM à l'utilisation standard et plus répandue des MLM sans adaptation pour le contexte *few-shot*. Nous utilisons NLStruct (Wajsbürt, 2021), une bibliothèque Python open-source [3] qui met en œuvre l'approche standard du *fine-tuning*. NLStruct utilise les représentations fournies par le modèle de langues pour encoder l'entrée, puis utilise un décodeur LSTM bidirectionnel et un CRF pour prédire itérativement les entités présentes dans l'entrée encodée, comme décrit par Gérardin *et al.* (2022). Cette démarche permet à NLStruct de traiter efficacement les entités imbriquées, très présentes dans certains des corpus d'étude. Nous entraînons le modèle pendant 20 epochs sur 80 % des données et utilisons les 20 % restants pour valider l'*early stopping*.

REN avec des modèles de langues causaux Dans nos expériences, nous invitons les modèles à baliser les mentions, et non de les lister. Nous discutons de ce choix plus en détail dans la section 5.2. La partie supérieure de la figure 1 montre un exemple de *prompt* de balisage, en mettant en évidence les différentes sections de celui-ci. Ci-dessous, nous décrivons 9 caractéristiques de formulation du *prompt* et de sélection des exemples qui y figurent.

1. **Langue du *prompt*** : Par défaut, nous construisons les *prompts* en anglais, car il s'agit de la langue la plus répandue dans tous les corpus d'apprentissage. Cette caractéristique consiste à faire plutôt aligner la langue du *prompt* sur celle de la phrase de test.
2. **Phases supplémentaires** : Par défaut, nous présentons 5 phrases annotées dans les annotées. Cette caractéristique permet de présenter 5 phrases supplémentaires (soit 10 phrases au total). La partie 5.2 discute de ce choix, ainsi que de la possibilité de présenter plus de démonstrations dans le *prompt*.
3. **Auto-vérification** : Par défaut, nous sélectionnons les 5 (ou 10) phrases les plus proches de la phrase test en termes de distance TF-IDF. Les mentions étiquetées par le modèle sont alors considérées comme les prédictions finales du modèle. Cette fonctionnalité sélectionne plutôt les 5 phrases contenant le plus d'entités du type ciblé et les présente dans un *prompt* initial. Intuitivement, ce *prompt* se traduit par un rappel plus élevé et une précision plus faible. Un deuxième *prompt* d'« auto-vérification » est ensuite construit sur les prédictions initiales du modèle afin d'éliminer les faux positifs. Un exemple de *prompt* d'auto-vérification est présenté dans la partie inférieure de la figure 1. Le nombre de démonstrations suit celui du *prompt* principal.

3. https://github.com/percevalw/nlstruct

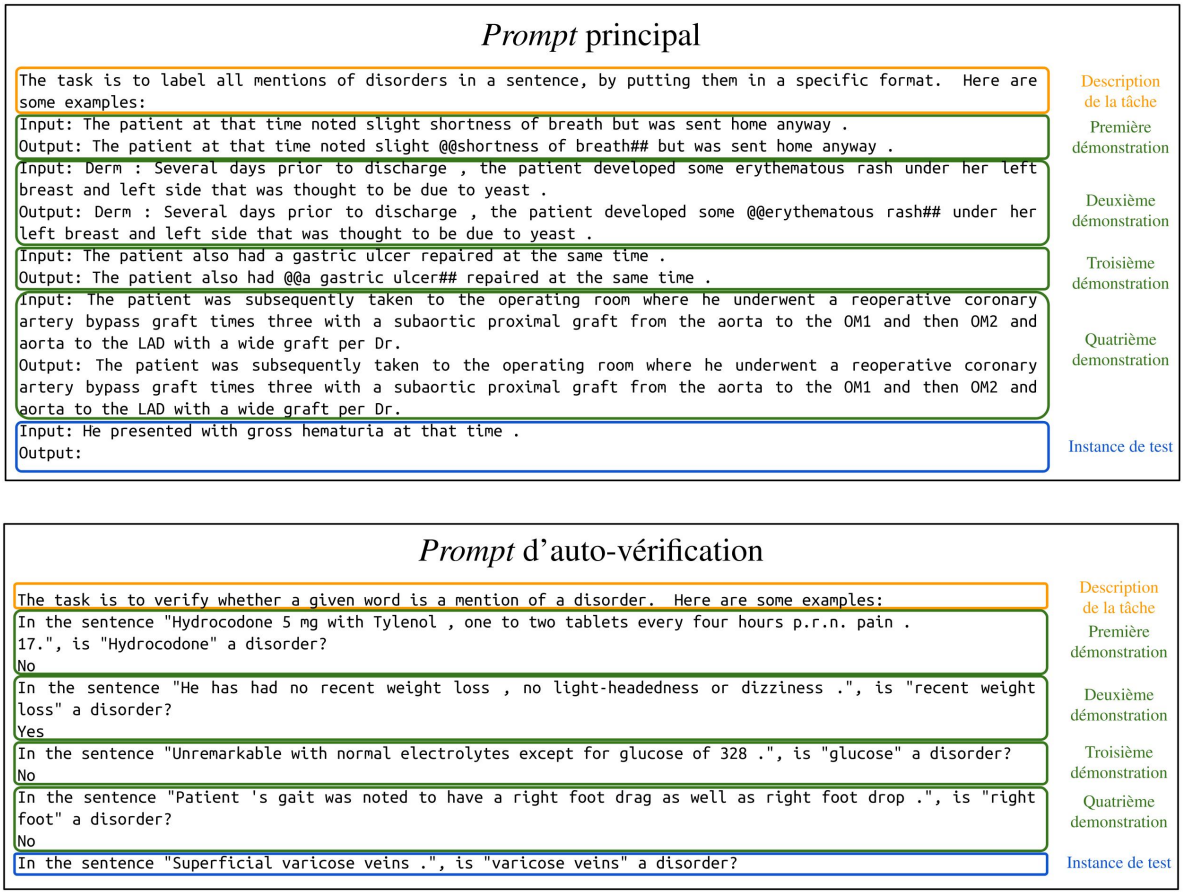


FIGURE 1 – Exemple d'un *prompt* de balisage, utilisée dans l'expérience principale (en haut) et d'un *prompt* d'auto-vérification (en bas) pour détecter les mentions DISO dans **n2c2-2019**

4. **Baliseurs** : Par défaut, nous suivons Wang *et al.* (2023b) qui invite le modèle à entourer les mentions de @@ et ##. Cette caractéristique l'invite plutôt à entourer les mentions de guillemets « et ».
5. **S'addresser à un spécialiste** : Par défaut, la première phrase est la description de la tâche présentée dans la figure 1. Cette caractéristique fait commencer le *prompt* par *You are an excellent <specialist>. You can identify all the mentions of <entity-type> in a sentence, by putting them in a specific format. Here are some examples you can handle :* à la place. Le *<specialist>* est un *linguist* ou un *clinician*, suivant le domaine de la tâche.
6. **Inclure les descriptions des types dans le *prompt*** : Cette caractéristique ajoute une description d'une phrase pour chaque type d'entité. Les descriptions complètes des entités utilisées figurent à l'annexe 3.
7. **Phrase d'introduction pour l'instance de test** : Par défaut, les démonstrations sont immédiatement suivies de l'instance de test. Cette fonctionnalité consiste à la précéder par *Identify all the mentions of <entity-type> in the following sentence, by putting <begin-tag> in front and a <end-tag> behind each of them.*.
8. **Demander une réponse longue pour l'auto-vérification** : Par défaut, le *prompt* d'auto-vérification demander *Yes* (respectivement *No*) comme réponse. Cette fonctionnalité demande *<mention> is a(n) <entity-type>, yes.* (respectivement *<mention> is not a(n) <entity-type>, no.*).

9. **Format dialogue** : Cette fonction remplace les *Input :* et *Output :* du *prompt* par des tirets pour imiter un format de dialogue.

Les performances de l'*in-context learning* varient considérablement en fonction de la formulation exacte du *prompt* (Lu *et al.*, 2022; Min *et al.*, 2022). En outre, le choix optimal de chacune de ces caractéristiques peut varier en fonction du modèle utilisé. Par exemple, intuitivement, les modèles dont le pré-entraînement est fortement concentrés sur la langue anglaise ont tendance à être plus performants avec un *prompt* en anglais qu'avec un *prompt* dans la langue du corpus.

Notre système vise à rechercher la meilleure combinaison de caractéristique pour chaque modèle, mais un *grid search* sur ces caractéristique nécessiterait $2^9 = 512$ expériences pour chaque modèle et pour chaque corpus. Afin de construire un système plus léger, nous avons choisi d'effectuer un *greedy search*. Nous itérons sur les caractéristiques dans cet ordre, en testant la valeur qui n'est pas celle par défaut et en la conservant si elle est plus performante que la valeur par défaut. Dans la section 5.2, nous comparons cette approche à un *grid search* pour un modèle sur un corpus.

De nombreux travaux sur l'apprentissage en « *few-shot* » avec les CLM optimisent les *prompts* sur de grands jeux de données de validation (Brown *et al.*, 2020; Tam *et al.*, 2021; Radford *et al.*, 2021; Qin & Eisner, 2021). Cela conduit à des résultats qui se révèlent (Perez *et al.*, 2021) trop optimistes. Une comparaison équitable entre les MLM et les CLM devrait les comparer avec l'accès au même (petit) nombre d'instances annotées, ce qui correspond à notre $k = 100$. Dans ce contexte d'absence d'entraînement, nous suivons Perez *et al.* (2021) en optimisant ces caractéristiques par une validation LOOCV (leave-one-out cross-validation).

Mesures Nous évaluons la performance des modèles à l'aide de deux mesures. Pour des raisons de simplicité, nous évaluons les modèles sur la base d'un score de performance global, la **micro-F1**. Il est calculé comme la micro-moyenne des F1-mesures de la détection de chaque type d'entité. Nous mesurons également l'**empreinte carbone** de chacune des approches. Nous utilisons GreenAlgorithms v2.2 (Lannelongue *et al.*, 2021) [4] pour estimer l'empreinte carbone de chaque expérience, sur la base de facteurs tels que la durée d'exécution, le matériel informatique et le lieu de production de l'électricité utilisée par notre installation informatique.

4 Résultats et discussion

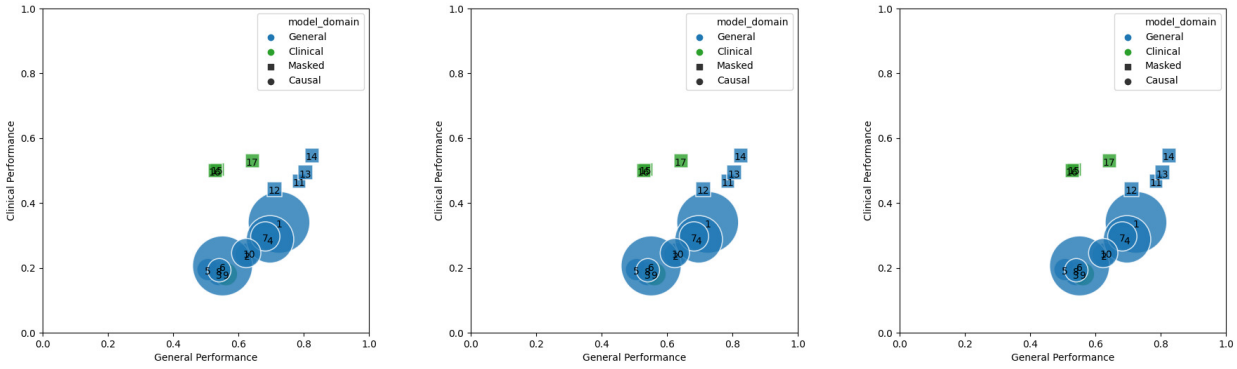
Mesures Le tableau 1 et la figure 2 décrivent les performances des modèles testés. L'annexe 4 détaille les estimations des émissions carbone pour toutes nos expériences. En particulier, nous estimons que l'expérience utilisant Mistral-7B sur CoNLL-2003 a généré 41 g d'équivalent CO2. (6 g pour l'optimisation du *prompt* et 35 g pour l'inférence sur l'ensemble de test). LLaMA-2-70B, environ 10 fois plus grand, est estimé avoir généré 191 g d'équivalent CO2. (44 g pour l'optimisation du *prompt* et 147 g pour l'inférence sur l'ensemble de test). L'expérience sur le modèle de langue masqué BERT-large est quant à elle estimée avoir généré 6 g d'équivalent CO2. (2 g pour le fine-tuning et l'entraînement et 4 g pour l'inférence sur l'ensemble de test).

Au total, on estime que les expériences décrites dans cet article ont généré environ 27 kg d'équivalent CO2 (25 kg pour les expériences principales et 2 kg pour l'ablation).

4. http ://calculator.green-algorithms.org/

	#	Modèle	Anglais					Français					Espagnol			
			WikiNER	CoNLL2003	E3C	n2c2	NCBI	WikiNER	QFP	E3C	EMEA	MEDLINE	WikiNER	CoNLL2002	E3C	CWL
Approches few-shot																
Causal	1	LLAMA-2-70B	0.728	0.721	0.312	0.309	0.400	0.740	0.400	0.483	0.201	0.312	0.805	0.616	0.021	0.339
	2	Mistral-7B	0.754	0.646	0.488	0.291	0.395	0.727	0.428	0.590	0.229	0.333	0.720	0.707	0.083	0.374
	3	BLOOM-7B1	0.524	0.557	0.279	0.113	0.151	0.148	0.206	0.320	0.197	0.120	0.470	0.419	0.051	0.117
	4	Falcon-40B	0.686	0.708	0.280	0.279	0.305	0.662	0.456	0.378	0.279	0.283	0.720	0.543	0.072	0.267
	5	GPT-J-6B	0.521	0.493	0.167	0.179	0.238	0.423	0.244	0.334	0.080	0.177	0.005	0.142	0.021	0.162
	6	OPT-66B	0.608	0.495	0.227	0.157	0.234	0.624	0.406	0.019	0.206	0.283	0.166	0.273	0.043	0.204
	7	Vicuna-13B	0.657	0.708	0.355	0.236	0.300	0.677	0.350	0.399	0.207	0.326	0.744	0.250	0.040	0.213
	8	Vicuna-7B	0.594	0.489	0.259	0.147	0.172	0.591	0.277	0.439	0.152	0.296	0.659	0.569	0.042	0.151
	9	Medalpaca-7B	0.537	0.586	0.272	0.138	0.132	0.529	0.142	0.259	0.162	0.252	0.581	0.490	0.088	0.220
	10	Vigogne-13B	0.593	0.655	0.252	0.176	0.309	0.515	0.250	0.464	0.099	0.142	0.580	0.561	0.010	0.198
Masked	11	mBERT	0.768	0.804	0.624	0.378	0.401	0.801	0.728	0.741	0.588	0.428	0.812	0.760	0.324	0.432
	12	XLM-R-large	0.786	0.826	0.637	0.462	0.471	0.811	0.781	0.762	0.629	0.531	0.797	0.781	0.325	0.528
	13	BERT-large	0.776	0.835	0.626	0.435	0.422	-	-	-	-	-	-	-	-	-
	14	RoBERTa-large	0.790	0.862	0.626	0.462	0.552	-	-	-	-	-	-	-	-	-
	15	Bio_ClinicalBERT	0.528	0.542	0.621	0.469	0.420	-	-	-	-	-	-	-	-	-
	16	ClinicalBERT	0.462	0.597	0.622	0.480	0.397	-	-	-	-	-	-	-	-	-
	17	MedBERT	0.613	0.673	0.607	0.478	0.504	-	-	-	-	-	-	-	-	-
	18	CamemBERT-large	-	-	-	-	-	0.829	0.793	0.768	0.661	0.577	-	-	-	-
	19	FlauBERT-large	-	-	-	-	-	0.826	0.778	0.760	0.635	0.542	-	-	-	-
	20	DrBERT-4GB	-	-	-	-	-	0.587	0.599	0.730	0.602	0.486	-	-	-	-
	21	CamemBERT-bio	-	-	-	-	-	0.782	0.761	0.779	0.636	0.549	-	-	-	-
	22	BETO	-	-	-	-	-	-	-	-	-	-	0.794	0.732	0.352	0.522
	23	PatanaBERT	-	-	-	-	-	-	-	-	-	-	0.802	0.769	0.343	0.487
	24	TulioBERT	-	-	-	-	-	-	-	-	-	-	0.804	0.798	0.340	0.482
	25	BSC-BioEHR	-	-	-	-	-	-	-	-	-	-	0.804	0.758	0.354	0.578
	26	BSC-Bio	-	-	-	-	-	-	-	-	-	-	0.804	0.775	0.358	0.552
Skyline en utilisant toutes les données à disposition																
		RoBERTa-large	0.919	0.939	0.718	0.712	0.815	-	-	-	-	-	-	-	-	-
		CamemBERT-large	-	-	-	-	-	0.928	0.834	0.828	0.748	0.713	-	-	-	-
		BETO	-	-	-	-	-	-	-	-	-	-	0.918	0.881	0.411	0.736

TABLE 1 – Mesures micro-F1 obtenues. Nous évaluons les modèles masqués monolingues uniquement dans les langues sur lesquelles ils ont été entraînés.



(a) Anglais (b) Français (c) Espagnol

FIGURE 2 – Performance dans les domaines général vs. clinique des modèles étudiés

Comparaison des performances des modèles Nous avons comparé différents modèles de langues masqués (MLM) et causaux (CLM) pour la reconnaissance d'entités nommées en nous concentrant sur les contextes à faibles ressources, typiques des applications biomédicales. Les résultats montrent que les modèles masqués de type BERT, bien que plus petits et nécessitant théoriquement une plus grande quantité de données d'apprentissage, surpassent systématiquement les CLM. Cette performance s'accompagne d'un impact environnemental beaucoup plus faible (les émissions de CO2 sont 10 à 50 fois inférieures pour les MLM par rapport aux CLM), et d'une plus grande consistance (par exemple, sur la tâche généraliste WikiNER en anglais, les 4 modèles du domaine général testés ont

obtenu des scores F1 compris entre 0,768 et 0,79). Par ailleurs, les MLMs spécialisés dans le domaine biomédical apportent peu d'amélioration significative dans les tâches spécialisées. Ce commentaire doit cependant être pondéré par la différence de taille entre les modèles : tous les modèles spécialisés ont seulement 110 millions de paramètres.

La reconnaissance d'entités nommées basée sur des représentations de type BERT a reçu beaucoup d'attention ces dernières années, et est sans aucun doute plus mature que l'utilisation des CLMs pour cette tâche. Nous avons exploré les techniques de reconnaissance d'entités nommées basées sur les CLMs existantes dans la littérature, avec nos connaissances actuelles. De nouvelles approches pourraient améliorer les performances à l'avenir, mais cette tâche reste difficile pour un modèle génératif en raison de ses contraintes syntaxiques et d'évaluation spécifiques. Ces résultats ne reflètent pas nécessairement les performances sur d'autres tâches, comme la classification.

Usage pratique des modèles de langues pour la REN à peu de ressources Nos expériences indiquent que les modèles de langues pour la reconnaissance des entités nommées cliniques ont actuellement des performances sous-optimales. Même les modèles MLM, *fine-tunés* simplement avec le peu de données à disposition, ne rivalisent pas avec les modèles entièrement supervisés. Les grands modèles entraînés avec l'ensemble de chaque corpus d'entraînement surpassent systématiquement les meilleurs résultats en *few-shot*, de 5 à 16 % pour le domaine général et de 8 à 48 % pour le domaine biomédical (*skylines* Table 1). Cependant, les performances peuvent suffire pour une utilisation en pré-annotation, accélérant ainsi l'annotation manuelle, par exemple dans un contexte d'*online learning* ou d'*active learning*.

Bruit aléatoire Dans les expériences MLM, les paramètres de la couche d'étiquetage REN ajoutée au modèle pré-entraîné sont initialisés de manière aléatoire. De même, dans les expériences CLM, les démonstrations dans les *prompts* sont ordonnées de façon aléatoire et les exemples négatifs dans les *prompts* d'auto-vérification sont sélectionnés de manière aléatoire, introduisant potentiellement du bruit dans nos mesures de performance. Répliquer toutes les expériences renforcerait nos conclusions (Reimers & Gurevych, 2017), mais il serait coûteux (25kg de CO2eq et 56 heures de calcul par réplication). Le grand nombre de modèles testés et de tâches traitées peut toutefois conforter les principales observations de cet article. Par exemple, nous utilisons l'ordre presque stochastique (ASO)[5] (Dror *et al.*, 2019) avec $\alpha = 0,05$ pour mesurer la significativité de la supériorité des MLM sur les CLM pour chaque ensemble de données séparément. Les MLM ne montrent pas toujours une supériorité significative sur les CLM pour la REN dans le domaine général (0,54 et 0,121 respectivement pour WikiNER anglais et CoNLL2003). Pour la REN clinique, les MLM sont nettement supérieurs aux CLM : les MLM dominent stochastiquement les CLM (ϵ_{min}=0) pour tous les corpus cliniques.

Conclusion Cette étude a évalué les performances de deux types de modèles de langues pour la reconnaissance d'entités en *few-shot* dans trois langues. Nos expériences révèlent que la performance du *few-shot learning* est significativement plus faible dans le domaine clinique que dans le domaine

5. Étant donné les scores de performance de deux algorithmes A et B, chacun étant exécuté plusieurs fois avec des paramètres différents, l'ASO calcule une valeur spécifique au test (ϵ_{min}) qui indique à quel point l'algorithme A est loin d'être significativement meilleur que l'algorithme B. Si la distance ϵ_{min} = 0,0, on peut affirmer que A domine stochastiquement B avec le niveau de signification prédéfini. La littérature interprète généralement ϵ_{min} < 0,5 comme un indicateur de supériorité significative de A sur B.

général. Alors que les modèles de langues masqués surpassent les modèles de langues causaux (avec une F1-mesure plus élevée et des émissions de CO2 plus faibles), leur utilisation devrait être restreinte à la pré-annotation plutôt qu'à l'extraction d'informations efficace.

Références

ALSENTZER E., MURPHY J., BOAG W., WENG W.-H., JINDI D., NAUMANN T. & MCDERMOTT M. (2019). Publicly available clinical BERT embeddings. In A. RUMSHISKY, K. ROBERTS, S. BETHARD & T. NAUMANN, Éds., *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, p. 72–78, Minneapolis, Minnesota, USA : Association for Computational Linguistics. DOI : 10.18653/v1/W19-1909.

ALY R., VLACHOS A. & MCDONALD R. (2021). Leveraging type descriptions for zero-shot named entity recognition and classification. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1516–1528, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.120.

ASHOK D. & LIPTON Z. (2023). Promptner : Prompting for named entity recognition.

BÁEZ P., VILLENA F., ROJAS M., DURÁN M. & DUNSTAN J. (2020). The Chilean waiting list corpus : a new resource for clinical named entity recognition in Spanish. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, p. 291–300, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.clinicalnlp-1.32.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.

CAMPILLOS L., DELÉGER L., GROUIN C., HAMON T., LIGOZAT A.-L. & NÉVÉOL A. (2018). A french clinical corpus with comprehensive semantic annotations : development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*, **52**, 571–601.

CARRINO C. P., LLOP J., PÀMIES M., GUTIÉRREZ-FANDIÑO A., ARMENGOL-ESTAPÉ J., SILVEIRA-OCAMPO J., VALENCIA A., GONZALEZ-AGIRRE A. & VILLEGAS M. (2022). Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, p. 193–199, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.bionlp-1.19.

CAÑETE J., CHAPERON G., FUENTES R., HO J.-H., KANG H. & PÉREZ J. (2020). Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

CHO H., CHOI W. & LEE H. (2017). A method for named entity normalization in biomedical articles : Application to diseases and plants. *BMC Bioinformatics*, **18**. DOI : 10.1186/s12859-017-1857-8.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised cross-lingual representation learning at scale. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds.,

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, p. 8440–8451, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

CUI L., WU Y., LIU J., YANG S. & ZHANG Y. (2021). Template-based named entity recognition using BART. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 1835–1845, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.161.

DEMNER-FUSHMAN D., CHAPMAN W. W. & MCDONALD C. J. (2009). What can natural language processing do for clinical decision support ? *Journal of Biomedical Informatics*, **42**(5), 760–772. Biomedical Natural Language Processing, DOI : https ://doi.org/10.1016/j.jbi.2009.08.007.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

DOĞAN R. I., LEAMAN R. & LU Z. (2014). Ncbi disease corpus : A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, **47**, 1–10. DOI : https ://doi.org/10.1016/j.jbi.2013.12.006.

DROR R., SHLOMOV S. & REICHART R. (2019). Deep dominance - how to properly compare deep neural models. In A. KORHONEN, D. R. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1 : Long Papers*, p. 2773–2785 : Association for Computational Linguistics. DOI : 10.18653/v1/p19-1266.

DU S. S., HU W., KAKADE S. M., LEE J. D. & LEI Q. (2021). Few-shot learning via learning the representation, provably.

ESCUDIÉ J.-B., RANCE B., MALAMUT G., KHATER S., BURGUN A., CELLIER C. & JANNOT A.-S. (2017). A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease : a case study on autoimmune comorbidities in patients with celiac disease. *BMC medical informatics and decision making*, **17**(1), 1–10.

FRITZLER A., LOGACHEVA V. & KRETOV M. (2019). Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, p. 993–1000, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3297280.3297378.

GAO L., BIDERMAN S., BLACK S., GOLDING L., HOPPE T., FOSTER C., PHANG J., HE H., THITE A., NABESHIMA N. *et al.* (2020). The pile : An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv :2101.00027*.

GE Y., GUO Y., DAS S., AL-GARADI M. A. & SARKER A. (2023). Few-shot learning for medical text : A review of advances, trends, and opportunities. *Journal of Biomedical Informatics*, **144**, 104458. DOI : https ://doi.org/10.1016/j.jbi.2023.104458.

GROUIN C., ROSSET S., ZWEIGENBAUM P., FORT K., GALIBERT O. & QUINTARD L. (2011). Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview. In N. IDE, A. MEYERS, S. PRADHAN & K. TOMANEK, Éds., *Proceedings of the 5th Linguistic Annotation Workshop*, p. 92–100, Portland, Oregon, USA : Association for Computational Linguistics.

GUPTA S., GARDNER M. & SINGH S. (2023). Coverage-based example selection for in-context learning. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational*

Linguistics : EMNLP 2023, p. 13924–13950, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.930.

GÉRARDIN C., WAJSBÜRT P., VAILLANT P., BELLAMINE A., CARRAT F. & TANNIER X. (2022). Multilabel classification of medical concepts for patient clinical profile identification. *Artificial Intelligence in Medicine*, **128**, 102311. DOI : https ://doi.org/10.1016/j.artmed.2022.102311.

HAN T., ADAMS L. C., PAPAIOANNOU J.-M., GRUNDMANN P., OBERHAUSER T., LÖSER A., TRUHN D. & BRESSEM K. K. (2023). Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv :2304.08247*.

HOU Y., CHE W., LAI Y., ZHOU Z., LIU Y., LIU H. & LIU T. (2020). Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 1381–1393, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.128.

HU N., ZHOU X., XU B., LIU H., XIE X. & ZHENG H.-T. (2023a). Vpn : Variation on prompt tuning for named-entity recognition. *Applied Sciences*, **13**(14). DOI : 10.3390/app13148359.

HU Y., AMEER I., ZUO X., PENG X., ZHOU Y., LI Z., LI Y., LI J., JIANG X. & XU H. (2023b). Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv :2303.16416*.

HUANG J., LI C., SUBUDHI K., JOSE D., BALAKRISHNAN S., CHEN W., PENG B., GAO J. & HAN J. (2021a). Few-shot named entity recognition : An empirical baseline study. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 10408–10423, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.813.

HUANG J., LI C., SUBUDHI K., JOSE D., BALAKRISHNAN S., CHEN W., PENG B., GAO J. & HAN J. (2021b). Few-shot named entity recognition : An empirical baseline study. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 10408–10423, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.813.

JIA C., LIANG X. & ZHANG Y. (2019). Cross-domain NER using cross-domain language modeling. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 2464–2474, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1236.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., CASAS D. D. L., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L. *et al.* (2023). Mistral 7b. *arXiv preprint arXiv :2310.06825*.

JIMENEZ GUTIERREZ B., MCNEAL N., WASHINGTON C., CHEN Y., LI L., SUN H. & SU Y. (2022). Thinking about GPT-3 in-context learning for biomedical IE ? think again. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2022*, p. 4497–4512, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-emnlp.329.

JOHNSON A. E., POLLARD T. J., SHEN L., LEHMAN L.-W. H., FENG M., GHASSEMI M., MOODY B., SZOLOVITS P., ANTHONY CELI L. & MARK R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, **3**(1), 1–9.

KORMILITZIN A., VACI N., LIU Q. & NEVADO-HOLGADO A. (2021). Med7 : A transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, **118**, 102086. DOI : https ://doi.org/10.1016/j.artmed.2021.102086.

LABRAK Y., BAZOGE A., DUFOUR R., ROUVIER M., MORIN E., DAILLE B. & GOURRAUD P.-A. (2023). DrBERT : A robust pre-trained model in French for biomedical and clinical domains. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 16207–16221, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.896.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LANNELONGUE L., GREALEY J. & INOUYE M. (2021). Green algorithms : quantifying the carbon footprint of computation. *Advanced science*, **8**(12), 2100707.

LAURENÇON H., SAULNIER L., WANG T., AKIKI C., VILLANOVA DEL MORAL A., LE SCAO T., VON WERRA L., MOU C., GONZÁLEZ PONFERRADA E., NGUYEN H. *et al.* (2022). The bigscience roots corpus : A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, **35**, 31809–31826.

LAYEGH A., PAYBERAH A. H., SOYLU A., ROMAN D. & MATSKIN M. (2023). Contrastner : Contrastive-based prompt tuning for few-shot ner. *arXiv preprint arXiv :2305.17951*.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised language model pre-training for French. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 2479–2490, Marseille, France : European Language Resources Association.

LEAMAN R., KHARE R. & LU Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, **57**, 28–37. DOI : https ://doi.org/10.1016/j.jbi.2015.07.010.

LEE D.-H., KADAKIA A., TAN K., AGARWAL M., FENG X., SHIBUYA T., MITANI R., SEKIYA T., PUJARA J. & REN X. (2022). Good examples make a faster learner : Simple demonstration-based learning for low-resource NER. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2687–2700, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.192.

LI J., SUN A., HAN J. & LI C. (2022). A survey on deep learning for named entity recognition. *IEEE Trans. on Knowl. and Data Eng.*, **34**(1), 50–70. DOI : 10.1109/TKDE.2020.2981314.

LIAO B., MENG Y. & MONZ C. (2023). Parameter-efficient fine-tuning without introducing new latency. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 4242–4260, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.233.

LINDBERG D. A., HUMPHREYS B. L. & MCCRAY A. T. (1993). The unified medical language system. *Yearbook of medical informatics*, **2**(01), 41–51.

LIU P., YUAN W., FU J., JIANG Z., HAYASHI H. & NEUBIG G. (2023). Pre-train, prompt, and predict : A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, **55**(9), 1–35.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

LIU Z., XU Y., YU T., DAI W., JI Z., CAHYAWIJAYA S., MADOTTO A. & FUNG P. (2021). Crossner : Evaluating cross-domain named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**(15), 13452–13460. DOI : 10.1609/aaai.v35i15.17587.

LU Y., BARTOLO M., MOORE A., RIEDEL S. & STENETORP P. (2022). Fantastically ordered prompts and where to find them : Overcoming few-shot prompt order sensitivity. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 8086–8098, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.556.

LUO Y.-F., HENRY S., WANG Y., SHEN F., UZUNER O. & RUMSHISKY A. (2020). The 2019 n2c2/UMass Lowell shared task on clinical concept normalization. *Journal of the American Medical Informatics Association*, **27**(10), 1529–e1. DOI : 10.1093/jamia/ocaa106.

LUO Y.-F., SUN W. & RUMSHISKY A. (2019). Mcn : A comprehensive corpus for medical concept normalization. *Journal of Biomedical Informatics*, **92**, 103132. DOI : https ://doi.org/10.1016/j.jbi.2019.103132.

MA J., BALLESTEROS M., DOSS S., ANUBHAI R., MALLYA S., AL-ONAIZAN Y. & ROTH D. (2022a). Label semantics for few shot named entity recognition. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Findings of the Association for Computational Linguistics : ACL 2022*, p. 1956–1971, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-acl.155.

MA R., ZHOU X., GUI T., TAN Y., LI L., ZHANG Q. & HUANG X. (2022b). Template-free prompt tuning for few-shot NER. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 5721–5732, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.420.

MAGNINI B., ALTUNA B., LAVELLI A., SPERANZA M. & ZANOLI R. (2021). The e3c project : European clinical case corpus. *Language*, **1**(L2), L3.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MCCRAY A., BURGUN A. & BODENREIDER O. (2001). Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, **84**, 216–20. DOI : 10.3233/978-1-60750-928-8-216.

MIN S., LYU X., HOLTZMAN A., ARTETXE M., LEWIS M., HAJISHIRZI H. & ZETTLEMOYER L. (2022). Rethinking the role of demonstrations : What makes in-context learning work ? In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 11048–11064, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.759.

NÉVÉOL A., GROUIN C., LEIXA J., ROSSET S. & ZWEIGENBAUM P. (2014). The quaero french medical corpus : A ressource for medical entity recognition and normalization. *Proc of BioTextMining Work*, p. 24–30.

NOTHMAN J., RINGLAND N., RADFORD W., MURPHY T. & CURRAN J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, **194**, 151–175. Artificial Intelligence, Wikipedia and Semi-Structured Resources, DOI : https ://doi.org/10.1016/j.artint.2012.03.006.

PENEDO G., MALARTIC Q., HESSLOW D., COJOCARU R., CAPPELLI A., ALOBEIDLI H., PANNIER B., ALMAZROUEI E. & LAUNAY J. (2023). The refinedweb dataset for falcon llm : outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv :2306.01116*.

PEREZ E., KIELA D. & CHO K. (2021). True few-shot learning with language models. In A. BEYGELZIMER, Y. DAUPHIN, P. LIANG & J. W. VAUGHAN, Éds., *Advances in Neural Information Processing Systems*.

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In M. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 2227–2237, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1202.

QIN G. & EISNER J. (2021). Learning how to ask : Querying LMs with mixtures of soft prompts. In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 5203–5212, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.410.

RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., KRUEGER G. & SUTSKEVER I. (2021). Learning transferable visual models from natural language supervision. In M. MEILA & T. ZHANG, Éds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 de *Proceedings of Machine Learning Research*, p. 8748–8763 : PMLR.

REIMERS N. & GUREVYCH I. (2017). Reporting score distributions makes a difference : Performance study of LSTM-networks for sequence tagging. In M. PALMER, R. HWA & S. RIEDEL, Éds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 338–348, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1035.

SCHICK T. & SCHÜTZE H. (2021). It's not just size that matters : Small language models are also few-shot learners. In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 2339–2352, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.185.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SHEN Y., TAN Z., WU S., ZHANG W., ZHANG R., XI Y., LU W. & ZHUANG Y. (2023). PromptNER : Prompt locating and typing for named entity recognition. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 12492–12507, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.698.

SHIN S., LEE S.-W., AHN H., KIM S., KIM H., KIM B., CHO K., LEE G., PARK W., HA J.-W. & SUNG N. (2022). On the effect of pretraining corpora on in-context learning by a large-scale

language model. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 5168–5186, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.380.

SRIVASTAVA A., RASTOGI A., RAO A. & CO AUTHORS (2023). Beyond the imitation game : Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

SUÁREZ P. J. O., ROMARY L. & SAGOT B. (2020). A monolingual approach to contextualized word embeddings for mid-resource languages. *arXiv preprint arXiv :2006.06202*.

SUN C., YANG Z., WANG L., ZHANG Y., LIN H. & WANG J. (2021). Biomedical named entity recognition using bert in the machine reading comprehension framework. *Journal of Biomedical Informatics*, **118**, 103799. DOI : https ://doi.org/10.1016/j.jbi.2021.103799.

SUNG M., JEONG M., CHOI Y., KIM D., LEE J. & KANG J. (2022). BERN2 : an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics*, **38**(20), 4837–4839. DOI : 10.1093/bioinformatics/btac598.

TAM D., R. MENON R., BANSAL M., SRIVASTAVA S. & RAFFEL C. (2021). Improving and simplifying pattern exploiting training. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 4980–4991, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.407.

TIEDEMANN J. (2012). Parallel data, tools and interfaces in OPUS. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, M. U. DOĞAN, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 2214–2218, Istanbul, Turkey : European Language Resources Association (ELRA).

TJONG KIM SANG E. F. (2002). Introduction to the CoNLL-2002 shared task : Language-independent named entity recognition. In *COLING-02 : The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

TJONG KIM SANG E. F. & DE MEULDER F. (2003). Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, p. 142–147.

TOMA A., LAWLER P. R., BA J., KRISHNAN R. G., RUBIN B. B. & WANG B. (2023). Clinical camel : An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv :2305.12031*.

TOUCHENT R., ROMARY L. & DE LA CLERGERIE E. (2023). CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé. In C. SERVAN & A. VILNAT, Éds., *18e Conférence en Recherche d'Information et Applications*
16e Rencontres Jeunes Chercheurs en RI
30e Conférence sur le Traitement Automatique des Langues Naturelles
25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, p. 323–334, Paris, France : ATALA. HAL : hal-04130187.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S. *et al.* (2023). Llama 2 : Open foundation and fine-tuned chat models. *arXiv preprint arXiv :2307.09288*.

UZUNER Ö., SOUTH B. R., SHEN S. & DUVALL S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, **18**(5), 552–556.

VASANTHARAJAN C., TUN K. Z., THI-NGA H., JAIN S., RONG T. & SIONG C. E. (2022). Medbert : A pre-trained language model for biomedical named entity recognition. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, p. 1482–1488. DOI : 10.23919/APSIPAASC55919.2022.9980157.

VILAR D., FREITAG M., CHERRY C., LUO J., RATNAKAR V. & FOSTER G. (2023). Prompting PaLM for translation : Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 15406–15427, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.859.

WAJSBÜRT P. (2021). *Extraction and normalization of simple and structured entities in medical documents*. Theses, Sorbonne Université. HAL : tel-03624928.

WAJSBÜRT P., SARFATI A. & TANNIER X. (2021). Medical concept normalization in french using multilingual terminologies and contextual embeddings. *Journal of Biomedical Informatics*, **114**, 103684. DOI : https ://doi.org/10.1016/j.jbi.2021.103684.

WANG B. & KOMATSUZAKI A. (2021). GPT-J-6B : A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

WANG G., LIU X., YING Z., YANG G., CHEN Z., LIU Z., ZHANG M., YAN H., LU Y., GAO Y. *et al.* (2023a). Optimized glycemic control of type 2 diabetes with reinforcement learning : a proof-of-concept trial. *Nature Medicine*, p. 1–10.

WANG S., SUN X., LI X., OUYANG R., WU F., ZHANG T., LI J. & WANG G. (2023b). Gpt-ner : Named entity recognition via large language models.

WANG Y., TONG H., ZHU Z. & LI Y. (2022). Nested named entity recognition : A survey. *ACM Trans. Knowl. Discov. Data*, **16**(6). DOI : 10.1145/3522593.

WANG Y., WANG L., RASTEGAR-MOJARAD M., MOON S., SHEN F., AFZAL N., LIU S., ZENG Y., MEHRABI S., SOHN S. & LIU H. (2018). Clinical information extraction applications : A literature review. *Journal of Biomedical Informatics*, **77**, 34–49. DOI : https ://doi.org/10.1016/j.jbi.2017.11.011.

WEI J., WANG X., SCHUURMANS D., BOSMA M., XIA F., CHI E., LE Q. V., ZHOU D. *et al.* (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, **35**, 24824–24837.

WORKSHOP B., SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIĆ S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F. *et al.* (2022). Bloom : A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv :2211.05100*.

YANG Y. & KATIYAR A. (2020). Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6365–6375, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.516.

YE F., HUANG L., LIANG S. & CHI K. (2023). Decomposed two-stage prompt learning for few-shot named entity recognition. *Information*, **14**(5). DOI : 10.3390/info14050262.

ZHANG S., ROLLER S., GOYAL N., ARTETXE M., CHEN M., CHEN S., DEWAN C., DIAB M., LI X., LIN X. V. *et al.* (2022). Opt : Open pre-trained transformer language models. *arXiv preprint arXiv :2205.01068*.

ZHENG L., CHIANG W.-L., SHENG Y., ZHUANG S., WU Z., ZHUANG Y., LIN Z., LI Z., LI D., XING E. *et al.* (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv :2306.05685*.

ZHU Y., KIROS R., ZEMEL R., SALAKHUTDINOV R., URTASUN R., TORRALBA A. & FIDLER S. (2015). Aligning books and movies : Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, p. 19–27.

5 Annexes

5.1 Modèles évalués

	#	Modèle	Nombre de paramètres	Taille des données d'entraînement	Corpus d'entraînement
Causal	1	LLAMA-2-70B[en] (Touvron *et al.*, 2023)	70B	2 billions de tokens	Un mélange de corpus disponibles publiquement, principalement en anglais
	2	Mistral-7B[?] (Jiang *et al.*, 2023)	7B	Non divulgué	Non divulgué
	3	BLOOM-7B1[en][fr][es] (Workshop *et al.*, 2022)	7B	1,6 To	ROOTS (Laurençon *et al.*, 2022), un mélange de corpus et de données pseudo-crawlées dans 59 langues
	4	Falcon-40B[en][fr][es]	40B	1 billion de tokens	RefinedWeb (Penedo *et al.*, 2023), un ensemble de données de Web filtrées et dédoublonnées
	5	GPT-J-6B[en] (Wang & Komatsuzaki, 2021)	6B	825 Gio	The Pile (Gao *et al.*, 2020), un mélange de corpus publics et de données Web en anglais
	6	OPT-66B[en] (Zhang *et al.*, 2022)	66B	180 milliards de tokens	Données crawlées sur le Web, principalement en anglais
	7	Vicuna-13B[en]* (Zheng *et al.*, 2023)	13B	125K conversations	LLAMA 2, affiné sur des conversations collectées sur ShareGPT.com, principalement en anglais
	8	Vicuna-7B[en]* (Zheng *et al.*, 2023)	7B	125K conversations	LLAMA 2, affiné sur des conversations collectées sur ShareGPT.com, principalement en anglais
	9	Medalpaca-7B[en]* (Han *et al.*, 2023)	7B	400K paires Q.R.	LLAMA 2, affiné sur des paires de questions-réponses médicales semi-générées en anglais
	10	Vigogne-13B[fr][en]*	13B	52K instructions	LLAMA 2, affiné sur des instructions en anglais automatiquement traduites en français
Masked	11	mBERT[en][fr][es] (Devlin *et al.*, 2019)	110M	Non divulgué	Un corpus comprenant 104 langues construit à partir de sources non divulguées
	12	XLM-R-large[en][fr][es] (Conneau *et al.*, 2020)	355M	2,5 To	Données CommonCrawl filtrées contenant 100 langues
	13	BERT-large[en] (Devlin *et al.*, 2019)	345M	3,3 milliards de mots	BookCorpus (Zhu *et al.*, 2015), un corpus composé de livres non publiés et de Wikipédia en anglais.
	14	RoBERTa-large[en] (Liu *et al.*, 2019)	355M	160 Gio	BooksCorpus (Zhu *et al.*, 2015), Wikipédia en anglais, et données Web crawlées
	15	Bio_ClinicalBERT[en] (Alsentzer *et al.*, 2019)	110M	2 millions de notes cliniques	MIMIC-III (Johnson *et al.*, 2016), une base de données contenant des dossiers médicaux électroniques de patients en soins intensifs hospitalisés
	16	ClinicalBERT[en] (Wang *et al.*, 2023a)	110M	1,2 milliard de mots	non divulgué
	17	MedBERT[en] (Vasantharajan *et al.*, 2022)	110M	57 millions de mots	Plusieurs corpus publics (y compris N2C2 (Luo *et al.*, 2020)) et articles médicaux crawlés depuis Wikipédia
	18	CamemBERT-large[fr] (Martin *et al.*, 2020)	335M	64 milliards de tokens	OSCAR (Suárez *et al.*, 2020), un corpus de données Web en français
	19	FlauBERT-large[fr] (Le *et al.*, 2020)	335M	13 milliards de tokens	Un mélange de Wikipédia français, de livres français et de données Web français
	20	DrBERT-4GB[fr] (Labrak *et al.*, 2023)	110M	1 milliard de mots	Un mélange de corpus biomédicaux disponibles publiquement en français (dont QuaeroFrenchMed (Névéol *et al.*, 2014)).
	21	CamemBERT-bio[fr] (Touchent *et al.*, 2023)	110M	413 millions de mots	Un mélange de corpus biomédicaux disponibles publiquement en français (dont E3C (Magnini *et al.*, 2021)).
	22	BETO[es] (Cañete *et al.*, 2020)	110M	3 milliards de mots	Wikipédia en espagnol et données espagnoles d'OPUS (Tiedemann, 2012)
	23	PatanaBERT[es]	110M	Non divulgué	Espagnol
	24	TulioBERT[es]	110M	Non divulgué	Espagnol
	25	BSC-BioEHR[es] (Carrino *et al.*, 2022)	110M	1,1 milliard de tokens	Un mélange de corpus biomédicaux, y compris des documents EHR et des données crawlées en espagnol
	26	BSC-Bio[es] (Carrino *et al.*, 2022)	110M	963 millions de tokens	Un mélange de corpus biomédicaux et de données crawlées en espagnol

TABLE 2 – Caractérisation des modèles de langage utilisés dans nos expériences en termes de paramètres et de corpus d'entraînement. Les modèles marqués avec [en] (respectivement [fr], [es]) sont fortement entraînés en anglais (respectivement en français, en espagnol). Les CLMs marqués avec * sont des versions affinées d'autres CLMs.

5.2 Ablation

Pour mieux comprendre la contribution de chaque étape de notre approche, nous avons mené une série d'expériences complémentaires.

***Prompts* de listage** Dans cette section, nous comparons les *prompts* de balisage adoptés aux *prompts* de listage. Dans les *prompts* de listage, les démonstrations listent simplement les mentions étiquetées. Le séparateur de liste est optimisé (de la même manière que les baliseurs) entre une virgule et un retour à la ligne. Le cas échéant, les phrases introductives demandent de lister les entités (et non de les baliser). Les résultats présentés dans le tableau 3 corroborent davantage notre choix de nous concentrer uniquement sur les *prompts* de balisage.

	Anglais					Français					Espagnol			
Modèle	WikiNER	CoNLL2003	E3C	n2c2	NCBI	WikiNER	QFP	E3C	EMEA	MEDLINE	WikiNER	CoNLL2002	E3C	CWL
Prompts de listage														
Mistral-7B	0.659	0.533	0.417	0.281	0.340	0.676	0.083	0.451	0.169	0.403	0.697	0.620	0.211	0.273
Prompts de balisage														
Mistral-7B	0.754	0.646	0.488	0.291	0.395	0.727	0.428	0.590	0.229	0.333	0.720	0.707	0.083	0.374

TABLE 3 – F1-mesures obtenues avec les *prompts* de listage vs. balisage

Échantillon et taille de l'échantillon Nous avons testé notre approche avec différents échantillons et différentes tailles d'échantillon pour un MLM : XLM-RoBERTa-large, et un CLM : Mistral-7B. Les résultats sont présentés dans le tableau 4. On peut noter que, bien que l'écart-type par rapport à p soit assez élevé, une différence significative est systématiquement observée entre les deux modèles sur des échantillons de même taille. Nous observons également que, lorsque le nombre d'instances annotées diminue, les performances du MLM chutent plus rapidement que celles du CLM.

	CoNLL2003			n2c2		
100 exemples annotés						
	p=1	*p=2*	*p=3*	*p=1*	*p=2*	*p=3*
Mistral-7B	0.646	0.626	0.714	0.291	0.178	0.215
XLM-R-large	0.826	0.814	0.786	0.462	0.478	0.526
50 exemples annotés						
	p=1	*p=2*	*p=3*	*p=1*	*p=2*	*p=3*
Mistral-7B	0.615	0.648	0.637	0.278	0.176	0.106
XLM-R-large	0.697	0.77	0.714	0.431	0.476	0.35
25 exemples annotés						
	p=1	*p=2*	*p=3*	*p=1*	*p=2*	*p=3*
Mistral-7B	0.509	0.599	0.52	0.152	0.252	0.116
XLM-R-large	0.487	0.588	0.637	0.393	0.361	0.283

TABLE 4 – 1-mesures obtenues avec différents échantillons et différentes tailles d'échantillons.

***Grid search* des caractéristiques** Afin d'évaluer la qualité du *greedy search* adopté pour trouver la meilleure combinaison de caractéristiques à incorporer dans le *prompt*, nous comparons cette méthode à un *grid search* naïf sur ces caractéristiques. Nous testons les 512 combinaisons des 9 caractéristiques identifiées, pour Mistral-7B sur CoNLL2003. Les scores trouvés par LOOCV varient entre 0,0 et 0,656 avec une valeur moyenne de 0,387 et une médiane de 0,46. La combinaison la plus

performante est : *phrases supplémentaires*, *auto-vérification*, *phrase d'introduction pour l'instance de test* et *demander une réponse longue pour l'auto-vérification*, qui est précisément la combinaison que nous avons trouvée initialement par le biais d'un *greedy search*, qui est environ 20 fois plus rapide et moins consommateur.

Nombre de démonstrations Le choix de limiter le nombre d'exemples annotés présentés aux modèles causaux, dans le *prompt*, à 10 maximum est dicté par deux contraintes.

Tout d'abord, les modèles imposent une limite sur le nombre de *tokens* passés en entrée, sur lesquels l'attention est calculée. La plupart des modèles utilisés ont 2048 *tokens* comme limite, mais Mistral-7B autorise jusqu'à 8096 *tokens*. Cette limite se traduit par une limite du nombre de phrases présentables dans le *prompt*, entre 40 et 50 pour Mistral-7B et entre 10 et 15 pour les modèles moins permissifs, selon les corpus et les *tokenizers*. Par exemple, considérons la tâche de détection de parties du corps dans la partie française d'E3C. Si l'on utilise Mistral-7B (et son *tokenizer*), la limite pratique est autour de 40 exemples, ce qui fait un *prompt* de 7779.5 *tokens* en moyenne. Si l'on utilise Bloom-7b1 (et son *tokenizer*), elle se situe autour de 11 exemples, ce qui fait 1851.5 *tokens* en moyenne.

En outre, l'amélioration apportée par l'ajout d'exemples ne semble pas conséquente, comme on peut l'observer dans le tableau 5, qui montre les résultats obtenus avec Mistral-7B en triplant le nombre d'exemples annotés. Notons que les améliorations marginales obtenues en triplant le nombre d'exemples, viennent à un coût considérable, surtout dans le contexte d'une complexité quadratique en fonction de la longueur du *prompt*.

Nous choisissons donc de limiter le prompt à 5/10 exemples, choisis selon le critère choisi (proximité TF-IDF à la phrase de référence ou nombre d'entité présentes). Au lieu de sélectionner les exemples de façon indépendante, Gupta *et al.* (2023) proposent de les sélectionner de façon interdépendante, afin d'améliorer la représentativité du *prompt*. Cette piste serait intéressante à implémenter dans notre système dans de futurs travaux.

	Anglais					Français					Espagnol			
Modèle	WikiNER	CoNLL2003	E3C	n2c2	NCBI	WikiNER	QFP	E3C	EMEA	MEDLINE	WikiNER	CoNLL2002	E3C	CWL
5/10 exemples														
Mistral-7B	0.754	0.646	0.488	0.291	0.395	0.727	0.428	0.590	0.229	0.333	0.720	0.707	0.083	0.374
15/30 exemples														
Mistral-7B	0.763	0.692	0.453	0.263	0.377	0.782	0.355	0.587	0.237	0.396	0.785	0.751	0.163	0.413

TABLE 5 – F1-mesures obtenues avec 5/10 exemples vs. avec 15/30 exemples

5.3 Description des types d'entités nommées

Etiquette	Nom du type (en singulier)	Description
PER	person names (a person's name)	These are names of persons such as real people or fictional characters.
FAC	facilities (a facility)	These are names of man-made structures such as infrastructure, buildings and monuments.
LOC	locations (a location)	These are names of geographical locations such as landmarks, cities, countries and regions.
ORG	organizations (an organization)	These are names of organizations such as companies, agencies and political parties.
FUNC	functions and jobs (a function or a job)	These are words that refer to a profession or a job.
ACTI	activities and behaviors (an activity or behavior)	These are words that refer to human activities, behaviors or events as well as governmental or regulatory activities.
ANAT	anatomy (an anatomy)	These are words that refer to the structure of the human body, its organs and their position, such as body parts or organs, systems, tissues, cells, body substances and embryonic structures.
CHEM	chemicals and drugs (a chemical or a drug)	These are words that refer to a substance or composition that has a chemical characteristic, especially a curative or preventive property with regard to human or animal diseases, such as drugs, antibiotics, proteins, hormones, enzymes and hazardous or poisonous substances.
CONC	concepts and ideas (a concept or an idea)	These are words that refer to a concept or an idea, such as a classification, an intellectual product, a language, a law or a regulation.
DEVI	medical devices (a device)	These are words that refer to a medical device used to administer care or perform medical research.
DISO	disorders (a disorder)	These are words that refer to an alteration of morphology, function or health of a living organism, animal or plant, such as congenital abnormalities, dysfunction, injuries, signs or symptoms or observations.
GENE	genes and molecular sequences (a gene or a molecular sequence)	These are words that refer to a gene, a genome or a molecular sequence.
GEOG	geographical areas (a geographical area)	These are words that refer to a country, a region or a city.
LIVB	living beings (a living being)	These are words that refer to a living being or a group of living beings, such as a person or a group of persons, a plant or a category of plants, an animal or a category of animals.
OBJC	objects (an object)	These are words that refer to anything animate or inanimate that affects the senses, such as physical manufactured objects.
OCCU	occupations (an occupation)	These are words that refer to a professional occupation or discipline.
ORGA	organizations (an organization)	These are words that refer to an organization such as healthcare related organizations.

TABLE 6 – Description des types d'entités nommées utilisées dans nos expériences sur l'anglais.

Etiquette	Nom du type (en singulier)	Description
PHEN	phenomema (a phenomemon)	These are words that refer to a phenomenon that occurs naturally or as a result of an activity, such as a biologic function.
PHYS	physiology (a physiology)	These are words that refer to any element that contributes to the mechanical, physical and biochemical functioning or organization of living organisms and their components.
PROC	procedures (a procedure)	These are words that refer to an activity or a procedure that contributes to the diagnosis or treatment of patients, the information of patients, the training of medical personnel or biomedical research.
EVENT	events (an event)	These are words that refer to actions, states, and circumstances that are relevant to the clinical history of a patient such as pathologies and symptoms, or more generally words like "enters", "reports" or "continue".
TIMEX3	time expressions (a time expression)	These are time expressions such as dates, times, durations, frequencies, or intervals.
RML	results and measurements (a result or a measurement)	These are test results, results of laboratory analyses, formulaic measurements, and measure values.
ACTOR	actors (an actor)	These are words that refer patients, healthcare professionals, or other actors relevant to the clinical history of a patient.
Abbreviation	abbreviations (an abbreviation)	These are words that refer to abbreviations.
Body_Part	body parts (a body part)	These are words that refer to organs and anatomical parts of persons.
Clinical_Finding	clinical findings (a clinical finding)	These are words that refer to observations, judgments or evaluations made about patients.
Diagnostic_Procedure	diagnostic procedures (a diagnostic procedure)	These are words that refer to tests that allow determining the condition of the individual.
Disease	diseases (a disease)	These are words that describe an alteration of the physiological state in one or several parts of the body, due to generally known causes, manifested by characteristic symptoms and signs, and whose evolution is more or less predictable.
Family_Member	family members (a family member)	These are words that refer to family members.
Laboratory_or_Test_Result	laboratory or test results (a laboratory or test result)	These are words that refer to any measurement or evaluation obtained from a diagnostic support examination.
Laboratory_Procedure	laboratory procedures (a laboratory procedure)	These are words that refer to tests that are performed on various patient samples that allow diagnosing diseases by detecting biomarkers and other parameters. Blood, urine, and other fluids and tissues that use biochemical, microbiological and/or cytological methods are considered.
Medication	medications (a medication)	These are words that refer to medications or drugs used in the treatment and/or prevention of diseases, including brand names and generics, as well as names for groups of medications.
Procedure	procedures (a procedure)	These are words that refer to activities derived from the care and care of patients.
Sign_or_Symptom	signs or symptoms (a sign or symptom)	These are words that refer to manifestations of a disease, determined by medical examination or perceived and expressed by the patient.
Therapeutic_Procedure	therapeutic procedures (a therapeutic procedure)	These are words that refer to activities or treatments that are used to prevent, repair, eliminate or cure the individual's disease.
CompositeMention	composite mentions of diseases (a composite mention of diseases)	These are words that refer to mentions of multiple diseases, such as "colorectal, endometrial, and ovarian cancers".
DiseaseClass	disease classes (a disease class)	These are words that refer to classes of diseases, such as "an autosomal recessive disease".
Modifier	modifiers (a modifier of diseases)	These are words that refer to modifiers of diseases, such as "primary" or "C7-deficient".
SpecificDisease	diseases (a disease)	These are words that refer to specific diseases, such as "diastrophic dysplasia".

TABLE 7 – Description des types d'entités nommées utilisées dans nos expériences sur l'anglais, suite.

Etiquette	Nom du type (en singulier)	Description
PER	de noms de personnes (un nom de personne)	Il s'agit des noms de personnes, qu'elles soient réelles ou fictives.
FAC	de productions humaines (une production humaine)	Il s'agit des noms de structures faites par les humains comme des infrastructures, des bâtiments ou des monuments.
LOC	de lieux (un lieu)	Il s'agit des noms de lieux comme des endroits, villes, pays ou régions.
ORG	d'organisations (une organisation)	Il s'agit des noms d'organisations comme des entreprises, des agences ou des partis politiques.
FUNC	de fonctions et métiers (une fonction ou un métier)	Il s'agit de mots qui se rapportent à une activité professionnelle.
ANAT	d'anatomie (une partie du corps)	Il s'agit d'une entité se rapportant à la structure du corps humain, ses organes et leur position. Il s'agit principalement des parties du corpus ou organes, des appareils, des tissus, des cellules, des substances corporelles et des organismes embryonaires.
CHEM	de médicaments et substances chimiques (un médicament ou une substance chimique)	Il s'agit d'une substance ou composition présentant des propriétés chimiques caractéristiques, en particulier des propriétés curatives ou préventives à l'égard des maladies humaines ou animales. Il s'agit principalement des médicaments disponibles en pharmacie, des antibiotiques, des proteines, des hormones, des substances dangereuses, des enzymes.
DEVI	de matériel (un matériel)	Il s'agit d'un matériel utilisé pour administrer des soins ou effectuer des recherches médicales.
DISO	de problèmes médicaux (un problème médical)	Il s'agit d'une altération de la morphologie, des fonctions, ou de la santé d'un organisme vivant, animal ou végétal. Il peut s'agir de malformations, de maladies, de blessure, de signe ou symptome ou d'une observation.
GEOG	de zones géographiques (une zone géographique)	Il s'agit d'un pays, une région, ou une ville.
LIVB	d'êtres vivants (un être vivant)	Il s'agit d'un être vivant ou groupe d'êtres vivants. Il peut s'agir d'une personne ou d'un groupe de personnes, d'une plante ou d'une catégorie de végétaux, d'un animal ou d'une catégorie d'animaux.
OBJC	d'objets (un objet)	Il s'agit de tout ce qui, animé ou inanimé, affecte les sens. Ici, il s'agit principalement d'objets physiques manufacturés.
PHEN	de phénomènes (un phénomène)	Il s'agit d'un phénomène qui se produit naturellement ou à la suite d'une activité. Il s'agit principalement de fonctions biologiques.
PHYS	de physiologie (une physiologie)	Il s'agit de tout élément contribuant au fonctionnement ou à l'organisation mécanique, physique et biochimique des organismes vivants et de leurs composants.
PROC	de procédures (une procédure)	Il s'agit d'une activité ou procédure contribuant au diagnostic ou au traitement des patients, à l'information des patients, la formation du personnel médical ou à la recherche biomédicale.
EVENT	d'événements (un événement)	Il s'agit d'une action, d'un état ou d'une circonstance qui est pertinent pour l'histoire clinique d'un patient. Il peut s'agir de pathologies et symptômes, ou plus généralement de mots comme "entre", "rapporte" ou "continue".
TIMEX3	d'expressions temporelles (une expression temporelle)	Il s'agit d'expressions temporelles comme des dates, heures, durées, fréquences, ou intervalles.
RML	de résultats et mesures (un résultat ou une mesure)	Il s'agit de résultats d'analyses de laboratoire, de mesures formelles, et de valeurs de mesure.
ACTOR	d'acteurs (un acteur)	Il s'agit de patients, de professionnels de santé, ou d'autres acteurs pertinents pour l'histoire clinique d'un patient.

TABLE 8 – Description des types d'entités nommées utilisées dans nos expériences sur le français.

Etiquette	Nom du type (en singulier)	Description
PER	nombres de personas (un nombre de persona)	Estos son nombres de personas, ya sean reales o personajes ficticios.
FAC	instalaciones (una instalación)	Estos son nombres de estructuras hechas por el hombre como infraestructura, edificios y monumentos.
LOC	lugares (un lugar)	Estos son nombres de ubicaciones geográficas como hitos, ciudades, países y regiones.
ORG	organizaciones (una organización)	Estos son nombres de organizaciones como empresas, agencias y partidos políticos.
ACTI	actividades y comportamientos (una actividad o comportamiento)	Estas son palabras que se refieren a actividades humanas, comportamientos o eventos, así como actividades gubernamentales o regulatorias.
ANAT	anatomía (una anatomía)	Estas son palabras que se refieren a la estructura del cuerpo humano, sus órganos y su posición, como partes del cuerpo u órganos, sistemas, tejidos, células, sustancias corporales y estructuras embrionarias.
CHEM	productos químicos y medicamentos (un producto químico o un medicamento)	Estas son palabras que se refieren a una sustancia o composición que tiene una característica química, especialmente una propiedad curativa o preventiva con respecto a las enfermedades humanas o animales, como medicamentos, antibióticos, proteínas, hormonas, enzimas y sustancias peligrosas o venenosas.
CONC	conceptos e ideas (un concepto o una idea)	Estas son palabras que se refieren a un concepto o una idea, como una clasificación, un producto intelectual, un idioma, una ley o un reglamento.
DEVI	dispositivos médicos (un dispositivo)	Estas son palabras que se refieren a un dispositivo médico utilizado para administrar atención o realizar investigaciones médicas.
DISO	trastornos (un trastorno)	Estas son palabras que se refieren a una alteración de la morfología, la función o la salud de un organismo vivo, animal o vegetal, como anomalías congénitas, disfunción, lesiones, signos o síntomas u observaciones.
GENE	genes y secuencias moleculares (un gen o una secuencia molecular)	Estas son palabras que se refieren a un gen, un genoma o una secuencia molecular.
GEOG	áreas geográficas (un área geográfica)	Estas son palabras que se refieren a un país, una región o una ciudad.
LIVB	seres vivos (un ser vivo)	Estas son palabras que se refieren a un ser vivo o un grupo de seres vivos, como una persona o un grupo de personas, una planta o una categoría de plantas, un animal o una categoría de animales.
OBJC	objetos (un objeto)	Estas son palabras que se refieren a cualquier cosa animada o inanimada que afecte los sentidos, como objetos físicos fabricados.
OCCU	ocupaciones (una ocupación)	Estas son palabras que se refieren a una ocupación o disciplina profesional.
ORGA	organizaciones (una organización)	Estas son palabras que se refieren a una organización, por ejemplo organizaciones relacionadas con la salud.
PHEN	fenómenos (un fenómeno)	Estas son palabras que se refieren a un fenómeno que ocurre naturalmente o como resultado de una actividad, por ejemplo una función biológica.

TABLE 9 – Description des types d'entités nommées utilisées dans nos expériences sur l'espagnol.

Etiquette	Nom du type (en singulier)	Description
PHYS	fisiología (una fisiología)	Estas son palabras que se refieren a cualquier elemento que contribuya al funcionamiento mecánico, físico y bioquímico o la organización de los organismos vivos y sus componentes.
PROC	procedimientos (un procedimiento)	Estas son palabras que se refieren a una actividad o un procedimiento que contribuye al diagnóstico o tratamiento de pacientes, la información de pacientes, la capacitación del personal médico o la investigación biomédica.
EVENT	eventos (un evento)	Estas son palabras que se refieren a acciones, estados y circunstancias que son relevantes para la historia clínica de un paciente, como patologías y síntomas, o más generalmente palabras como "entra", "reporta" o "continúa".
TIMEX3	expresiones de tiempo (una expresión de tiempo)	Estas son expresiones de tiempo como fechas, horas, duraciones, frecuencias o intervalos.
RML	resultados y mediciones (un resultado o una medida)	Estos son resultados de análisis de laboratorio, mediciones formales y valores de medición.
ACTOR	actores (un actor)	Estas son palabras que se refieren a pacientes, profesionales de la salud u otros actores relevantes para la historia clínica de un paciente.
Abbreviation	abreviaciones (una abreviación)	Estas son los casos de siglas y acrónimos.
Body_Part	partes del cuerpo (una parte del cuerpo)	Estas son palabras que se refieren a òrganos y partes anatómicas de personas.
Clinical_Finding	hallazgos clínicos (un hallazgo clínico)	Estas son palabras que se refieren a observaciones, juicios o evaluaciones que se hacen sobre los pacientes.
Diagnostic_Procedure	procedimientos diagnósticos (un procedimiento diagnóstico)	Estas son palabras que se refieren a exámenes que permiten determinar la condición del individuo.
Disease	enfermedades (una enfermedad)	Estas son palabras que describen una alteración del estado fisiológico en una o varias partes del cuerpo, por causas en general conocidas, manifestada por síntomas y signos característicos, y cuya evolución es más o menos previsible.
Family_Member	miembros de la familia (un miembro de la familia)	Estas son palabras que se refieren a miembros de la familia.
Laboratory_or_Test_Result	resultados de exámenes de laboratorio u otras pruebas (un resultado de un examen de laboratorio u otra prueba)	Estas son palabras que se refieren a cualquier medición o evaluación obtenida a partir de un exámen de apoyo diagnóstico.
Laboratory_Procedure	procedimientos de laboratorio (un procedimiento de laboratorio)	Estas son palabras que se refieren a exámenes que se realizan en diversas muestras de pacientes que permiten diagnosticar enfermedades mediante la detección de biomarcadores y otros parámetros. Se consideran los análisis de sangre, orina, y otros fluidos y tejidos que emplean métodos bioquímicos, microbiológicos y/o citológicos.
Medication	medicamentos o drogas (un medicamento o una droga)	Estas son palabras que se refieren a medicamentos o drogas empleados en el tratamiento y/o prevención de enfermedades, incluyendo marcas comerciales y genéricos, así como también nombres para grupos de medicamentos.
Procedure	procedimientos (un procedimiento)	Estas son palabras que se refieren a actividades derivadas de la atención y el cuidado de los pacientes.
Sign_or_Symptom	signos o síntomas (un signo o un síntoma)	Estas son palabras que se refieren a manifestaciones de una enfermedad, determinadas mediante la exploración médica o percibidas y expresadas por el paciente.
Therapeutic_Procedure	procedimientos terapéuticos (un procedimiento terapéutico)	Estas son palabras que se refieren a actividades o tratamientos que es empleado para prevenir, reparar, eliminar o curar la enfrmedad del individuo.

TABLE 10 – Description des types d'entités nommées utilisées dans nos expériences sur l'espagnol, suite.

5.4 Empreinte carbone

	#	Modèle	Anglais					Français					Espagnol			
			WikiNER	CoNLL2003	E3C	n2c2	NCBI	WikiNER	QFP	E3C	EMEA	MEDLINE	WikiNER	CoNLL2002	E3C	CWL
Approches few-shot																
Causal	1	LLAMA-2-70B	46	44	126	233	54	85	131	129	273	284	41	76	114	344
	2	Mistral-7B	4	6	12	24	8	5	8	14	13	25	7	5	11	27
	3	BLOOM-7B1	4	6	10	26	9	8	13	9	26	20	4	8	8	18
	4	Falcon-40B	49	45	56	176	45	31	58	75	162	129	33	25	82	99
	5	GPT-J-6B	7	6	8	23	7	5	8	13	21	17	6	6	13	28
	6	OPT-66B	73	50	120	253	96	38	64	138	273	240	57	52	106	247
	7	Vicuna-13B	10	11	20	52	11	11	12	18	33	40	10	11	22	51
	8	Vicuna-7B	6	8	14	17	6	5	10	10	24	14	8	6	13	27
	9	Medalpaca-7B	8	4	17	24	10	7	14	11	19	21	5	8	15	26
	10	Vigogne-13B	14	14	29	37	11	13	20	26	36	39	11	14	32	44
Masked	11	mBERT	2	1	2	2	2	2	2	2	1	1	1	2	1	2
	12	XLM-R-large	2	2	2	1	2	2	2	2	2	2	1	1	1	2
	13	BERT-large	2	1	2	2	2	-	-	-	-	-	-	-	-	-
	14	RoBERTa-large	1	2	2	2	2	-	-	-	-	-	-	-	-	-
	15	Bio_ClinicalBERT	2	2	1	2	1	-	-	-	-	-	-	-	-	-
	16	ClinicalBERT	1	1	2	2	1	-	-	-	-	-	-	-	-	-
	17	MedBERT	2	2	1	1	1	-	-	-	-	-	-	-	-	-
	18	CamemBERT-large	-	-	-	-	-	1	1	1	2	2	-	-	-	-
	19	FlauBERT-large	-	-	-	-	-	2	2	2	2	2	-	-	-	-
	20	DrBERT-4GB	-	-	-	-	-	2	2	2	2	2	-	-	-	-
	21	CamemBERT-bio	-	-	-	-	-	1	2	2	2	2	-	-	-	-
	23	BETO	-	-	-	-	-	-	-	-	-	-	2	1	1	1
	23	PatanaBERT	-	-	-	-	-	-	-	-	-	-	2	2	2	2
	24	TulioBERT	-	-	-	-	-	-	-	-	-	-	1	2	2	1
	25	BSC-BioEHR	-	-	-	-	-	-	-	-	-	-	2	2	2	2
	26	BSC-Bio	-	-	-	-	-	-	-	-	-	-	2	2	2	2
Skyline en utilisant toutes les données à disposition																
		RoBERTa-large	647	68	5	12	24	-	-	-	-	-	-	-	-	-
		CamemBERT-large	-	-	-	-	-	595	15	4	5	8	-	-	-	-
		BETO	-	-	-	-	-	-	-	-	-	-	579	41	3	21

TABLE 11 – Ce tableau présente les émissions carbone (en g) de l'optimisation de chaque modèle sur sur le jeu de validation de chaque corpus. Pour les CLMs , il s'agit du *greedy search* sur les potentiels caractéristiques du *prompt* via la validation croisée. Pour les MLM, cela correspond au *fine-tuning* des paramètres du modèle (et à l'apprentissage des couches de classification introduites).

	#	Modèle	Anglais					Français					Espagnol			
			WikiNER	CoNLL2003	E3C	n2c2	NCBI	WikiNER	QFP	E3C	EMEA	MEDLINE	WikiNER	CoNLL2002	E3C	CWL
Approches few-shot																
Causal	1	LLAMA-2-70B	812	147	36	196	33	508	11	13	92	47	514	201	11	198
	2	Mistral-7B	234	35	8	59	21	148	3	4	27	20	261	50	2	32
	3	BLOOM-7B1	220	33	8	44	16	255	3	5	38	29	261	47	2	46
	4	Falcon-40B	600	109	26	144	46	722	9	19	155	70	752	154	9	157
	5	GPT-J-6B	146	17	4	53	20	245	2	6	14	26	154	40	3	53
	6	OPT-66B	765	139	33	185	63	971	12	27	179	93	993	196	12	217
	7	Vicuna-13B	314	47	11	63	24	363	5	8	61	46	502	67	4	74
	8	Vicuna-7B	146	17	4	53	20	246	2	6	14	26	155	65	3	53
	9	Medalpaca-7B	192	24	5	39	14	98	2	2	17	13	172	53	1	21
	10	Vigogne-13B	322	49	11	65	24	245	5	6	44	33	361	68	3	66
Masked	11	mBERT	14	4	<1	2	<1	15	1	<1	1	1	13	2	<1	2
	12	XLM-R-large	14	4	<1	2	<1	15	1	<1	1	1	13	2	<1	2
	13	BERT-large	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
	14	RoBERTa-large	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
	15	Bio_ClinicalBERT	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
	16	ClinicalBERT	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
	17	MedBERT	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
	18	CamemBERT-large	-	-	-	-	-	15	1	<1	1	1	-	-	-	-
	19	FlauBERT-large	-	-	-	-	-	15	1	<1	1	1	-	-	-	-
	20	DrBERT-4GB	-	-	-	-	-	17	1	<1	1	1	-	-	-	-
	21	CamemBERT-bio	-	-	-	-	-	15	1	<1	1	1	-	-	-	-
	22	BETO	-	-	-	-	-	-	-	-	-	-	13	2	<1	2
	23	PatanaBERT	-	-	-	-	-	-	-	-	-	-	13	2	<1	2
	24	TulioBERT	-	-	-	-	-	-	-	-	-	-	13	2	<1	2
	25	BSC-BioEHR	-	-	-	-	-	-	-	-	-	-	13	2	<1	2
	26	BSC-Bio	-	-	-	-	-	-	-	-	-	-	13	2	<1	2
Skyline en utilisant toutes les données à disposition																
		RoBERTa-large	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
		CamemBERT-large	-	-	-	-	-	15	1	<1	1	1	-	-	-	-
		BETO	-	-	-	-	-	-	-	-	-	-	13	2	<1	2

TABLE 12 – Ce tableau présente les émissions carbone (en g) de l'inférence, en utilisant chaque modèle, sur sur le jeu de test de chaque corpus.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Réduction des répétitions dans la Traduction Automatique Neuronale

Marko Avila Anabel Rebollo Josep Crego
4 rue du Port aux Vins, F-92 150 Suresnes, France
{mavila,arebollo,jcrego}@chapsvision.com

RÉSUMÉ

Actuellement, de nombreux systèmes TAL utilisent des décodeurs neuronaux pour la génération de textes, qui font preuve d'une capacité impressionnante à générer des textes approchant les niveaux de fluidité humaine. Toutefois, dans le cas des réseaux de traduction automatique, ils sont souvent confrontés à la production de contenu répétitif, également connu sous le nom de diction répétitive ou de répétition de mots, un aspect pour lequel ils n'ont pas été explicitement entraînés. Bien que cela ne soit pas intrinsèquement négatif, cette répétition peut rendre l'écriture monotone ou maladroite si elle n'est pas utilisée intentionnellement pour l'emphase ou des fins stylistiques. La répétition de mots a été traitée par des méthodes post-hoc pendant l'inférence, contraignant le réseau à examiner des hypothèses auxquelles le système avait initialement attribué une plus faible probabilité. Dans cet article, nous implémentons une méthode qui consiste à pénaliser les répétitions lors de l'apprentissage et qui s'inspire des principes du *label smoothing*. Conformément à cette méthode, nous modifions la distribution de la vérité terrain afin d'orienter le modèle de manière à décourager ces répétitions. Les résultats de nos expériences montrent que les méthodes proposées permettent de contrôler le problème de la répétition dans les moteurs neuronaux de traduction automatique sans compromis en termes d'efficacité ou de qualité des traductions.

ABSTRACT

Reducing Repetitions in Neural Machine Translation

Many contemporary NLP systems rely on neural decoders for text generation, which demonstrate an impressive ability to generate text approaching human fluency levels. However, in the case of neural machine translation networks, they often grapple with the production of repetitive content, also known as repetitive diction or word repetition, an aspect they weren't explicitly trained to address. While not inherently negative, this repetition can make writing seem monotonous or awkward if not used intentionally for emphasis or stylistic purposes. Repetitions have been addressed through post-hoc methods during inference, compelling the network to consider hypotheses it initially assigned lower probability. In this paper, we implement a repetition penalty method applied at learning inspired by the principles of label smoothing. In line with label smoothing, we modify the ground-truth distribution to steer the model towards discouraging repetitions. Experiments show the ability of the proposed methods in reducing repetitions within neural machine translation engines, without compromising efficiency or translation quality.

MOTS-CLÉS : Traduction automatique neuronale ; Génération de texte ; Répétitions.

KEYWORDS: Neural machine translation ; Text generation ; Repetitions.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 198–210.

1 Introduction

This study addresses the issue of word repetition in machine translation, which involves the repeated occurrence of words, phrases, or ideas throughout the translation process. More specifically, we focus on repetitions that typically occur when translating synonyms or semantically equivalent phrases found in the source sentence. These repetitions lead to diminished readability of the text, potentially causing boredom or confusion for the reader and creating the perception of verbose or awkward writing. Consider, for instance, the following French sentence :

— *nous avons **lutté** contre l'infodémie en **combattant** les mythes par des informations fiables.*

and its two corresponding English translations :

— *We have **combated** the infodemia by **combating** myths with reliable information.*

— *We have **fought** the infodemia by **combating** myths with reliable information.*

Since both translations are grammatically correct and semantically equivalent, the first one is clumsy due to word repetition [1], while the second one effectively avoids repetition by suggesting alternative translations *fought* and *combating*, for the French words *lutté* and *combattant*, resulting in a smoother and more preferable translation.

In neural machine translation, repetition often arises when the model faces input sentences containing synonyms, leading these synonymous terms to be translated into identical words. Although lacking numerical support for our observation, the repetition issue becomes more salient when utilizing a model trained across multiple domains, highlighting the dearth of lexical diversity. We attribute this phenomenon to a lack of diversity in the decoder module. Although this type of repetition may occur with low frequency, it is highly concerning as it vividly illustrates the lack of fluency in translations.

However, repetitions do not always have a negative impact on readability. Without aiming to be exhaustive : i) repetitions play a role when summarizing information or reinforcing a concept ; ii) common expressions are formed using word repetitions, and altering them to eliminate repetition would alter their intended meaning ; iii) in highly specialized domains, expressions convey precise meanings that disallow being reformulated. The following examples illustrate these observations :

i) *once **closed**, the door stays **closed***
ii) ***over** and **over** ; **to be** or not **to be** ; **step** by **step***
iii) *the congenital **muscular** dystrophy in newborns presenting with **muscular** hypotonia*

As previously introduced, finding suitable alternatives without altering the meaning of a sentence can be a challenging task. In this work, we propose a method applied in training designed to teach the model to discourage certain repetitions, thereby alleviating the need for difficult decisions during inference. Next, we summarize the main contributions of this work :

— We propose a method that discourages repetitions during the training phase by adjusting the ground-truth distribution so as to penalize repetitions more severely.
— We introduce a technique for gathering examples containing both, acceptable repetitions and repetitions that hinder fluency, which are then utilized during the training phase.
— We build a curated test set that includes various types of word repetitions found in machine translations. Evaluation on this test set provides deeper insights into the repetitions issue.

Repetitions can manifest in various forms, including single words, phrases, and larger segments of content. However, this work concentrates on repetitions manifested through the repetition of linguistic

1. Note that repetitions can diminish readability, even when they occur as inflectional variants. Is the case of the repeated words in our example, *combated* and *combating*.

words, which are more commonly observed in machine translations. Note that linguistic words are typically decomposed into multiple tokens as taken into account by neural networks.

2 Related Work

The fluency levels achieved by LLMs are widely acknowledged to be high, primarily owing to the extensive availability of monolingual datasets, which surpasses that of standard neural machine translation (NMT) models trained solely on parallel texts. To the best of our knowledge, no dedicated research has been conducted on addressing the repetition issue tackled in this work within NMT systems. Closely related, (Welleck *et al.*, 2019) describe a method to train neural language models that in addition to maximizing likelihood to model the overall sequence probability distribution, also includes an unlikelihood term in the loss function to correct known biases such as repeated tokens. (Li *et al.*, 2020) use the same approach to control copy effect and repetitions observed in dialogue tasks. (Su *et al.*, 2022) present a contrastive solution to encourage diversity while maintaining coherence in the generated text.

Various studies have addressed diversity in neural MT systems, which is a closely related topic. Sampling predictions from the output distribution can be an effective decoding strategy for back-translation, as described by (Edunov *et al.*, 2018), or sampling from less likely tokens (Holtzman *et al.*, 2020). Results show that such techniques enlarge diversity and richness of the generated translations when compared to data generated by beam or greedy search, but introduce semantic inconsistency in translations. In (Lin *et al.*, 2022) is proposed a multi-candidate optimization framework for augmenting diversity. The authors propose to guide an NMT model to learn more diverse translations from its candidate translations based on reinforcement learning. During training, the model generates multiple candidate translations, of which rewards are quantified according to their diversity and quality.

A different approach attempts to condition the decoding procedure with diverse signals. Typically, (Shu *et al.*, 2019) use syntactic codes to condition the translation process. (Lachaux *et al.*, 2020) replace the syntactic codes with latent domain variables derived from target sentences. Similarly, (Schioppa *et al.*, 2021) use prefix-based control tokens and vector-based interventions for controlling output translations from a NMT system. In the context of paraphrase generation (Vahtola *et al.*, 2023) propose a translation-based guided paraphrase generation model that learns useful features for promoting surface form variation in generated paraphrases.

3 Adjusting the ground-truth distribution

Throughout the training process, at every time-step t, neural machine translation networks generate predictions over the target-side vocabulary based on the input x and previous predictions $y_{<t}$:

$$p_t^i = p(y_t^i | x, y_{<t}),\ i \in [1, ..., V]$$

where V indicates the size of the target vocabulary.

The loss function evaluates the neural network's capacity to model the training data by comparing

t	1	2	3	4	5	6
r	I	like	cookies	and	cookies	.

$\mathcal{M}$

.	0	0	0	0	0	0
I	0	0	0	0	0	0
and	0	0	0	0	0	0
like	0	0	0	0	0	0
cookies	0	0	0	0	1	0

FIGURE 1 – Matrix $\mathcal{M}$ for the ground-truth r ='*I like cookies and cookies.*'. Rows t and r represent respectively the time-step and the corresponding ground-truth token. A reduced model vocabulary (matrix rows) is used to facilitate reading.

its predictions to a reference target vector $r = [r_1, r_2, ..., r_T]$, where T denotes the sequence length. This loss is utilized to update the network's parameters, aiming to minimize the observed error in the model. The loss at time-step t is usually computed as the cross-entropy between the model predictions $p_t = [p_t^1, ..., p_t^V]$ and the ground-truth distribution $q_t = [q_t^1, ..., q_t^V]$:

$$\mathcal{L}_t = -\sum_{i=1}^{V} q_t^i \, log(p_t^i) \tag{1}$$

Note that the vector q_t is a one-hot encoding representation of r_t, with all entries set to 0 except for the token indicated by r_t, which is set to 1. Addressing the over-fitting risk illustrated by the previous q_t distribution, label smoothing (Szegedy *et al.*, 2015; Müller *et al.*, 2019) (LS) is widely employed to achieve a smoother distribution :

$$q_t^{\epsilon LS} = (1 - \epsilon) q_t + \frac{\epsilon}{V} \tag{2}$$

with ϵ being a commonly small hyper-parameter. [2]

LS can be interpreted as penalizing the probability of the ground-truth class by a factor of $1 - \epsilon$, while evenly distributing the removed probability mass among all classes, ϵ/V. Building upon a strategy akin to label smoothing, we make additional adjustments to the ground-truth distribution and reduce the likelihood of repeated tokens, with the goal of enabling the model to learn to predict repetitions with lower probability. We introduce a matrix, denoted as $\mathcal{M}_{V \times T}$, which indicates whether the ground-truth token r_t is also present in the preceding time-steps. [3] Figure 1 illustrates an example of matrix $\mathcal{M}$ with ground-truth *I like cookies and cookies.* as translation of the French sentence *J'aime les cookies et les biscuits.* with a model vocabulary of 5 tokens (matrix rows). Both French terms *cookies* and *biscuits* are correctly translated into English as *cookies*, yet this choice clearly reduces the fluency and clarity of the translation. As it can be seen, only $\mathcal{M}_{[i=5,t=5]}$ is set to 1 since only r_5 ='*cookies*' occurs in a preceding time-step ($t = 3$).

We consequently update the ground-truth distribution following :

$$q_t^{\epsilon LS \alpha \mathcal{M}} = (1 - \epsilon)(1 - \alpha \mathcal{M}_t) \; q_t + \frac{\epsilon}{V} \tag{3}$$

2. $\epsilon = 0$ yields the initial distribution q_t, whereas $\epsilon = 1$ implies a uniform distribution.
3. Note that repetitions are computed over words while matrix $\mathcal{M}$ refers to tokens $r \in V$ for each time-step $t \in T$.

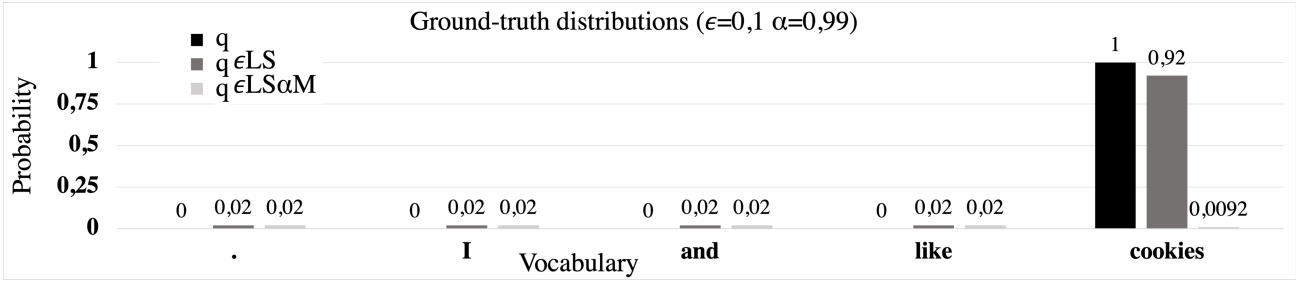


FIGURE 2 – Ground-truth distributions for the 5^{th} time-step of our example : the original one-hot encoding q ; adjusted with label smoothing $q^{\epsilon LS}$; and further adjusted with repetitions $q^{\epsilon L\alpha\mathcal{M}}$.

where α is a hyper-parameter, and $\alpha\mathcal{M}$ is used as a penalty, much like ϵ in the case of LS. Note that only the label smoothing probabilities discounted are distributed among all classes. As a result, time-steps with repeated tokens (such as $t = 5$ in our example) do not constitute proper probability distributions, as their sum does not add to 1. Figure 2 illustrates ground-truth distributions for our example at time-step $t = 5$: the original one-hot encoding q ; the original distribution adjusted using label smoothing $q^{\epsilon LS}$, and further adjusted using repetitions $q^{\epsilon LS\alpha\mathcal{M}}$.[4] A significant challenge with the aforementioned techniques that modify q distribution with repetitions is their limited impact on the training process, primarily caused by the scarcity of repeated tokens in datasets. In the following section, we present alternative approaches to address this challenge.

4 Gathering Examples with Repetitions

As previously depicted, our intention is to instruct the model to minimize certain repetitions while preserving others deemed necessary for an accurate translation. To achieve this, we must compile a relatively large dataset of examples that demonstrate this behavior to the model. We initially focus on repetitions of content words such as *nouns*, *adjectives*, *verbs*, and *adverbs*. Function words, which serve a distinct grammatical role in a sentence, are excluded from this analysis. Current MT networks reliably generate these words based on their understanding of grammatical correctness.

Given that training corpora contain a relatively small number of repetitions, and these are manually curated, often comprising only acceptable repetitions, we opt to focus solely on repetitions found in machine translations. Accordingly, we translate the source sentences (src) of our training examples to generate synthetic translations (hyp). Repetitions that detrimentally impact fluency are only considered if they appear in such translation. Identifying examples containing repetitions following the previous morpho-syntactic patterns is straightforward. However, the challenge lies in discerning which repetitions degrade the fluency of translations and which do not. We follow the next filtering steps to select repetitions degrading fluency :

— We first word align the source (src) and synthetic (hyp) sentences and eliminate repetitions in the synthetic sentences that also align with repeated words in the source sentence. Word alignments are performed by the Giza++ (Och & Ney, 2003) toolkit[5]. This approach is

4. As previously discussed, distribution $q^{\epsilon LS\alpha\mathcal{M}}$ does not form a proper distribution since probabilities do not add to 1 $(0,02 + 0,02 + 0,02 + 0,02 + 0,0092 = 0,0892)$. We leave for future experiments the normalization of the output scores in order to allow for a valid probability distribution.

5. https://github.com/moses-smt/giza-pp.

	Degrading	*Acceptable*
src	The home was modest and frugal	I want to know what you mean
hyp	La maison était modeste et modeste	Je veux savoir ce que tu veux dire
tgt	La maison était modeste et économe	Je veux savoir ce que tu veux dire

TABLE 1 – Two synthetic translations containing repeated words (underlined) and their corresponding source (hyp) and reference translation (tgt) sentences. Shades of grey indicate word alignments between repeated words and their alignments in the src/tgt sentences.

based on the premise that if repeated words are necessary in a human-generated sentence, the corresponding translation may similarly require the repetition. This holds true for repetitions used to reinforce a concept or in highly specialized domains.

— Next, we also align the synthetic (hyp) and target (tgt) sentences word by word and remove repetitions of the synthetic sentences (hyp) aligned to repeated words in the reference translation. The same previous rationale remains consistent, now encompassing examples of common expressions that necessitate word repetitions.

The resulting set of examples with repetitions from src/hyp training pairs will be regarded as instances that the model needs to learn to discourage. Consequently, we utilize them for training after annotating the repeated words in their respective $\mathcal{M}$ matrices. Table 1 illustrates the procedure previously outlined for two synthetic translations (hyp) containing repetitions.

The repeated word on the left-side example is the French adjective *modeste* while the verb *veux* is repeated in the right-side example. Once word alignments are computed between synthetic translations and their respective source and target sentences, the example on the right is not classified as a repetition degrading fluency. This is because both instances of *veux* are aligned with repeated words in the target reference translation, suggesting that the repetition is motivated. Concerning the example on the left side, neither the source nor the reference target sentences contain repeated words, suggesting that the repetition stems from a lack of diversity when translating *modest* and *frugal*, thus hindering fluency.

We employ the left-side example to train the model to identify it as a repetition to be avoided. The corresponding matrix $\mathcal{M}$ marks the second occurrence of the word *modeste* with a 1, thereby incurring in a significant loss if the model predicts it with high probability. The example on the right is used without penalization, thus instructing the model to reproduce the repetition.

It's worth noting that the presented approach does not require any alterations to the network architecture and maintains the same training and inference efficiency.

5 Experimental Framework

We evaluate the proposed methods in an English-to-French translation task. Thus, we utilize English-French parallel corpora freely obtained from the Opus website [6]. We strive for balanced utilization across various domains and ensure the inclusion of clean parallel data whenever possible. Due to the extensive volume of French-English parallel sentences accessible we randomly choose a subset exceeding 7 million examples that we employ as *Training set*.

For testing, we make use of English-French *News-test* (2008 to 2013) datasets made available

6. http://opus.nlpl.eu

	Training set		*Repetition-test*	
Type	*Degrading*	*Acceptable*	*Degrading*	*Acceptable*
Noun	170,169	356,105	25	34
Verb	36,111	41,949	24	33
Adjective	22,834	51,599	26	30
Adverb	4,016	6,097	26	1
Total	233,130	455,750	101	98

TABLE 2 – Number of repetitions that degrade fluency and those which do not, found in both the *Training* and *Repetition-test* corpora. Occurrences are also displayed considering the morpho-syntactic function of repetitions.

through the WMT'2014 translation shared task [7]. In addition, we use a held-out *Repetition-test* composed of reference English and their corresponding French machine translations that feature at least one repeated word on the target (French) side for a more nuanced analysis of repetition. Machine translations were obtained with our baseline NMT model (referred in Appendix B as *baseline*). The *Repetition-test* set primarily serves to assess our models' performance in handling repetition issues, while we employ the *News-test* set to evaluate overall translation accuracy. Further details of the train and test datasets used are given in Appendix A.

We translate the English side of the *Training* and *Repetition-test* sets following the procedure detailed in Section 4 to identify repetitions which hinder fluency (*Degrading*) and those which do not (*Acceptable*). Table 2 displays the frequency of repetitions identified within the French translations. The table presents the occurrence of both types of repetitions, accompanied by an analysis of their frequency concerning the morpho-syntactic function of the repetitions [8]. Sentences are morpho-syntactically analyzed using the spaCy [9] toolkit.

Our NMT model is built using an in-house implementation of the state-of-the-art Transformer architecture (Vaswani *et al.*, 2017). Details of the network and training work are given in Appendix B.

6 Results and Analysis

To evaluate the methods presented in this paper we consider the previous *baseline* model that we update with 15K additional iterations for two different configurations of the ground-truth distribution :

$q^{\epsilon LS}$ follows the same configuration than the *baseline* model with label smoothing set to $\epsilon = 0.1$.

$q^{\epsilon LS\alpha\mathcal{M}}$ further penalizes the ground-truth distribution with repetition penalties as detailed in Section 3 with $\epsilon = 0.1$ and for different values of α.

Note that for both configurations, we use 7.6M reference sentence pairs detailed in Table A (*Training set*) together with the synthetic translations containing repetitions predicted *Degrading* and *Acceptable* of Table 2, summing up to 7.6M + 233K + 455K sentence pairs. It's essential to find a balance between the number of sentences in each training set (reference and synthetic) to uphold overall quality while teaching the model to minimize specific repetitions.

7. https://www.statmt.org/wmt14/translation-task.html
8. Only French adverbs ending with suffix **ment* are considered.
9. https://spacy.io/ with the French fr_core_news_lg model.

Configuration	*Repetition-test* BLEU	COMET	*Degrading*	*Acceptable*	*News-test* BLEU	COMET
$q^{\epsilon LS}$	45.15	37.36	99	95	32.60	26.50
$q^{\epsilon LS\alpha\mathcal{M}}, 1-\alpha = 10^{-2}$	45.54	39.10	81	89	32.45	26.05
$q^{\epsilon LS\alpha\mathcal{M}}, 1-\alpha = 10^{-3}$	45.63	40.31	79	87	32.44	26.18
$q^{\epsilon LS\alpha\mathcal{M}}, 1-\alpha = 10^{-6}$	45.65	40.13	77	86	32.50	26.30
$beam, \beta = 0$	35.03	31.28	0	0	20.66	10.65
$beam, topk = 10$	44.34	37.88	94	85	32.45	25.91
$GPT3.5$	29.70	29.59	25	43	29.98	27.60
$NLLB$	34.13	25.37	51	57	31.98	23.64

TABLE 3 – Translation accuracy results and number of repetitions present in translations performed by models under different configurations. Two different test sets are considered. ϵ is always set to 0.1.

Configuration $beam$ employs the *baseline* model and performs inference following two strategies to reduce repetitions and improve diversity :

β with a penalty applied at each inference time-step t whenever token y_t appears repeated in the hypothesis prefix $y_0, ..., y_{t-1}$. Probability p_t is reduced by factor $0 \leq \beta \leq 1$. Thus, reducing the likelihood of such hypotheses.

$topk$ Sampling predictions from the k most likely tokens of the output distribution. This is an effective decoding strategy typically used for increasing diversity when building back-translation datasets.

We also assess the effectiveness of two large language models (LLM) with translation capabilities to overcome the repetition issue :

$GPT3.5$ consists of the *GPT3.5-turbo* version of the OpenAI LLM. Built upon the Generative Pre-trained Transformer architecture (Radford & Sutskever, 2018) which employs only a transformer decoder. Following an auto-regressive approach, the model ensures that the generated text maintains coherence and relevance to the context provided by the input text. Translations are conducted using the OpenAI API, while emphasizing the importance of minimizing word repetitions through the provided prompt [10].

$NLLB$ is a family of machine translation models based on the Transformer encoder-decoder architecture, enabling translation between any of the 202 language varieties (NLLB Team *et al.*, 2022). We use the *nllb-200-distilled-600M* [11] version and perform translations with the efficient CTranslate2 [12] inference toolkit.

To evaluate the presented methods, we report BLEU and COMET results computed by `sacrebleu` [13] (Post, 2018) and `comet-score` [14] (Rei *et al.*, 2020) respectively over both test sets. Concerning *Repetition-test*, we also report the number of word repetitions that hinder fluency, *Degrading*, and those deemed acceptable, *Acceptable*, measured in translation hypotheses.

10. Prompt = *Translate the following text from English to French, ensuring that the translated output maintains coherence and fluency while minimizing the repetition of words or phrases. Pay attention to using synonyms, varied sentence structures, and appropriate linguistic devices to enhance the overall quality of the translation. Feel free to creatively adapt the language to achieve a natural and engaging tone in the target language.I want you to only reply the traduction, do not write explanations*
11. https://huggingface.co/facebook/nllb-200-distilled-600M
12. https://github.com/OpenNMT/CTranslate2
13. https://github.com/mjpost/sacrebleu
14. https://github.com/Unbabel/COMET

Models fine-tuned from the *baseline* network exhibit nearly identical quality scores across the *News-test* sets. This suggests that training with the method presented to adjust the ground-truth distribution does not compromise translation quality. On the contrary, unlike Configuration $q^{\epsilon LS}$, Configurations $q^{\epsilon LS\alpha\mathcal{M}}$ demonstrate a significant decrease in the number of repetitions that degrade fluency over the *Repetition-test*, while retaining most of the acceptable repetitions in the translated output. Note also the increase in quality over the *Repetition-test* set as measured by COMET score ($\sim 40 > 37.62$). Adjusting α does not seem to have a significant impact on reducing repetitions that degrade fluency. Decreasing its value gradually (lightly) reduces the occurrence of such repetitions. As expected, the number of acceptable repetitions remains unchanged since the training input signal with acceptable repetitions remains constant across all α values.

Regarding inference-based configurations, *beam* with $\beta = 0$ effectively eliminates all repetitions but at the expense of a notable decrease in translation quality, in the case of $topk = 10$ the number of repetitions is lightly reduced as well as global accuracy.

Results from both LLMs demonstrate a reduced number of repetitions, suggesting an elevated level of diversity and fluency of such models. However, the translation quality scores of LLMs do not align with those achieved by the models presented in this study in either of the test sets, especially translations obtained by GPT-3.5. These findings are consistent with those presented by (Bawden & Yvon, 2023) where the authors note the challenge of controlling translations performed by BLOOM [15], a multilingual LLM.

Table 4 illustrates reference translations (src and tgt) together with translations by models $q^{\epsilon LS}$ and $q^{\epsilon LS\alpha\mathcal{M}}$. The first (top) examples exhibit the ability of model $q^{\epsilon LS\alpha\mathcal{M}}$ to avoid *degrading* repetitions. The last examples contain *acceptable* repetitions hypothesized by both models.

7 Conclusions and Further Work

We have introduced a method to reduce the occurrence of repetitions in translation hypotheses, which significantly affects the readability of the generated texts. Additionally, we have proposed a straightforward approach to identify repetitions in machine translations that detract from fluency. The method is solely implemented during fine-tuning at the conclusion of the training phase, without any modifications to the inference process. Experiments indicate the ability of our proposed methods in reducing the repetition problem. Additional experiments are necessary to confirm the applicability of the proposed methods across various language pairs and dataset conditions. We aim to further study the impact of the ratio between the number of reference sentences and synthetic translations that include repetitions during the training process. Additionally, we plan to analyze the influence of the distance (measured in number of words) between repetitions and explore the possibility of replacing the binary penalty in matrix $\mathcal{M}$ with a softer approach.

Remerciements

The work presented in this paper was supported by the EU Horizon 2020 Programme for TRACE project (Grant Agreement No. 101022004).

15. https://huggingface.co/bigscience/bloom

src	(h) liabilities, including unliquidated obligations ;
tgt	h) les dettes, y compris les engagements non réglés ;
$q^{\epsilon LS}$	(h) les engagements, y compris les engagements non réglés ;
$q^{\epsilon LS\alpha\mathcal{M}}$	(h) les **passifs**, y compris les **engagements** non réglés ;
src	We talked about the tourism, hospitality and hotel sectors.
tgt	On a parlé des secteurs du tourisme, de l'hébergement et de l'hôtellerie.
$q^{\epsilon LS}$	Nous avons parlé des secteurs du tourisme, de l'hôtellerie et de l'hôtellerie.
$q^{\epsilon LS\alpha\mathcal{M}}$	Nous avons parlé des secteurs du tourisme, de l'**accueil** et de l'**hôtellerie**.
src	The home was modest and frugal.
tgt	C'était une maison modeste.
$q^{\epsilon LS}$	La maison était modeste et modeste.
$q^{\epsilon LS\alpha\mathcal{M}}$	La maison était **modeste** et **économe**.
src	Courts will tackle the question anyway, often obliquely or indirectly.
tgt	Les tribunaux vont se poser cette question de toute façon, souvent à mots couverts ou indirectement.
$q^{\epsilon LS}$	Les tribunaux aborderont la question de toute façon, souvent indirectement ou indirectement.
$q^{\epsilon LS\alpha\mathcal{M}}$	Les tribunaux aborderont la question de toute façon, souvent de façon **oblique** ou **indirecte**.
src	Technology travels fast and is swiftly adopted.
tgt	Les technologies se distribuent rapidement et sont rapidement adoptées.
$q^{\epsilon LS}$	La technologie voyage rapidement et est rapidement adoptée.
$q^{\epsilon LS\alpha\mathcal{M}}$	La technologie voyage **vite** et est **rapidement** adoptée.
src	Parliament must be able to exercise its power of scrutiny.
tgt	Le Parlement européen doit exercer son pouvoir de contrôle.
$q^{\epsilon LS}$	Le Parlement doit pouvoir exercer son pouvoir de contrôle.
$q^{\epsilon LS\alpha\mathcal{M}}$	Le Parlement doit **être en mesure d'**exercer son **pouvoir** de contrôle.
src	Interferometer apparatus and interferometric method
tgt	Appareil interféromètre et procédé interférométrique
$q^{\epsilon LS}$	Appareil interférométrique et procédé interférométrique
$q^{\epsilon LS\alpha\mathcal{M}}$	Appareil interférométrique et procédé interférométrique
src	Cardiac murmur, heart rate increased
tgt	Souffle cardiaque, augmentation de la fréquence cardiaque
$q^{\epsilon LS}$	Souffle cardiaque, fréquence cardiaque augmentée
$q^{\epsilon LS\alpha\mathcal{M}}$	Souffle cardiaque, fréquence cardiaque augmentée

TABLE 4 – Models configured with ($q^{\epsilon LS\alpha\mathcal{M}}$) and without ($q^{\epsilon LS}$) penalization exhibit varying performance when encountering repetitions (outlined using blue). Reference source (src) and target (tgt) translations are also indicated. The initial (top) examples include repetitions that *degrade* fluency, whereas the final (bottom) examples feature *acceptable* repetitions.

Références

BAWDEN R. & YVON F. (2023). Investigating the translation performance of a large multilingual language model : the case of BLOOM. In M. NURMINEN, J. BRENNER, M. KOPONEN, S. LATOMAA, M. MIKHAILOV, F. SCHIERL, T. RANASINGHE, E. VANMASSENHOVE, S. A. VIDAL, N. ARANBERRI, M. NUNZIATINI, C. P. ESCARTÍN, M. FORCADA, M. POPOVIC, C. SCARTON & H. MONIZ, Éds., *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, p. 157–170, Tampere, Finland : European Association for Machine Translation.

EDUNOV S., OTT M., AULI M. & GRANGIER D. (2018). Understanding back-translation at scale. In E. RILOFF, D. CHIANG, J. HOCKENMAIER & J. TSUJII, Éds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 489–500, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1045.

HOLTZMAN A., BUYS J., DU L., FORBES M. & CHOI Y. (2020). The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020* : OpenReview.net.

KINGMA D. P. & BA J. (2014). Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*.

LACHAUX M.-A., JOULIN A. & LAMPLE G. (2020). Target conditioning for one-to-many generation. In T. COHN, Y. HE & Y. LIU, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 2853–2862, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.256.

LI M., ROLLER S., KULIKOV I., WELLECK S., BOUREAU Y.-L., CHO K. & WESTON J. (2020). Don't say that ! making inconsistent dialogue unlikely with unlikelihood training. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 4715–4728, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.428.

LIN H., YANG B., YAO L., LIU D., ZHANG H., XIE J., ZHANG M. & SU J. (2022). Bridging the gap between training and inference : Multi-candidate optimization for diverse neural machine translation. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Findings of the Association for Computational Linguistics : NAACL 2022*, p. 2622–2632, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-naacl.200.

MÜLLER R., KORNBLITH S. & HINTON G. (2019). *When Does Label Smoothing Help ?*, In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc. : Red Hook, NY, USA.

NLLB TEAM, COSTA-JUSSÀ M. R., CROSS J., ÇELEBI O., ELBAYAD M., HEAFIELD K., HEFFERNAN K., KALBASSI E., LAM J., LICHT D., MAILLARD J., SUN A., WANG S., WENZEK G., YOUNGBLOOD A., AKULA B., BARRAULT L., MEJIA-GONZALEZ G., HANSANTI P., HOFFMAN J., JARRETT S., SADAGOPAN K. R., ROWE D., SPRUIT S., TRAN C., ANDREWS P., AYAN N. F., BHOSALE S., EDUNOV S., FAN A., GAO C., GOSWAMI V., GUZMÁN F., KOEHN P., MOURACHKO A., ROPERS C., SALEEM S., SCHWENK H. & WANG J. (2022). No language left behind : Scaling human-centered machine translation.

OCH F. J. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29**(1), 19–51.

POST M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 186–191, Belgium, Brussels : Association for Computational Linguistics.

RADFORD, ALEC N. K. S. T. & SUTSKEVER I. (2018). Improving language understanding with unsupervised learning. *Technical Report*.

REI R., STEWART C., FARINHA A. C. & LAVIE A. (2020). COMET : A neural framework for MT evaluation. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 2685–2702, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.213.

SCHIOPPA A., VILAR D., SOKOLOV A. & FILIPPOVA K. (2021). Controlling machine translation for multiple attributes with additive interventions. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 6676–6696, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.535.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Neural machine translation of rare words with subword units. In K. ERK & N. A. SMITH, Éds., *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1715–1725, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1162.

SHU R., NAKAYAMA H. & CHO K. (2019). Generating diverse translations with sentence codes. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1823–1827, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1177.

SU Y., LAN T., WANG Y., YOGATAMA D., KONG L. & COLLIER N. (2022). A contrastive framework for neural text generation. In A. H. OH, A. AGARWAL, D. BELGRAVE & K. CHO, Éds., *Advances in Neural Information Processing Systems*.

SZEGEDY C., VANHOUCKE V., IOFFE S., SHLENS J. & WOJNA Z. (2015). Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 2818–2826.

VAHTOLA T., CREUTZ M. & TIEDEMANN J. (2023). Guiding zero-shot paraphrase generation with fine-grained control tokens. In A. PALMER & J. CAMACHO-COLLADOS, Éds., *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, p. 323–337, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.starsem-1.29.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.

WELLECK S., KULIKOV I., ROLLER S., DINAN E., CHO K. & WESTON J. (2019). Neural text generation with unlikelihood training. *ArXiv*, **abs/1908.04319**.

A Corpora Statistics

Table 5 presents various statistics of the corpora used in this work, including the total number of sentences, vocabularies, words, and average sentence length. Statistics are computed after performing a light tokenization aiming to split-off punctuation.

Side	Sentences	Vocabulary	Words	Length
Training set				
English	7.6M	755K	174M	22.9
French		839K	208M	27.3
News-test				
English	16,071	27K	401K	24.1
French		32K	468K	29.1
Repetition-test				
English	199	1,323	2,521	12.6
French		1,352	3,215	16.1

TABLE 5 – Corpora statistics. M and K stand for millions and thousands respectively.

size of word embedding	512
size of hidden layers	512
size of inner feed forward layer	2,048
number of heads	8
number of layers	6
batch size	4,000 (tokens)
batch accumulation	25 (batches)

TABLE 6 – Network hyperparameters.

B NMT Network

Table 6 indicate the hyper-parameters employed to build our translation network.

For optimization work we use the lazy Adam algorithm (Kingma & Ba, 2014). We set warmup steps to 4,000 and update learning rate for every 8 iterations. All models are trained using a single NVIDIA V100 GPU.

We limit the source and target sentence lengths to 150 tokens based on BPE (Sennrich *et al.*, 2016) preprocessing in both source and target sides. We use a joint vocabulary of $32K$ tokens for both source and target sides. In inference we use a beam size of 5.

Our *baseline* English-to-French model is trained during more than 3 million iterations using all the parallel data available in the Opus website (see Appendix A).

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Régression logistique parcimonieuse pour l'extraction automatique de règles de grammaire

Santiago Herrera[1] Caio Corro[2] Sylvain Kahane[1,3]
(1) Université Paris Nanterre, CNRS, Modyco, 200 Avenue de la République, 92001, Nanterre, France
(2) Sorbonne Université, CNRS, ISIR, 4 Place Jussieu, 75005, Paris, France
(3) Institut Universitaire de France
s.herrera@parisnanterre.fr, caio.corro@isir.upmc.fr, sylvain@kahane.fr

RÉSUMÉ

Nous proposons une nouvelle approche pour extraire et explorer des motifs grammaticaux à partir de corpus arborés, dans le but de construire des règles de grammaire syntaxique. Plus précisément, nous nous intéressons à deux phénomènes linguistiques, l'accord et l'ordre des mots, en utilisant un espace de recherche étendu et en accordant une attention particulière au classement des règles. Pour cela, nous utilisons un classifieur linéaire entraîné avec une pénalisation L1 pour identifier les caractéristiques les plus saillantes. Nous associons ensuite des informations quantitatives à chaque règle. Notre méthode permet de découvrir des règles de différentes granularités, certaines connues et d'autres moins. Dans ce travail, nous nous intéressons aux règles issues d'un corpus du français.

ABSTRACT

Sparse Logistic Regression with High-order Features for Automatic Grammar Rule Extraction from Treebanks

We propose a novel approach to extract and explore significant fine-grained grammar patterns and potential syntactic grammar rules from treebanks, in order to create an easy-to-understand corpus-based grammar. More specifically, we extract descriptions and rules across different languages for two linguistic phenomena, agreement and word order, using a large search space and paying special attention to the ranking order of the extracted rules. For that, we use a sparse logistic regression classifier to extract the most salient features that predict the linguistic phenomena under study. We associate statistical information to each rule. Our method discovers both known and significant grammar rules that are less well known. In this paper, we focus on rules extracted from a French treebank.

MOTS-CLÉS : Extraction de grammaire, règles de grammaire, grammaire fondée sur des corpus, grammaire quantitative, régression logistique, pénalité L1.

KEYWORDS: Grammar extraction, grammar rules, corpus based grammar, quantitative grammar, sparse logistic, L1 regularization.

1 Introduction

Les grammaires descriptives sont des ressources précieuses pour la recherche linguistique, mais leur construction est longue et difficile. Les collections de corpus arborés comme Universal Dependencies

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 211–218.

(De Marneffe *et al.*, 2021) ont permis le développement de méthodes d'extraction automatique de grammaires à partir de corpus. Dans ce travail, nous proposons une nouvelle approche fondée sur la régression logistique parcimonieuse. Nos contributions peuvent se résumer de la façon suivante :

— nous proposons une nouvelle formalisation des règles grammaticales visant l'extraction automatique à partir de corpus arborés ;
— nous étudions l'utilisation de la régression logistique L1 pour extraire et classer les règles ;
— par rapport aux travaux précédents (notamment Chaudhary *et al.* 2020, 2022), nous augmentons l'expressivité de nos règles, ce qui permet d'obtenir des règles plus fines et d'avoir aussi un outil plus adapté à la fouille de règles

Nous nous évaluons sur le français, l'espagnol et wolof pour trois phénomènes : les accords en genre, en nombre et l'ordre des mots. Cependant, nous ne reportons que les résultats sur la position du sujet en français dans ce résumé. [1]

2 Règle de grammaire

Une grammaire est un ensemble de contraintes régulières qu'une langue impose à ses locuteurs, contraintes que nous appelons souvent des règles. Une règle de grammaire décrit le contexte linguistique particulier dans lequel un motif grammatical est privilégié ou non, par rapport à d'autres motifs possibles. Dans la pratique, ces règles sont de nature probabiliste (dans l'interprétation fréquentiste), peuvent être plus ou moins fines et peuvent se chevaucher.

Par exemple, une règle simple du français est : « *le sujet nominal d'un verbe se place avant son gouverneur* ». Cependant, cette règle n'est pas déterministe. Une description plus précise serait : « *pour une dépendance syntaxique de type* sujet*, le dépendant se place avant son gouverneur dans plus de* 97% *des cas* » [2]. En pratique, il est intéressant de comprendre dans quels cas il y a une divergence avec la règle dominante, par exemple : « *le sujet nominal dans une subordonnée relative se place après son gouverneur dans* $\approx 23\,\%$ *des cas* » (voir la Figure 1).

Nous proposons de définir une règle de grammaire comme une construction composée de trois motifs :

— S est un motif donné qui définit la portée (angl. *scope*) de la règle, c'est-à-dire quelles sont les constructions qui nous intéressent (p. ex. nous pouvons nous focaliser sur les relations de type *sujet* uniquement).
— Q identifie le phénomène ou la question linguistique qui nous intéresse (p. ex. l'inversion du sujet).
— P est un motif prédicteur ou déclencheur de Q dans les limites de S (p. ex. le fait que le sujet soit dans une subordonnée relative).

Nous formalisons donc une règle de grammaire par une formule de la forme suivante :

$$\mathrm{S} \implies (\mathrm{P} \overset{\alpha\,\%}{\Longrightarrow} \mathrm{Q})\,.$$

où α est la fréquence avec laquelle le motif P déclenche le motif Q au sein de S. Une reformulation naturelle de notre exemple serait « parmi tous les sujets, ceux qui se trouvent dans une proposition

1. Une version en anglais de ce travail a été accepté à la conférence LREC-Coling 2024 : https://arxiv.org/abs/2403.17534.
2. Distribution tirée de la version SUD du corpus arborés GSD du français (Guillaume *et al.*, 2019).

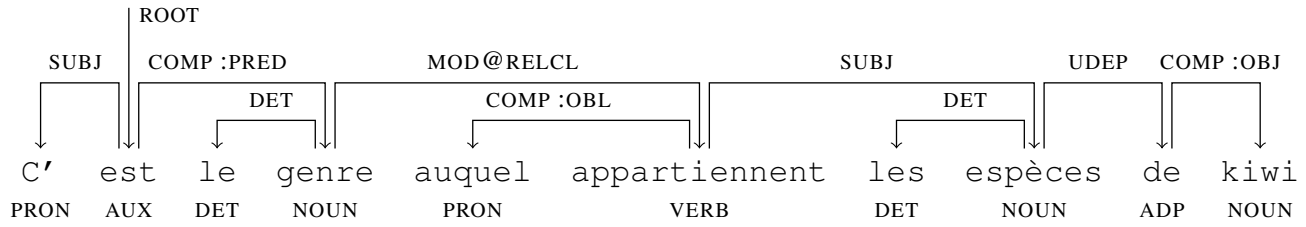


FIGURE 1 – Exemple tiré de la version SUD du corpus arborés GSD du français. Deux propositions sont reliées par une relation MOD@RELCL avec des sujets dans des positions différentes. Dans la première (resp. seconde) proposition, le sujet occupe une position préverbale (resp. post-verbale).

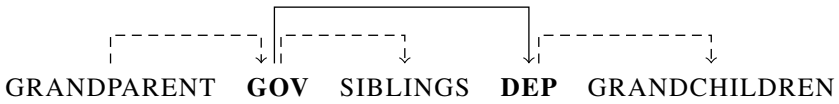


FIGURE 2 – L'espace de recherche est définit autour d'une relation entre un nœud gouverneur (GOV), et son dépendant (DEP), le gouverneur de GOV (GRANDPARENT) et les autres dépendants de GOV (SIBLINGS), ainsi que les dépendants de DEP (GRANDCHILDREN). On utilise presque toutes les informations linguistiques encodées dans ces nœuds, y compris les traits morphosyntaxiques, certains lemmes et les relations de dépendances, ainsi que les positions relatives de ces nœuds (qui peuvent être différentes de celles de la figure).

relative sont inversés dans plus de 23% des cas ».

Cette formalisation s'inspire des règles (de correspondance) de la théorie Sens-Texte (Mel'čuk, 1988) [3]. Cependant, nous nous concentrons uniquement sur les relations au niveau de la syntaxe de surface. De plus, cette formalisation se traduit facilement en une tâche de classification par apprentissage automatique.

En pratique, les motifs S et Q sont déterminés manuellement, car ils définissent les phénomènes linguistiques auxquels nous allons nous intéresser. Cependant, le nombre potentiel de motifs P est très important et il est donc difficile de les explorer manuellement. En particulier, nous utilisons presque toutes les informations linguistiques encodées dans les corpus arborés : pour toutes les dépendances filtrées par S, notre espace de recherche pour P considère toute information relative au grand-parent (gouverneur du gouverneur), aux codépendants du gouverneur et aux enfants du dépendant (voir Figure 2).

3 Méthode d'extraction

Plusieurs études ont démontré la nécessité d'avoir une approche quantitative pour étudier des phénomènes syntaxiques (p. ex. Bresnan *et al.*, 2004; Thuilier, 2012). Aujourd'hui, il est possible d'extraire automatiquement des règles syntaxiques en utilisant des corpus annotés en syntaxe comme *Universal Dependencies* (UD, De Marneffe *et al.*, 2021) ou *Surface Syntactic UD* (SUD, Gerdes *et al.*, 2018). Les travaux de Chaudhary *et al.* (2020, 2022) proposent d'utiliser des arbres de décision,

3. Dans la notation de Mel'čuk (1988), la règle s'écrirait « S $\Longrightarrow$ Q | P » et se lirait « S est susceptible de correspondre à Q dans le contexte P » (le sujet peut être placé après le verbe quand il est dans une relative).

comme technique d'apprentissage automatique, à cette fin. Nous nous différencions de ces travaux en proposant une nouvelle formalisation du problème (section précédente) et car nous utilisons la régression logistique parcimonieuse, qui est plus sensible à des changements même peu prononcés de la distribution. Cela nous permet de simultanément extraire et classer les règles, tout en explorant un large espace de recherche. Nous obtenons ainsi des règles de grammaire quantitative plus expressives et fines. De plus, notre approche ne nécessite pas de régler des hyperparamètres d'apprentissage.

3.1 Régression logistique et pénalisation L1

Dans cette section, nous décrivons brièvement notre méthode d'extraction des motifs P fondée sur la régression logistique parcimonieuse. Nous renvoyons les lecteurices vers Shalev-Shwartz & Ben-David (2014, Sec. 9 & 25) et Bach *et al.* (2012) pour une introduction détaillée à ce sujet.

Modèle de classification. Soit F l'ensemble des motifs possibles pour P, qui correspondent aux caractéristiques d'entrée de notre classifieur. Nous considérons comme caractéristiques tous les motifs composés d'un ou de deux attributs dans l'espace de recherche de la Figure 2, par exemple GRANDCHILDREN.UPOS=ADP et GOV.REL_SYNT=MOD,SIBLINGS.UPOS=ADP. Soit $\boldsymbol{x} \in \{0,1\}^F$ un vecteur booléen qui indique quelles sont les caractéristiques présentent dans l'entrée. Notre modèle de classification définie la probabilité d'observer un certain motif Q, par ex. un sujet inversé, de la façon suivante :

$$\text{P("inversion du sujet"}|\boldsymbol{x}) = \sigma\left(\boldsymbol{a}^\top \boldsymbol{x} + b\right) \tag{1}$$

où $\boldsymbol{a} \in \mathbb{R}^F$ et $b \in \mathbb{R}$ sont les paramètres du modèle, et $\sigma(w) = \frac{\exp(w)}{1+\exp(w)}$ est la fonction sigmoïde. Par construction du modèle, les caractéristiques qui ont un poids 0 dans le vecteur $\boldsymbol{a}$ ne contribuent pas à la distribution du modèle.

Nous apprenons les paramètres du modèle en minimisant l'objectif suivant :

$$\min_{\boldsymbol{a}\in\mathbb{R}^F, b\in\mathbb{R}} \quad \frac{1}{|D|} \sum_{(\boldsymbol{x},y)\in D} \ell\left(\boldsymbol{a}^\top \boldsymbol{x} + b, y\right) \quad + \quad \lambda r(\boldsymbol{a})\,, \tag{2}$$

où D est le corpus d'entraînement filtré par S, ℓ est la fonction de perte, r est une pénalité de régularisation de poids $\lambda \geq 0$. En particulier, nous utilisons :

- $\ell(w, y) = -yw + \log(1 + \exp w)$, l'opposé de la log-vraisemblance ;
- $r(\boldsymbol{a}) = \|\boldsymbol{a}\|_1 = \sum_{f\in F} |a_f|$, la norme L1 qui favorise les solutions contenant des 0 dans le vecteur $\boldsymbol{a}$ (Bach *et al.*, 2012).

Nous utilisons la bibliothèque SKGLM (Bertrand *et al.*, 2022) pour apprendre les paramètres.

Chemin de régularisation. Plus le paramètre λ est élevé, plus le nombre de 0 dans le vecteur $\boldsymbol{a}$ sera élevé. Nous entraînons donc une séquence de modèles en faisant varier le poids de la régularisation L1 afin d'identifier les motifs P du plus important au moins important. Nous initialisons λ à une valeur assez grande pour que toutes les valeurs dans $\boldsymbol{a}$ soit nulles. Ensuite, nous entraînons la séquence de modèles en diminuant petit à petit la valeur de λ. Plus un motif apparaît tôt avec un score non nul, plus il est considéré comme important. La séquence de solutions obtenues en diminuant la valeur de λ est appelée le chemin de régularisation (Markowitz, 1952; Osborne *et al.*, 2000; Efron *et al.*, 2004).

3.2 Filtrage et analyse des règles

Bien que le chemin de régularisation nous donne un classement des motifs P, les coefficients de $\boldsymbol{a}$ ne constituent pas en soi des facteurs explicatifs, contrairement aux idées reçues (Achen, 2005). C'est pour cela que, dans un premier temps, pour chaque motif extrait, nous vérifions s'il est déclencheur du motif Q ou ¬Q, autrement dit, si le motif fait pencher la distribution d'un côté ou de l'autre.

Dans un deuxième temps, nous calculons la précision et la couverture/rappel de la règle, et nous appliquons un test statistique pour mieux analyser les règles. La précision et la couverture de la règle sont particulièrement importantes, car nous faisons l'hypothèse qu'une règle de grammaire fiable couvre largement le phénomène linguistique d'intérêt (ex. l'inversion du sujet) et qu'elle comporte peu de faux positifs.

Règle	**Couverture**	**Précision**
$S \wedge P \overset{\alpha\,\%}{\Longrightarrow} Q$	$\frac{\#(S \wedge P \wedge Q)}{\#(S \wedge Q)}$	$\frac{\#(S \wedge P \wedge Q)}{\#(S \wedge P)}$
$S \wedge P \overset{\alpha\,\%}{\Longrightarrow} \neg Q$	$\frac{\#(S \wedge P \wedge \neg Q)}{\#(S \wedge \neg Q)}$	$\frac{\#(S \wedge P \wedge \neg Q)}{\#(S \wedge P)}$

L'utilisation de tests et de mesures statistiques permet de pondérer les règles selon différentes mesures et ainsi d'avoir une grammaire quantitative fondée sur les occurrences d'un corpus. Il convient de noter que ces mesures doivent être contextualisées. Une règle comme celle de l'inversion du sujet dans une relative n'atteindra jamais une couverture de $100\,\%$ (et sera probablement bien inférieure).

4 Résultats pour l'inversion du sujet

Toutes les expériences utilisent les corpus arborés SUD dans sa version 2.13. Notre méthode étant indépendante des langues étudiées, nous avons extrait des règles d'accord et d'ordre pour plusieurs langues ayant un treebank en dépendance. Dans ce résumé, nous nous limitons à exposer quelques règles positives pour l'ordre du sujet en français (Table 1) qui servent à évaluer nos résultats. Notre méthode peut contribuer à la recherche de règles pour des langues peu décrites, mais aussi pour des langues bien décrites où certaines généralisations ou certains cas particuliers ont pu nous échapper. Le code pour reproduire les expériences est disponible en ligne [4] ainsi qu'un outil permettant de visualiser les résultats sur différents corpus [5].

La règle la plus saillante pour l'inversion du sujet, ainsi que la deuxième et quatrième, indiquent que les sujets dans les incises (PARATAXIS:INSERT), dans une structure de discours rapporté (ex : "La route sera longue, prévient le représentant du pape."), sont toujours inversés. Bien entendu, ces règles montrent que l'extraction dépend de ce qui a été annoté : le treebank du français utilisé mentionne explicitement les propositions incises, en les distinguant des autres cas de parataxe.

La cinquième règle capture l'exemple privilégié dans ce résumé : être dans une relative (GOV.DEEP_REL=MOD@RELCL) déclenche à un certain degré l'inversion du sujet. On sait qu'il s'agit d'un sujet nominal car il est déterminé (GRANDCHILDREN.REL_SYNT=DET). Avoir capturé

4. https://github.com/FilippoC/grex-lrec-coling-2024
5. https://autogramm.github.io/grex-lrec-coling-2024

P de la règle	λ	couverture	précision
gov.rel_synt=parataxis:insert,grandparent.position=before_gov	0,009	24,9	100
gov.rel_synt=parataxis:insert	0,007	24,9	100
grandchildren.rels_synt=det,grandparent.position=before_gov	0,005	34,1	9,7
gov.VerbForm=Fin,gov.rel_synt=parataxis:insert	0,005	24,9	100
gov.rel_deep=mod@relcl,grandchildren.rels_synt=det	0,004	16,8	23,7
gov.VerbForm=Fin,siblings.upos=PRON	0,004	26,3	5,4
dep.rel_deep=subj@expl,gov.Person=3	0,003	13,2	10,4
dep.Person=3,dep.rel_deep=subj@expl	0,003	13,2	10,4
gov.VerbForm=Fin,grandparent.Number=Sing	0,003	38,7	7,4
grandchildren.rels_synt=det,siblings.upos=PRON	0,003	17,8	9,5

TABLE 1 – Les dix règles d'ordre des mots les plus saillantes qui favorisent l'inversion du sujet par rapport à son gouverneur, extraites de la version 2.13 du treebank SUD GSD du français et classées en fonction de l'ordre donné par le classifieur linéaire. Bien que les règles négatives (qui prédisent l'ordre canonique sujet-verbe) sont intéressantes, nous incluons seulement les règles positives. Voir la section 3.1 pour l'interprétation de λ. La couverture et la précision sont exprimées en pourcentage.

cette règle représente un bon test de l'expressivité de notre méthode étant donné sa faible précision (23,7%) et couverture (16,8%).

La troisième règle, que nous avions préalablement sautée, englobe celles déjà mentionnées. Elle comprend un motif P très général signalant une tendance syntaxique : le sujet déterminé se place plus souvent après son gouverneur par rapport à la distribution positionelle de base dans des propositions subordonnées. La subordination est ici encodée par l'existence d'un gouverneur du verbe (grandparent.position=before_gov).

La sixième et la dixième règle correspondent au cas des sujets où le verbe possède un dépendant pronominal. Il s'agit en fait d'une situation où prédomine les relatives et les interrogatives (ex : "Que dit l'Église sur ce point délicat ?"). Les règles présentent néanmoins une formulation inhabituelle et inattendue.

La septième règle, comme la huitième, capture le cas des sujets explétifs (SUBJ@EXPL), qui se trouvent être effectivement plus souvent post-verbaux dans les interrogatives (ex. "Ces chiffres sont-ils élevés ?" ; "Qu'est-ce qui va augmenter ?).

On aura noté qu'on obtient souvent des règles similaires déclenchées par des motifs qui se recoupent. Autrement dit, nous obtenons une hiérarchie de résultats plus ou moins fins. Il serait possible de repérer de tels motifs et de les filtrer, mais il n'est pas forcément évident de décider automatiquement qu'elle est la formulation la plus naturelle pour l'utilisateur et une étude plus approfondie est nécessaire. À notre connaissance, les règles détaillées n'ont pas été capturées par des travaux précédents (Chaudhary *et al.*, 2022).

Outre l'intérêt descriptif, les règles extraites ont aussi un intérêt comparatif. Il serait envisageable de réaliser des études contrastives entre plusieurs langues ou entre plusieurs corpus pour repérer les motifs saillants d'une langue ou d'un corpus par rapport à d'autres. La comparaison de motifs entre plusieurs langues ou familles de langues permettrait aussi de faire de la typologie quantitative avec un niveau de description plus fin que d'habitude. Par exemple, il rendrait possible non seulement de comparer l'ordre des constituants majeurs, mais aussi de comparer la distribution complète des sujets

dans une famille de langue. Ces expériences prometteuses requièrent des études spécifiques et sont hors portée de ce travail.

5 Limitations

Les règles extraites dépendent du schéma d'annotation, de la nature du corpus et de sa qualité. Ainsi, quelques motifs extraits expriment des propriétés du corpus ou des décisions théoriques, plutôt que des règles de grammaire à proprement parler. Nous sommes aussi limités par ce qui est effectivement annoté et nos règles sont donc circonscrites à la syntaxe de surface, c'est-à-dire aux règles de bonne formation de l'arbre syntaxique (règles d'accord) et aux règles de correspondance entre l'arbre et la chaîne linéaire (règle d'ordre des mots).

En outre, certains résultats observés montrent des préférences d'usage de la langue, et non des règles de grammaire. Par exemple, on note une tendance à l'accord du verbe avec son objet, du fait que les objets singuliers sont significativement appariés à des sujets singuliers (de même pour les pluriels). À l'inverse, la règle d'accord d'un participe passé avec son objet lorsque celui-ci le précède n'a pas été retrouvée, probablement parce que le pronom relatif *que* ne comporte pas de trait de nombre et de genre dans l'annotation (ce qui est par ailleurs un choix d'annotation tout à fait raisonnable si on s'en tient à l'annotation morphosyntaxique).

6 Conclusion

Nous proposons une nouvelle méthode d'extraction de règles grammaticales à partir de corpus arborés. Notre approche est fondée sur (1) une définition formelle de ce qu'est une règle grammaticale syntaxique et (2) l'utilisation d'un modèle linéaire de classification pour extraire et classer les règles via le chemin de régularisation. De plus, nous avons réussi à montrer avec notre analyse qu'il est important d'étendre l'espace de recherche pour augmenter l'expressivité des règles et pour capturer des phénomènes linguistiques qui sont hors de portée des travaux précédents (Chaudhary *et al.*, 2020, 2022; Blache *et al.*, 2016). Notre méthode est également mieux adaptée à la fouille de règles.

Nous espérons aussi, avec ce travail, contribuer à l'élaboration de grammaires descriptives et aussi de contribuer au rapprochement entre le TAL ou la Linguistique Computationnelle (CL) et la linguistique théorique, ainsi qu'entre le TAL/CL et la linguistique de terrain.

Références

ACHEN C. H. (2005). Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science*, **22**(4), 327–339.

BACH F., JENATTON R., MAIRAL J., OBOZINSKI G. *et al.* (2012). Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, **4**(1), 1–106.

BERTRAND Q., KLOPFENSTEIN Q., BANNIER P.-A., GIDEL G. & MASSIAS M. (2022). Beyond l1 : Faster and better sparse models with skglm. In *NeurIPS*.

BLACHE P., RAUZY S. & MONTCHEUIL G. (2016). MarsaGram : an excursion in the forests of parsing trees. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 2336–2342, Portorož, Slovenia : European Language Resources Association (ELRA).

BRESNAN J., CUENI A., NIKITINA T. & BAAYEN R. (2004). Predicting the dative alternation. *Cognitive foundations of interpretation*.

CHAUDHARY A., ANASTASOPOULOS A., PRATAPA A., MORTENSEN D. R., SHEIKH Z., TSVETKOV Y. & NEUBIG G. (2020). Automatic extraction of rules governing morphological agreement. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 5212–5236, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.422.

CHAUDHARY A., SHEIKH Z., MORTENSEN D. R., ANASTASOPOULOS A. & NEUBIG G. (2022). Autolex : An automatic framework for linguistic exploration.

DE MARNEFFE M.-C., MANNING C. D., NIVRE J. & ZEMAN D. (2021). Universal dependencies. *Computational linguistics*, **47**(2), 255–308.

EFRON B., HASTIE T., JOHNSTONE I. & TIBSHIRANI R. (2004). Least angle regression. *Annals of Statistics*.

GERDES K., GUILLAUME B., KAHANE S. & PERRIER G. (2018). Sud or surface-syntactic universal dependencies : An annotation scheme near-isomorphic to ud. In *Universal Dependencies Workshop (UDW)*.

GUILLAUME B., DE MARNEFFE M.-C. & PERRIER G. (2019). Conversion et améliorations de corpus du français annotés en Universal Dependencies. *Revue TAL : traitement automatique des langues*, **60**(2), 71–95. HAL : hal-02267418.

MARKOWITZ H. (1952). Portfolio selection. *Journal of Finance*.

MEL'ČUK I. (1988). *Dependency Syntax : Theory and Practice*. State University of New York Press.

OSBORNE M. R., PRESNELL B. & TURLACH B. A. (2000). On the lasso and its dual. *Journal of Computational and Graphical statistics*, **9**(2), 319–337.

SHALEV-SHWARTZ S. & BEN-DAVID S. (2014). *Understanding machine learning : From theory to algorithms*. Cambridge university press.

THUILIER J. (2012). *Contraintes préférentielles et ordres des mots en français*. Thèse de doctorat, Paris 7. Dirigée par Laurence Danlos Laurence et Benoît Crabbé. Linguistique théorique, descriptive et automatique.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

SEC : contexte émotionnel phrastique intégré pour la reconnaissance émotionnelle efficiente dans la conversation

Barbara Gendron[1, 2] Gaël Guibon[1]
(1) LORIA, Université de Lorraine, CNRS
(2) Université du Luxembourg
prénom.nom@loria.fr

RÉSUMÉ

L'essor des modèles d'apprentissage profond a apporté une contribution significative à la reconnaissance des émotions dans les conversations (ERC). Cependant, cette tâche reste un défi important en raison de la pluralité et de la subjectivité des émotions humaines. Les travaux antérieurs sur l'ERC fournissent des modèles prédictifs utilisant principalement des représentations de la conversation basées sur des graphes. Dans ce travail, nous proposons une façon de modéliser le contexte conversationnel que nous incorporons à une stratégie d'apprentissage de métrique, avec un processus en deux étapes. Cela permet d'effectuer l'ERC dans un scénario de classification flexible et d'obtenir un modèle léger et efficace. En utilisant l'apprentissage de métrique à travers une architecture de réseau siamois, nous obtenons un score de macroF1 de 57,71% pour la classification des émotions dans les conversations sur le jeu de données DailyDialog, ce qui surpasse les travaux connexes. Ce résultat état-de-l'art est prometteur en ce qui concerne l'utilisation de l'apprentissage de métrique pour la reconnaissance des émotions, mais est perfectible au regard du microF1 obtenu.

ABSTRACT

Context-Aware Metric Learning for Efficient Emotion Recognition in Conversation

The advent of deep learning models has made a considerable contribution to the achievement of Emotion Recognition in Conversation (ERC). However, this task still remains an important challenge due to the plurality and subjectivity of human emotions. Previous work on ERC provides predictive models using mostly graph-based conversation representations. In this work, we propose a way to model the conversational context that we incorporate into a metric learning training strategy, with a two-step process. This allows to perform ERC in a flexible classification scenario and to end up with a lightweight yet efficient model. Using metric-learning through a Siamese Network architecture, we achieve 57.71% in macroF1 score for emotion classification in conversation on DailyDialog dataset, which outperforms the related work. This state-of-the-art result is promising regarding the use of metric-learning for emotion recognition, yet perfectible compared to the microF1 score obtained.

MOTS-CLÉS : apprentissage profond, reconnaissance d'émotions en conversation, apprentissage de métrique.

KEYWORDS: deep learning, emotion recognition in conversation, metric learning.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 219–233.

1 Introduction

La communication médiée par ordinateur (CMO) est en constante évolution et de nouveaux moyens de communication apparaissent régulièrement. Avec l'avènement des agents conversationnels, il devient nécessaire de détecter les émotions au sein d'une conversation. Bien que plusieurs modalités soient désormais prises en compte dans le processus de communication, la modalité textuelle reste essentielle pour une communication quotidienne rapide et facile, par le biais d'outils comme les applications de messagerie ou les médias sociaux. La modalité textuelle est toutefois ambiguë, car elle ne préserve pas le contexte extra-linguistique présent par exemple dans les conversations dyadiques. L'une des principales ambiguïtés est l'état émotionnel de l'orateur, souvent mal interprété par les humains à travers des messages courts et non polis. Cela motive la reconnaissance d'émotions en conversation (*Emotion Recognition in Conversation*, ERC) qui vise non seulement à l'identification des émotions dans les messages, mais aussi à la prise en compte du contexte conversationnel pour reconnaître les émotions. L'ERC s'est révélée être un défi, notamment en ce qui concerne la représentation du contexte (Ghosal *et al.*, 2021). Récemment, les modèles multimodaux et les approches basées sur les graphes se sont multipliés. Ils représentent souvent le contexte conversationnel à travers un profil des locuteurs, ce qui est performant mais moins efficace, en plus de dépendre des étiquettes émotionnelles. Les approches existantes sont principalement supervisées et confrontées à un fort déséquilibre des étiquettes en raison de la rareté de certaines émotions.

Dans cet article, nous adressons ces deux défis en incorporant le contexte conversationnel dans un scénario d'apprentissage de métrique (que l'on désignera par *metric learning*), tout en contrôlant le déséquilibre des données de plusieurs façons. Dans notre cas, afin de rendre notre modèle utilisable pour d'autres émotions que les 6 émotions primaires (Ekman *et al.*, 1969), nous n'utilisons pas l'apprentissage contrastif supervisé (Khosla *et al.*, 2020) dans notre méthode. L'apprentissage de métrique permet justement de s'abstraire d'une dépendance aux définitions strictes des émotions, ce qui se révèle indispensable pour des émotions fines. Pour cela, nous mettons à jour le modèle en utilisant à la fois les prédictions d'étiquettes isolées (fonction de perte d'entropie croisée), et l'attribution d'étiquettes contextuelles relatives (fonction de perte contrastive). Ce processus en deux étapes est assez simple pour des énoncés isolés. Cependant, à notre connaissance, la représentation contextuelle par apprentissage contrastif pour l'ERC n'a pas encore été utilisée. Ceci représente notre principale contribution puisque nous présentons un modèle qui peut atteindre des performances compétitives par rapport à l'état-de-l'art tout en étant utilisable avec de nouvelles étiquettes d'émotions. Ainsi, notre modèle peut être appliqué et adapté dans de multiples contextes nécessitant la reconnaissance d'émotions à différents niveaux de granularité.

Notre principale contribution réside dans le développement d'une stratégie de *metric learning* pour la reconnaissance d'émotions en utilisant le contexte conversationnel. Le modèle présenté exploite des plongements lexicaux à l'échelle de la phrase et déploie de l'attention à l'aide d'un Transformer (Vaswani *et al.*, 2017; Devlin *et al.*, 2019) pour obtenir une représentation contextuelle de chaque tour de parole (que nous désignerons également par "énoncés" dans ce qui suit). Nous utilisons ici des réseaux siamois (Koch *et al.*, 2015) mais l'approche peut être adaptée à n'importe quel modèle de *metric learning*. Nous démontrons en outre que notre approche est plus performante que certains des derniers grands modèles de langage (*Large Language Models*, LLMs) tels que les versions allégées de Falcon (Penedo *et al.*, 2023) ou LLaMA 2 (Touvron *et al.*, 2023). En outre, notre méthode est efficace dans le sens où elle implique des modèles légers, adaptables et rapidement entraînables, qui donnent des scores état-de-l'art sur DailyDialog en macroF1 avec 57.71% et des résultats satisfaisants en microF1 avec 57.75%.

Dans les sections suivantes, nous passons d'abord en revue les travaux relatifs à l'ERC (Section 2). Nous présentons ensuite notre méthodologie (Section 3) et décrivons le dispositif expérimental que nous utilisons (Section 4). Nous évaluons ensuite nos modèles par rapport à une référence sans contexte conversationnel et aux modèles état-de-l'art pour l'ERC dans la section 5. Enfin, nous exposons nos principales conclusions ainsi que des perspectives pour les travaux futurs dans la section 6. Nous mettrons à disposition notre code et nos modèles sur *GitHub* et *HuggingFace models*.

2 État de l'art

ERC. Bien que la plupart des travaux en ERC tirent parti de la multimodalité (Song *et al.*, 2022; Li *et al.*, 2022; Hu *et al.*, 2022), certains modèles ont été développés pour l'ERC sur des conversations textuelles uniquement, que ce soit des données multimodales limités au texte tels que IEMOCAP (Busso *et al.*, 2008) ou MELD (Poria *et al.*, 2019), ou sur des données uniquement textuelles comme dans DailyDialog (Li *et al.*, 2017). L'apprentissage profond permet des progrès significatifs en ERC sur le texte, en commençant par l'utilisation de réseaux récurrents (RNN) (Rumelhart *et al.*, 1985; Jordan, 1986) par Poria *et al.* (2017). D'autres travaux utilisant des structures récurrentes ont suivi, comme DialogueRNN (Majumder *et al.*, 2019; Ghosal *et al.*, 2020). Ce modèle tire parti du mécanisme d'attention (Bahdanau *et al.*, 2014) implémenté dans le Transformer (Vaswani *et al.*, 2017). Les méthodes basées sur les graphes sont également efficaces comme le montre (Ghosal *et al.*, 2019), non seulement en tant que telles, mais aussi en incluant des connaissances externes (Lee & Choi, 2021).

Les travaux existants en ERC s'appuient principalement sur l'évaluation de leur modèle en microF1 en excluant l'étiquette neutre (pas d'émotion), souvent majoritaire. Cependant, des travaux récents se passent de cette évaluation pour se concentrer uniquement sur la macroF1 (Pereira *et al.*, 2023), tandis que d'autres ont considéré le coefficient de corrélation de Matthews comme une métrique adaptée à cette tâche (Guibon *et al.*, 2021). Dans ce travail, nous nous concentrons sur DailyDialog, qui consiste en des conversations générées artificiellement par l'Homme sur les préoccupations de la vie quotidienne, avec un étiquetage des émotions au niveau du tour de parole. Liang *et al.* (2022) proposent un modèle basé sur un réseau neuronal de graphes (*Graph Neural Network*, GNN) et un champ aléatoire conditionnel (*Conditional Random Field*, CRF) qui atteint 64,01% en microF1.

Bien qu'il soit connu pour ne pas fournir les meilleures performances par rapport aux approches d'apprentissage frugal (Dumoulin *et al.*, 2021), le *metric learning* permet une meilleure généralisation grâce à un entraînement plus robuste (Finn *et al.*, 2017; Antoniou *et al.*, 2019), ce qui est particulièrement adapté à la détection d'émotions humaines complexes et variées (Plutchik, 2001).

Metric learning. Hospedales *et al.* (2022) expliquent que le méta-apprentissage consiste en un *méta-optimiseur* qui décrit les mises à jour du méta-apprenant, une *méta-représentation* qui stocke les connaissances acquises et un *méta-objectif* orienté vers la tâche souhaitée. Cette configuration basée sur l'optimisation fournit des algorithmes complets souvent basés sur des scénarios épisodiques (Ravi & Larochelle, 2016; Finn *et al.*, 2017; Mishra *et al.*, 2017) qui reflètent l'idée d'"apprendre à apprendre". Mais cela implique des calculs de gradient au second ordre, ce qui est coûteux. Des solutions palliatives comme la différenciation implicite (Lorraine *et al.*, 2020) impliquent toujours un compromis entre performance et coût mémoire (Hospedales *et al.*, 2022). C'est pourquoi des variantes sont apparues, telles que le *metric learning*, dont le méta-objectif est l'apprentissage de la méta-représentation elle-même. Par exemple, les réseaux siamois (Koch *et al.*, 2015) tirent parti du partage des paramètres entre des sous-réseaux identiques pour apprendre une distance entre les

données. Les réseaux de relations (*Relation Networks*, Sung *et al.* (2018)) considèrent également une métrique de distance, s'écartant de l'approche euclidienne. Les réseaux de correspondance (*Matching Networks*, Vinyals *et al.* (2016)) exploitent des exemples d'apprentissage pour identifier les plus proches voisins pondérés. Les réseaux prototypiques (*Prototypical Networks*, Snell *et al.* (2017)) calculent les représentations moyennes des classes et les comparent avec la similarité cosinus. Son adaptation pour l'ERC en apprentissage frugal a commencé à partir de Guibon *et al.* (2021).

Dans ce travail, nous utilisons un réseau siamois. Ce modèle à l'architecture simple est facilement contrôlable et évolutif, mais on pourrait tout à fait adapter l'approche à des structures plus complexes. Les réseaux siamois ont été utilisés en TAL pour la détection d'intentions dans le texte (Ren & Xue, 2020), en vision par ordinateur pour la reconnaissance faciale (Hayale *et al.*, 2023) et dans l'apprentissage de représentations complexes (Jin *et al.*, 2021).

3 Méthodologie

Nous utilisons le *metric learning* pour l'apprentissage relatif des émotions, ce qui permet d'extraire des méta-informations des données. Notre réseau siamois comporte trois sous-réseaux identiques, dont les sorties sont comparées à l'aide de la fonction de coût par triplet (Schultz & Joachims, 2003), dénommée ci-après *triplet loss*. Initialement appliquée aux problèmes de vision par ordinateur (Chechik *et al.*, 2010; Schroff *et al.*, 2015), la *triplet loss* est définie sur un triplet d'échantillons de données (a, p, n) de sorte que si a et p appartiennent à la même classe et n à une classe différente, alors :

$$\mathcal{L}(a, p, n) = \max\{d(a, p) - d(a, n) + \texttt{marge}, 0\}$$

où le paramètre `marge` est un nombre strictement positif.

En considérant la *triplet loss*, plusieurs stratégies s'offrent à nous : récupérer les triplets les plus difficiles, lorsque le positif est loin de l'ancre, tandis que l'ancre est proche du négatif ; ou encore ignorer les triplets les plus faciles, c'est-à-dire lorsque le positif est plus proche de l'ancre. Compte tenu de la taille limitée de nos données, nous abordons la stratégie globale en considérant chaque triplet. Bien que la *triplet loss* puisse être utilisée dans plusieurs stratégies, nous n'abordons la stratégie globale qu'en considérant chaque triplet dans nos données, en raison de leur taille limitée.

Représentations isolées. L'objectif de nos expériences étant de caractériser la contribution du contexte conversationnel à la prédiction des émotions dans le cadre d'un apprentissage contrastif, nous avons développé, dans un premier temps, un modèle sur des énoncés isolés. Il s'agit formellement de prédire l'émotion des énoncés indépendamment de leur contexte. Pour ce faire, nous considérons d'abord une projection pour chaque mot de l'énoncé vers sa représentation FastText associée (Bojanowski *et al.*, 2017). À partir de ces plongements lexicaux, les triplets (a, p, n) susmentionnés sont échantillonnés aléatoirement et donnés en entrée au réseau siamois, dont le sous-réseau s'améliore dans la prédiction des émotions au fur et à mesure de la rétro-propagation de la *triplet loss*.

Représentations contextuelles. Dans le cas contextuel, nous construisons des représentations d'énoncés contextuels à partir d'un encodage de type BERT (Devlin *et al.*, 2019). Les plongements lexicaux utilisés sont à l'échelle de la phrase et non du mot car ils fournissent des représentations d'énoncés plus légères. Une fois que le dialogue est représenté avec la série de plongements pré-entraînés qui lui est associée, ces sorties sont concaténées pour former une représentation du dialogue, et les informations contextuelles sont prises en compte en déployant l'attention sur elles. Concrètement, une couche

d'encodeur de Transformer est appliquée aux plongements gelés concaténés. Cette représentation contextuelle du dialogue est ensuite divisée au niveau des marqueurs `[SEP]` pour aboutir à des représentations contextuelles au niveau de l'énoncé, sur lesquelles on prédit l'émotion. Afin d'adapter les représentations contextuelles des énoncés à l'objectif de prédiction des émotions, nous ajoutons un classificateur d'émotions pré-entraîné sur les données d'entraînement de DailyDialog, qui participe également à la rétro-propagation. En parallèle, les représentations contextuelles sont optimisées selon l'objectif lié à la *triplet loss*. Ceci est illustré en figure 1.

Ce scénario permet l'apprentissage des émotions individuelles et relatives, de telle sorte que chaque phase d'apprentissage renforce l'autre. Grâce à ce cadre de méta-apprentissage, des méta-informations sur les émotions sont extraites, et nous pouvons nous attendre à ce que ce modèle soit capable de réaliser une classification pertinente sur de nouvelles étiquettes en apprentissage frugal.

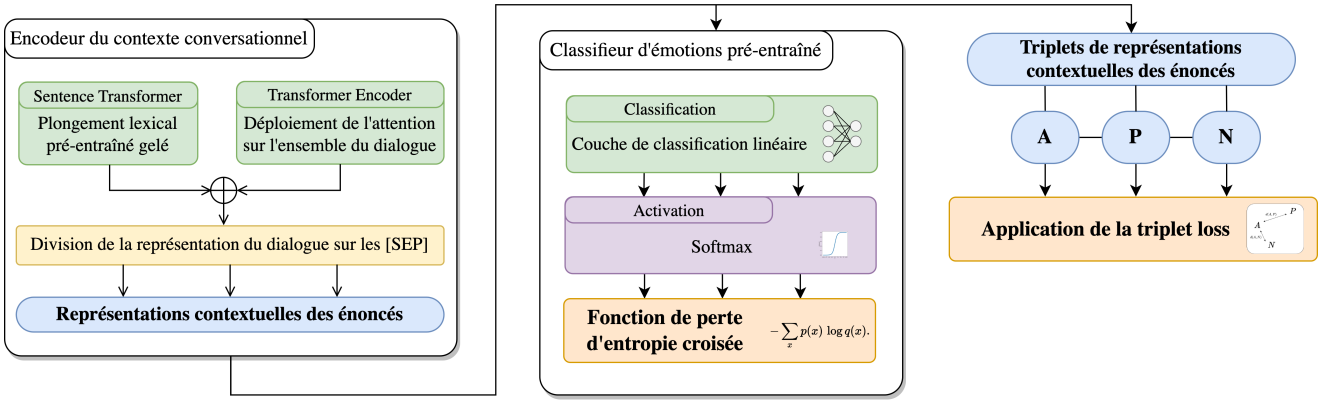


FIGURE 1 – Trois principales étapes de l'entraînement pour la prédiction d'émotions en contexte. Les fonctions de coût sont toutes deux rétro-propagées afin d'améliorer progressivement l'encodeur.

4 Protocole expérimental

Données. Toutes les expériences ont été menées sur DailyDialog (Li *et al.*, 2017) qui fournit 13 118 dialogues sur la vie quotidienne avec un étiquetage des émotions au niveau de l'énoncé. Ce jeu de données est relativement petit, ce qui permet de manipuler les entrées facilement et d'exécuter rapidement des tests. Il existe six étiquettes émotionnelles (colère, dégoût, peur, joie, tristesse et surprise) et une étiquette neutre. Pour la prédiction des émotions, l'évaluation est effectuée uniquement sur les étiquettes émotionnelles conformément aux travaux antérieurs (Ghosal *et al.*, 2021; Zhong *et al.*, 2019). Nous utilisons les sous-ensembles originaux (entraînement, validation et test) de (Li *et al.*, 2017). Les principales caractéristiques de DailyDialog sont données table 1.

Langue	Type	Max Msg/Conv	Moy Msg/Conv	Labels	Labels*	Nb. Conv
Anglais	Artificiel	35	8	7	6	13 118

TABLE 1 – Statistiques de DailyDialog (Li *et al.*, 2017). Labels* exclut l'étiquette neutre.

Spécificités du modèle. Pour le modèle avec énoncés isolés, nous considérons deux types de sous-réseaux : les couches linéaires simples et les couches récurrentes à mémoire court et long terme (*Long*

Short-Term Memory, LSTM, Hochreiter & Schmidhuber (1997)). Dans le cas contextuel, le sous-réseau est un encodeur de Transformer. Nous avons utilisé trois modèles de Transformers pré-entraînés au niveau de la phrase disponibles dans la bibliothèque Python `sentence transformers`[1] : MPNet (Song *et al.*, 2020), MiniLM (Wang *et al.*, 2020) et RoBERTa (Liu *et al.*, 2019).

Spécificités de l'entraînement. Que ce soit pour le modèle avec énoncés isolés ou pour le modèle contextuel, la prédiction de l'émotion est au niveau de l'énoncé. Les triplets sont donc toujours des triplets d'énoncés. Cela occasionne un problème d'équilibre des classes comme le montre la distribution des émotions de DailyDialog figure 3. Ainsi, le rééquilibrage des classes induit par l'échantillonnage des triplets selon une distribution uniforme n'atténue pas suffisamment les biais pendant l'apprentissage et empêche la fonction de coût de converger. Nous avons mis en œuvre un échantillonneur pondéré par l'inverse des fréquences des étiquettes afin de tenir compte de la rareté de certaines étiquettes telles que `fear` (peur) ou `disgust` (dégoût).

Évaluation quantitative. Nous considérons à la fois la performance et la pertinence de l'entraînement afin que bénéficier des capacités de généralisation du méta-apprentissage. Ainsi, nous avons choisi, en plus des mesures de performance habituelles, une métrique très exigeante : le coefficient de corrélation de Matthews (MCC) (Cramér, 1946). Il mesure la corrélation de Pearson (Pearson, 1895) entre classe prédite et classe réelle. Le MCC est défini dans (Matthews, 1975) comme suit :

$$\text{MCC} = \frac{TP/N - S \times P}{\sqrt{PS(1-S)(1-P)}} \tag{1}$$

Comparaison avec les LLMs. Afin de mettre en perspective les résultats de nos modèles isolés et contextuels, nous comparons nos modèles avec des LLMs état-de-l'art, à savoir LlaMA 2 (Touvron *et al.*, 2023) et Falcon (Penedo *et al.*, 2023). Les deux sont considérés dans leur version adaptée aux instructions (*instruction fine-tuning*) et évalués en génération sur un seul essai (*zero-shot*). Nous avons développé une requête (*cf.* fig. 2, *prompt* en anglais) demandant une prédiction sur le dernier énoncé de chaque dialogue du jeu de test de DailyDialog. Les requêtes sont conçues de sorte à ce que le modèle ne génère qu'une seule étiquette. La requête est la même pour chaque modèle du même type (LlaMA ou Falcon). Il est difficile de trouver une bonne requête pour Falcon car le modèle génère `happiness` (joie) sur l'ensemble des données, à l'exception d'un dialogue.

```
Here is a dialog :

-  Hello , Miao Li , Where are you going ?
-   Hello , I am going to the store to buy some fruit .
-   Oh , Would you do me a favor ?
-   Yes ?
-   Please mail this letter for me on your way to the store .
-   Sure . Do you want it to be registered ?
-   Yes , I think so . There are some pictures in it . It would be a great
pity if they were lost .
-   Yes , I will be glad to mail your letter .
-   Thanks .
-   you are welcome .

Regarding its conversational context, give me the appropriate emotion to
describe this utterance : "Yes , I think so . There are some pictures in
it . It would be a great pity if they were lost .", using only one of the
following labels: happiness, sadness, anger, surprise, fear, disgust, no
emotion. Predicted label :
```

(a) Requête utilisée pour Llama2

```
Here is a dialog :

-  Hello , Miao Li , Where are you going ?
-   Hello , I am going to the store to buy some fruit .
-   Oh , Would you do me a favor ?
-   Yes ?
-   Please mail this letter for me on your way to the store .
-   Sure . Do you want it to be registered ?
-   Yes , I think so . There are some pictures in it . It would be a great
pity if they were lost .
-   Yes , I will be glad to mail your letter .
-   Thanks .
-   you are welcome .

Regarding its conversational context, return the appropriate emotion for
the last utterance among: sadness, happiness, anger, surprise, fear and
disgust. If none of them properly correspond, return 'no emotion'.
```

(b) Requête utilisée pour Falcon

FIGURE 2 – Requêtes pour Llama2 et Falcon

1. https://www.sbert.net/

5 Résultats

Nom du modèle	macroF1*	microF1*	MCC
Modèles état-de-l'art en ERC			
CNN+cLSTM (Poria *et al.*, 2017)	–	50.24	–
KET (Zhong *et al.*, 2019)	–	53.37	–
COSMIC (Ghosal *et al.*, 2020)	51.05	58.48	–
RoBERTa (Ghosal *et al.*, 2020)	48.20	55.16	–
Rpe-RGAT (Ishiwatari *et al.*, 2020)	–	54.31	–
Glove-DRNN (Ghosal *et al.*, 2021)	41.80	55.95	–
roBERTa-DRNN (Ghosal *et al.*, 2021)	49.65	57.32	–
CNN (Ghosal *et al.*, 2021)	36.87	50.32	–
DAG-ERC (Shen *et al.*, 2021)	–	59.33	–
TODKAT (Zhu *et al.*, 2021)	<u>52.56</u>	58.47	–
SKAIG (Li *et al.*, 2021)	51.95	59.75	–
Sentic GAT (Tu *et al.*, 2022)	–	54.45	–
CauAIN (Zhao *et al.*, 2022)	–	58.21	–
DialogueRole (Ong *et al.*, 2022)	–	60.95	–
S+PAGE (Liang *et al.*, 2022)	–	**64.07**	–
DualGAT (Zhang *et al.*, 2023)	–	<u>61.84</u>	–
CD-ERC (Pereira *et al.*, 2023)	51.23	–	–
LLMs			
Llama2-7b (Touvron *et al.*, 2023)	09.70	24.92	0.08
Llama2-13b (Touvron *et al.*, 2023)	22.26	43.37	0.15
Falcon-7b (Penedo *et al.*, 2023)	07.54	42.75	0.01
Notre approche			
SentEmoContext	**57.71**	57.75	**0.49**

TABLE 2 – Résultats en ERC sur DailyDialog, en utilisant le jeu de test de l'article d'origine. DRNN réfère à DialogueRNN. L'astérisque (*) indique l'exclusion de l'étiquette neutre.

Travaux connexes. Le tableau 2 donne les résultats en ERC sur DailyDialog. On observe une lente progression depuis 2017 où Poria *et al.* (2017) propose d'évaluer en microF1 en excluant la classe neutre (majoritaire). Ce modèle est une première référence pour cette tâche, obtenant 50,24% en microF1. En revanche, le modèle état-de-l'art actuel atteint maintenant 64,07% en microF1 (Liang *et al.*, 2022), ce qui représente une amélioration d'environ 14 points en 6 ans. Comme le montre la table 2, la communauté a suivi ce schéma d'évaluation. Cependant, nous pensons qu'il est important de prendre également en compte le macroF1, à l'exclusion de la classe majoritaire, car il montre la performance globale sur toutes les émotions. Certains travaux l'ont proposé à partir de 2020 (Ghosal *et al.*, 2020), conduisant à un gain de 2,5 points en 3 ans. Cela renforce l'affirmation selon laquelle l'ERC est une tâche difficile.

Notre modèle. SentEmoContext obtient 57,75% en microF1, un résultat décent mais quelque peu modeste comparé aux travaux connexes. La table 2 donne la performance moyenne de notre modèle sur 10 exécutions. Notre modèle est état-de-l'art en macroF1 avec 57,71%, surpassant CD-ERC (Pereira *et al.*, 2023) de 6,48 points, ce qui est considérable étant donné qu'ils ne se sont concentrés que sur

cette métrique, et TODKAT (Zhu *et al.*, 2021) de 5,15 points. Nous évaluons également notre modèle en MCC multi-classe (Baldi *et al.*, 2000) pour assurer sa pertinence, sans qu'elle soit comparable aux travaux connexes car ils ne donnent pas cette métrique. Il fournit ici un bon indicateur de la qualité de la classification, en minimisant l'effet des données hautement déséquilibrées des conversations. Étant donné que le MCC varie de -1 à 1, et que 0 indique le caractère aléatoire, un MCC de 0,49 indique que notre approche est à la fois équilibrée et précise en termes de prédictions.

Notre modèle est très performant car nous n'avons besoin que de 20 minutes par époque et nous l'entraînons en utilisant seulement 5 époques. Ceci dénote des approches existantes qui utilisent plusieurs flux par locuteur (Pereira *et al.*, 2023), la modélisation graphique pour la représentation du contexte et des connaissances (Zhong *et al.*, 2019; Li *et al.*, 2021), ou d'autres représentations lourdes dans leur modèle (Liang *et al.*, 2022). Notre modèle est stable avec un écart-type de seulement 0,01 en moyenne pour chaque métrique, ce qui renforce la qualité de cette approche efficace.

Comparaison avec les classifieurs d'émotions au niveau de l'énoncé. La table 3 montre les résultats de la classification directe des émotions sur les énoncés. Pour cette tâche, nous n'avons pris en compte que les 6 étiquettes d'émotion, en excluant complètement l'étiquette neutre. Ainsi, nous voulons déterminer la différence entre notre approche et la prédiction d'émotions isolées. Cela sert également d'étude d'ablation pour notre modèle SentEmoContext puisque cette étape fait partie de son entraînement. En table 3, nous voyons que notre modèle exploite le contexte conversationnel et le *metric learning* pour augmenter toutes les métriques. Notamment, la différence en termes de macroF1 montre l'importance de la *triplet loss* dans notre modèle. En effet, les classifieurs d'émotions sont entraînés en utilisant des lots équilibrés sur la distribution du jeu de données d'entraînement et une fonction de perte d'entropie croisée pondérée. Les résultats montrent que cela n'est pas suffisant pour traiter des données extrêmement déséquilibrées telles que des conversations.

Nom du modèle	**macroF1**	**microF1**	**MCC**
Classifieur d'émotions pré-entrainé sur les tours de parole			
all-MiniLM-L6-v2	20.22	33.11	0.40
all-mpnet-base-v2	14.43	32.90	0.37
Notre approche			
SentEmoContext	**57.71**	**57.75**	**0.49**

TABLE 3 – Comparaison avec un classifieur d'émotions agissant au niveau du tour de parole.

LLMs. Les résultats des LLMs sur un seul essai sont donnés en table 2. Ceux-ci servent d'indication sur la performance de tels modèles (allégés) en ERC. Même si ces modèles génératifs ne sont pas conçus pour cette tâche spécifique, ils réussissent toujours à surpasser les classifieurs d'émotions d'énoncés de la table 3, ce qui peut être considéré comme une manifestation des capacités émergentes des LLMs – *emerging abilities* (Srivastava *et al.*, 2022).

Facteur du déséquilibre des classes. Alors que la table 1 montre les caractéristiques du jeu de données, elle omet la principale caractéristique des étiquettes d'émotions : un fort déséquilibre. En ERC, les principales difficultés sont la définition des étiquettes, du contexte mais aussi le déséquilibre qui limite l'apprentissage des émotions en contexte. La figure 3 montre la distribution des étiquettes dans DailyDialog, sans l'étiquette neutre. Étant donné que cette dernière est l'étiquette majoritaire et qu'elle est exclue des métriques d'évaluation par l'ensemble de la communauté ERC, le fait que même dans les étiquettes d'émotion les données soient aussi déséquilibrées s'avère être un défi et

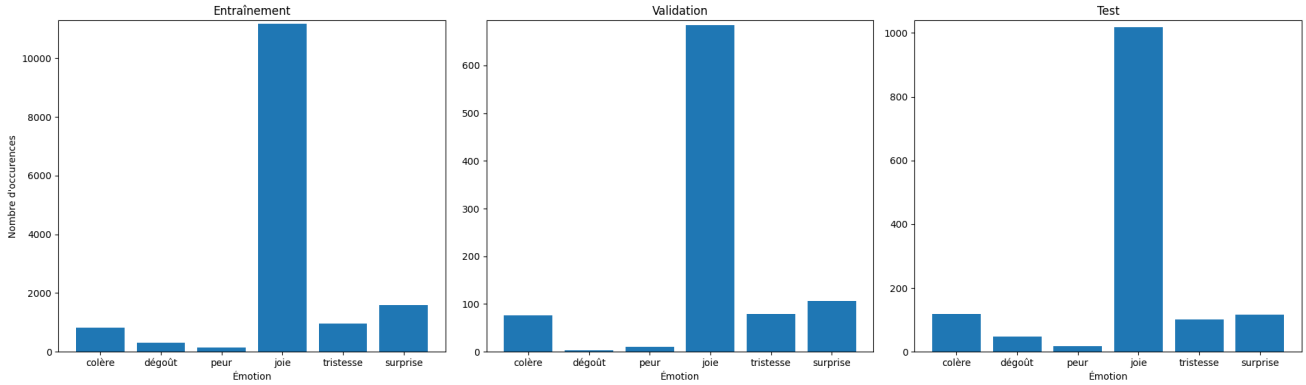


FIGURE 3 – Distributions des étiquettes émotionnelles dans les sous-ensembles de DailyDialog.

doit être abordé. Nous nous appuyons sur Guibon *et al.* (2023) pour traiter le déséquilibre en deux phases : tout d'abord, nous équilibrons les lots en fonction de la fréquence des étiquettes dans le jeu d'apprentissage. Ensuite, nous pondérons la fonction de perte d'entropie croisée du classificateur d'émotions en tenant compte du déséquilibre restant dans chaque lot. Ce travail traite également le déséquilibre en considérant des triplets, car nous éliminons alors le facteur de déséquilibre tout en utilisant les états cachés qui proviennent d'une représentation équilibrée. Nous pensons que cela explique en partie l'efficacité et l'efficience de notre modèle.

6 Discussion

Limitations des LLMs. La première limitation rencontrée avec les LLMs est la nécessité de GPUs à haute mémoire pour les tester. Ceci explique pourquoi en table 2 nous considérons seulement leurs versions légères. Alors que Llama2 7b et 13b ont donné des réponses dans un bon format, avec une seule étiquette, Falcon ne s'est pas comporté comme nous le souhaitions. Pour pallier ce problème, nous considérons la première émotion mentionnée dans la sortie. Il est également important de noter que nous n'avons pas voulu utiliser ChatGPT d'OpenAI car nous n'avons pas un contrôle clair sur la version du modèle, la taille et l'approche utilisée derrière l'API, mais aussi parce que nous voulions utiliser des systèmes open source pour pouvoir diffuser nos modèles à la communauté.

La fenêtre contextuelle constitue une autre limitation. En ERC, la taille du contexte est essentielle, mais avec les LLMs, l'ajout d'exemples dans la requête pour effectuer un apprentissage frugal prendrait beaucoup de place dans le contexte global, la requête faisant partie du contexte. Ceci explique notre décision de ne considérer que l'apprentissage en un seul essai pour les LLMs, même s'il conviendrait de considérer également l'apprentissage frugal sur cette tâche spécifique.

Taille et efficacité du modèle. Notre modèle est efficace. Il donne des résultats état-de-l'art en macroF1 et de bons résultats en microF1, alors qu'il s'entraîne relativement vite et ne nécessite pas beaucoup d'époques pour converger. Nous pensons que cette efficacité, ainsi que la mémoire limitée nécessaire à l'apprentissage, est due à la rétro-propagation en deux étapes et au fait que nous utilisons des représentations intégrées à l'énoncé avec des Transformers au niveau de la phrase. Ainsi, notre modèle peut traiter efficacement de longs contextes conversationnels avec un coût limité en mémoire. En outre, la table 4 montre la différence entre les modèles que nous avons utilisés, en termes de taille, de paramètres et de nombre de couches. Notre modèle est relativement petit si l'on considère les

	Transformers		LLMs			Notre approche
Modèle	MiniLM	MPNet	Llama2-7b	Llama2-13b	Falcon-7b	SentEmoContext
Tokens	1bn+	1bn+	2T	2T	1.5T	4M
Taille	80 MB	420 MB	13 GB	25 GB	15 GB	604,8 MB
Paramètres	22M	110M	7B	13B	7B	157M

TABLE 4 – Aperçu de la taille des modèles. Les modèles LLaMA sont deux versions de LLaMA 2. MiniLM et MPNet sont les mêmes que ceux présentés en table 3.

avancées récentes et les travaux connexes en ERC, mais aussi par rapport aux LLMs.

Représentation relative des étiquettes. Notre approche apprend deux fois à partir des données, d'abord en utilisant un cadre supervisé, puis en tenant compte des distances relatives entre les représentations, en mettant à jour par la *triplet loss*. Cela permet d'utiliser notre modèle pour différents jeux de données de conversations avec différentes étiquettes. La seule exigence pour étendre la portée de ce modèle serait de considérer une autre stratégie d'échantillonnage des triplets en ignorant les étiquettes, telle que la stratégie *batch-hard* (Do *et al.*, 2019).

7 Conclusion

Dans cet article, nous présentons notre modèle SentEmoContext issu d'une approche combinant la représentation au niveau de l'énoncé, le *metric learning* et les réseaux siamois à l'aide de la *triplet loss*. Ce modèle représente efficacement le contexte conversationnel, atteint un score état-de-l'art en macroF1 de 57.71%, et un microF1 satisfaisant de 57.75% en ERC sur DailyDialog. Nous proposons également d'utiliser le coefficient de corrélation de Matthew afin de mieux évaluer cette tâche.

Avec SentEmoContext , nous utilisons l'apprentissage contrastif avec un échantillonage pour limiter le déséquilibre des classes. Nous utilisons Sentence BERT pour minimiser la mémoire nécessaire tout en représentant l'ensemble du contexte conversationnel. Cela conduit à un apprentissage plus robuste et plus efficace qui ne nécessite pas beaucoup d'époques pour obtenir des résultats satisfaisants. Nous montrons également que les LLMs open source de taille modeste sont en retard en ERC, car cette tâche nécessite d'incorporer beaucoup de contexte dans la requête et n'est pas spécifiquement pertinente pour les modèles génératifs.

Dans nos travaux futurs, nous envisageons d'appliquer cette approche à des données conversationnelles proposant des étiquettes légèrement différentes, car notre modèle apprend les émotions de manière relative. Nous prévoyons donc de l'adapter à un cadre davantage lié au méta-apprentissage.

Remerciements

Les expériences présentées dans cet article ont été réalisées sur le banc d'essai Grid'5000, soutenu par un groupement d'intérêt scientifique hébergé par Inria et comprenant le CNRS, RENATER et plusieurs universités ainsi que d'autres organisations (voir https://www.grid5000.fr).

Références

ANTONIOU A., EDWARDS H. & STORKEY A. (2019). How to train your maml. Seventh International Conference on Learning Representations, ICLR 2019.

BAHDANAU D., CHO K. & BENGIO Y. (2014). Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015*, **1409**.

BALDI P., BRUNAK S., CHAUVIN Y., ANDERSEN C. & NIELSEN H. (2000). Assessing the accuracy of prediction algorithms for classification : An overview. *Bioinformatics (Oxford, England)*, **16**, 412–24.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146. DOI : 10.1162/tacl_a_00051.

BUSSO C., BULUT M., LEE C.-C., KAZEMZADEH A., MOWER E., KIM S., CHANG J. N., LEE S. & NARAYANAN S. S. (2008). Iemocap : interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, **42**(4), 335–359.

CHECHIK G., SHARMA V., SHALIT U. & BENGIO S. (2010). Large scale online learning of image similarity through ranking. *J. Mach. Learn. Res.*, **11**, 1109–1135.

CRAMÉR H. (1946). *Mathematical Methods of Statistics (PMS-9), Volume 9*. Princeton : Princeton University Press. DOI : doi :10.1515/9781400883868.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DO T.-T., TRAN T., REID I., KUMAR V., HOANG T. & CARNEIRO G. (2019). A theoretically sound upper bound on the triplet loss for improving the efficiency of deep distance metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 10404–10413.

DUMOULIN V., HOULSBY N., EVCI U., ZHAI X., GOROSHIN R., GELLY S. & LAROCHELLE H. (2021). A unified few-shot classification benchmark to compare transfer and meta learning approaches. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

EKMAN P., SORENSON E. R. & FRIESEN W. V. (1969). Pan-cultural elements in facial displays of emotion. DOI : https ://doi.org/10.1126/science.164.3875.86.

FINN C., ABBEEL P. & LEVINE S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, p. 1126–1135 : JMLR.org.

GHOSAL D., MAJUMDER N., GELBUKH A., MIHALCEA R. & PORIA S. (2020). COSMIC : COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 2470–2481, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.224.

GHOSAL D., MAJUMDER N., MIHALCEA R. & PORIA S. (2021). Exploring the role of context in utterance-level emotion, act and intent classification in conversations : An empirical study. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 1435–1449, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.124.

GHOSAL D., MAJUMDER N., PORIA S., CHHAYA N. & GELBUKH A. (2019). DialogueGCN : A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 154–164, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1015.

GUIBON G., LABEAU M., FLAMEIN H., LEFEUVRE L. & CLAVEL C. (2021). Few-shot emotion recognition in conversation with sequential prototypical networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic.

GUIBON G., LABEAU M., LEFEUVRE L. & CLAVEL C. (2023). An adaptive layer to leverage both domain and task specific information from scarce data. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**(6), 7757–7765. DOI : 10.1609/aaai.v37i6.25940.

HAYALE W., NEGI P. S. & MAHOOR M. H. (2023). Deep siamese neural networks for facial expression recognition in the wild. *IEEE Transactions on Affective Computing*, **14**(2), 1148–1158. DOI : 10.1109/TAFFC.2021.3077248.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long Short-Term Memory. *Neural Computation*, **9**(8), 1735–1780. DOI : 10.1162/neco.1997.9.8.1735.

HOSPEDALES T., ANTONIOU A., MICAELLI P. & STORKEY A. (2022). Meta-learning in neural networks : A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**(9), 5149–5169.

HU G., LIN T.-E., ZHAO Y., LU G., WU Y. & LI Y. (2022). UniMSE : Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 7837–7851, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.534.

ISHIWATARI T., YASUDA Y., MIYAZAKI T. & GOTO J. (2020). Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 7360–7370, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.597.

JIN M., ZHENG Y., LI Y.-F., GONG C., ZHOU C. & PAN S. (2021). Multi-scale contrastive siamese networks for self-supervised graph representation learning. In Z.-H. ZHOU, Éd., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, p. 1477–1483 : International Joint Conferences on Artificial Intelligence Organization. Main Track, DOI : 10.24963/ijcai.2021/204.

JORDAN M. I. (1986). Serial order : a parallel distributed processing approach. technical report, june 1985-march 1986.

KHOSLA P., TETERWAK P., WANG C., SARNA A., TIAN Y., ISOLA P., MASCHINOT A., LIU C. & KRISHNAN D. (2020). Supervised contrastive learning. *Advances in neural information processing systems*, **33**, 18661–18673.

KOCH G., ZEMEL R. & SALAKHUTDINOV R. (2015). Siamese neural networks for one-shot image recognition.

LEE B. & CHOI Y. S. (2021). Graph based network with contextualized representations of turns in dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 443–455, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.36.

LI J., LIN Z., FU P. & WANG W. (2021). Past, present, and future : Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the Association*

for Computational Linguistics : EMNLP 2021, p. 1204–1214, Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-emnlp.104.

LI Y., SU H., SHEN X., LI W., CAO Z. & NIU S. (2017). DailyDialog : A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 986–995, Taipei, Taiwan : Asian Federation of Natural Language Processing.

LI Z., TANG F., ZHAO M. & ZHU Y. (2022). EmoCaps : Emotion capsule based model for conversational emotion recognition. In *Findings of the Association for Computational Linguistics : ACL 2022*, p. 1610–1618, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-acl.126.

LIANG C., XU J., LIN Y., YANG C. & WANG Y. (2022). S+PAGE : A speaker and position-aware graph neural network model for emotion recognition in conversation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 148–157, Online only : Association for Computational Linguistics.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach.

LORRAINE J., VICOL P. & DUVENAUD D. (2020). Optimizing millions of hyperparameters by implicit differentiation. In S. CHIAPPA & R. CALANDRA, Éds., *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 de *Proceedings of Machine Learning Research*, p. 1540–1552 : PMLR.

MAJUMDER N., PORIA S., HAZARIKA D., MIHALCEA R., GELBUKH A. & CAMBRIA E. (2019). Dialoguernn : An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**(01), 6818–6825. DOI : 10.1609/aaai.v33i01.33016818.

MATTHEWS B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et biophysica acta*, **405 2**, 442–51.

MISHRA N., ROHANINEJAD M., CHEN X. & ABBEEL P. (2017). A simple neural attentive meta-learner. In *International Conference on Learning Representations*.

ONG D., SU J., CHEN B., LUU A. T., NARENDRANATH A., LI Y., SUN S., LIN Y. & WANG H. (2022). Is discourse role important for emotion recognition in conversation ? *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**(10), 11121–11129. DOI : 10.1609/aaai.v36i10.21361.

PEARSON K. (1895). Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, **58**(347-352), 240–242.

PENEDO G., MALARTIC Q., HESSLOW D., COJOCARU R., CAPPELLI A., ALOBEIDLI H., PANNIER B., ALMAZROUEI E. & LAUNAY J. (2023). The refinedweb dataset for falcon llm : Outperforming curated corpora with web data, and web data only.

PEREIRA P., MONIZ H., DIAS I. & CARVALHO J. P. (2023). Context-dependent embedding utterance representations for emotion recognition in conversations. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, p. 228–236, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.wassa-1.21.

PLUTCHIK R. (2001). The Nature of Emotions. *American Scientist*, **89**(4), 344. DOI : 10.1511/2001.4.344.

PORIA S., CAMBRIA E., HAZARIKA D., MAJUMDER N., ZADEH A. & MORENCY L.-P. (2017). Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual*

Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), p. 873–883, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1081.

PORIA S., HAZARIKA D., MAJUMDER N., NAIK G., CAMBRIA E. & MIHALCEA R. (2019). MELD : A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 527–536, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1050.

RAVI S. & LAROCHELLE H. (2016). Optimization as a model for few-shot learning. In *International Conference on Learning Representations*.

REN F. & XUE S. (2020). Intention detection based on siamese neural network with triplet loss. *IEEE Access*, **8**, 82242–82254. DOI : 10.1109/ACCESS.2020.2991484.

RUMELHART D. E., HINTON G. E. & WILLIAMS R. J. (1985). *Learning internal representations by error propagation*. Rapport interne, California Univ San Diego La Jolla Inst for Cognitive Science.

SCHROFF F., KALENICHENKO D. & PHILBIN J. (2015). Facenet : A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 815–823. DOI : 10.1109/CVPR.2015.7298682.

SCHULTZ M. & JOACHIMS T. (2003). Learning a distance metric from relative comparisons. In S. THRUN, L. SAUL & B. SCHÖLKOPF, Éds., *Advances in Neural Information Processing Systems*, volume 16 : MIT Press.

SHEN W., WU S., YANG Y. & QUAN X. (2021). Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1551–1560, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.123.

SNELL J., SWERSKY K. & ZEMEL R. (2017). Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 4080–4090, Red Hook, NY, USA : Curran Associates Inc.

SONG K., TAN X., QIN T., LU J. & LIU T.-Y. (2020). Mpnet : Masked and permuted pre-training for language understanding. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 16857–16867 : Curran Associates, Inc.

SONG X., HUANG L., XUE H. & HU S. (2022). Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 5197–5206, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.347.

SRIVASTAVA A. *et al.* (2022). Beyond the imitation game : Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv :2206.04615*.

SUNG F., YANG Y., ZHANG L., XIANG T., TORR P. H. & HOSPEDALES T. M. (2018). Learning to compare : Relation network for few-shot learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 1199–1208. DOI : 10.1109/CVPR.2018.00131.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S., BIKEL D., BLECHER L., FERRER C. C., CHEN M., CUCURULL G., ESIOBU D., FERNANDES J., FU J., FU W., FULLER B., GAO C., GOSWAMI V., GOYAL N., HARTSHORN A., HOSSEINI S., HOU R., INAN H., KARDAS M., KERKEZ V., KHABSA M., KLOUMANN I., KORENEV A., KOURA P. S., LACHAUX M.-A., LAVRIL T., LEE J., LISKOVICH D., LU Y., MAO Y., MARTINET X., MIHAYLOV T., MISHRA P., MOLYBOG I., NIE

Y., POULTON A., REIZENSTEIN J., RUNGTA R., SALADI K., SCHELTEN A., SILVA R., SMITH E. M., SUBRAMANIAN R., TAN X. E., TANG B., TAYLOR R., WILLIAMS A., KUAN J. X., XU P., YAN Z., ZAROV I., ZHANG Y., FAN A., KAMBADUR M., NARANG S., RODRIGUEZ A., STOJNIC R., EDUNOV S. & SCIALOM T. (2023). Llama 2 : Open foundation and fine-tuned chat models.

TU G., WEN J., LIU C., JIANG D. & CAMBRIA E. (2022). Context- and sentiment-aware networks for emotion recognition in conversation. *IEEE Transactions on Artificial Intelligence*, **3**(5), 699–708. DOI : 10.1109/TAI.2022.3149234.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need.

VINYALS O., BLUNDELL C., LILLICRAP T., KAVUKCUOGLU K. & WIERSTRA D. (2016). Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, p. 3637–3645, Red Hook, NY, USA : Curran Associates Inc.

WANG W., WEI F., DONG L., BAO H., YANG N. & ZHOU M. (2020). Minilm : Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 5776–5788 : Curran Associates, Inc.

ZHANG D., CHEN F. & CHEN X. (2023). DualGATs : Dual graph attention networks for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 7395–7408, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.408.

ZHAO W., ZHAO Y. & LU X. (2022). Cauain : Causal aware interaction network for emotion recognition in conversations. In L. D. RAEDT, Éd., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, p. 4524–4530 : International Joint Conferences on Artificial Intelligence Organization. Main Track, DOI : 10.24963/ijcai.2022/628.

ZHONG P., WANG D. & MIAO C. (2019). Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 165–176, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1016.

ZHU L., PERGOLA G., GUI L., ZHOU D. & HE Y. (2021). Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1571–1582, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.125.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Une approche par graphe pour l'analyse syntaxique en dépendances de bout en bout de la parole

Adrien Pupier, Maximin Coavoux, Benjamin Lecouteux, Jérôme Goulian
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
first.last@univ-grenoble-alpes.fr

RÉSUMÉ

Effectuer l'analyse syntaxique du signal audio –plutôt que de passer par des transcriptions de l'audio– est une tache récemment proposée par Pupier *et al.* (2022), dans le but d'incorporer de l'information prosodique dans le modèle d'analyse syntaxique et de passer outre les limitations d'une approche cascade qui consisterait à utiliser un système de reconnaissance de la parole (RAP) puis un analyseur syntaxique. Dans cet article, nous effectuons un ensemble d'expériences visant à comparer les performances de deux familles d'analyseurs syntaxiques : (i) l'approche par graphe et (ii) la réduction à une tâche d'étiquetage de séquence ; directement sur la parole. Nous évaluons notre approche sur un corpus arboré du français parlé. Nous montrons que (i) l'approche par graphe obtient de meilleurs résultats globalement et (ii) qu'effectuer l'analyse syntaxique directement depuis la parole obtient de meilleurs résultats qu'une approche par cascade de systèmes, malgré 30% de paramètres en moins.

ABSTRACT

A graph-based parser for end-to-end dependency parsing of speech

Direct dependency parsing of the speech signal –as opposed to parsing speech transcriptions– has recently been proposed as a task (Pupier *et al.*, 2022), as a way of incorporating prosodic information into the parsing system and bypassing the limitations of a pipeline approach that would consist of using first an Automatic Speech Recognition (ASR) system and then a syntactic parser. In this article, we report on a set of experiments aiming at assessing the performance of two parsing paradigms (graph-based parsing and sequence labeling based parsing) on speech parsing. We perform this evaluation on a large treebank of spoken French, featuring realistic spontaneous conversations. Our findings show that (i) the graph based approach obtain better results across the board, (ii) parsing directly from speech outperforms a pipeline approach, despite having 30% fewer parameters.

MOTS-CLÉS : Analyse syntaxique, parole, modèle pré-entraîné, bout-en-bout.

KEYWORDS: Syntactic parsing, Speech, Pre-trained model, end-to-end.

1 Introduction

L'analyse syntaxique est une tâche centrale en traitement automatique des langues (TAL). Au sein de la communauté du TAL, celle-ci a été plutôt effectuée sur des données textuelles, ou plus rarement sur des transcriptions de la parole. Pourtant, la parole est la principale forme de communication entre humain·es, et constitue le type de données linguistiques le plus naturel, ce qui motive le développement de systèmes de TAL pour la parole, à la fois pour des objectifs applicatifs (par exemple : le sous-titrage

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 234–244.

automatique) que pour des objectifs de recherche (par exemple : construire des corpus de grande taille annotés en syntaxe). La majorité des travaux existant sur l'analyse syntaxique de transcriptions de la parole se concentrent sur l'anglais et sur la détection et la suppression de disfluences (Charniak & Johnson, 2001; Johnson & Charniak, 2004; Rasooli & Tetreault, 2013; Honnibal & Johnson, 2014; Jamshid Lou *et al.*, 2019). Autrement dit, il s'agit de 'normaliser' les transcriptions de manière à pouvoir les utiliser en aval avec des systèmes de TAL entraînés sur du texte 'standard'. Utiliser uniquement des transcriptions en entrée est un choix naturel d'un point de vue du TAL : il est possible d'utiliser des analyseurs syntaxiques état de l'art sans modifications. Cependant, les transcriptions prédites peuvent être très bruitées, particulièrement dans le cas des conversations spontanées. De plus, les transcriptions sont des abstractions contenant beaucoup moins d'informations que le signal audio. La prosodie, ainsi que les pauses dans un énoncé, sont des indices importants pour l'analyse syntaxique (Price *et al.*, 1991) qui sont complètement absentes des transcriptions. Ainsi, nous effectuons cette tâche en utilisant seulement le signal audio en entrée. En particulier nous proposons un nouvel algorithme d'analyse syntaxique en dépendances pour la parole utilisant l'approche par graphe (*graph-based*).

Avec la popularisation des méthodes auto-supervisées et les réseaux neuronaux profonds, les domaines de la parole et du texte utilisent désormais une méthodologie similaire (Chrupała, 2023) : affiner les paramètres d'un modèle auto-supervisé générique sur une tâche applicative particulière. Cette convergence méthodologique a suscité l'intérêt d'autres applications des modèles de l'audio pour aller au-delà de la 'simple' reconnaissance de la parole. Ainsi, aborder les tâches classiques du TAL directement depuis l'audio est un enjeu important pour concevoir des outils de TAL robustes au bruit présent dans la parole. De plus, de nombreuses langues n'ont pas de forme écrite, ce qui proscrit d'utiliser des outils classiques du TAL et motive le développement d'outils de documentation travaillant directement sur le signal audio.

En résumé, nos contributions sont les suivantes :

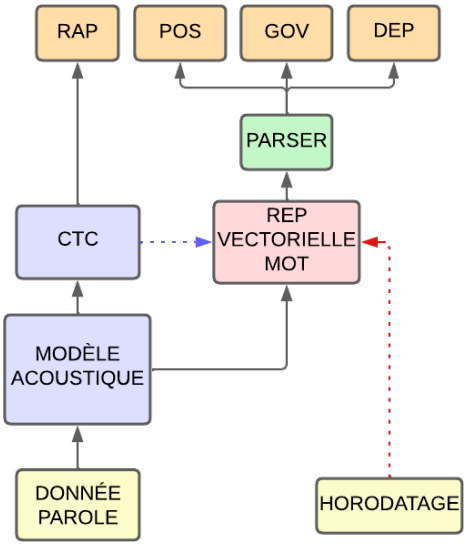
— Nous présentons un algorithme bout-en-bout d'analyse syntaxique basé sur les graphes ;
— Nous évaluons notre analyseur sur Orféo (Benzitoun *et al.*, 2016), un corpus arboré du français parlé contenant essentiellement de la parole spontanée ; nous comparons le parser à une approche cascade ainsi qu'à un analyseur basé sur une réduction du problème à l'étiquetage de séquence (Strzyz *et al.*, 2019) ;
— Nous publions le code à l'adresse suivante : `https://github.com/Pupiera/Growing_tree_on_sound`.

2 Analyseur syntaxique et modèle pré-entraîné

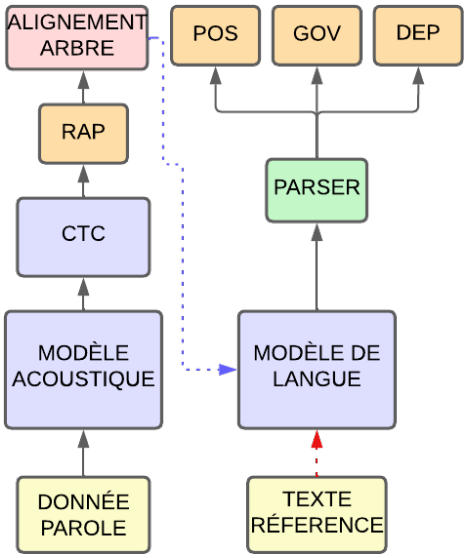
Nous définissons l'analyse syntaxique de la parole comme la tâche consistant à prédire un arbre de dépendance depuis un signal audio correspondant à un énoncé [1]. Les nœuds de l'arbre syntaxique étant constitués des tokens, nous réalisons la reconnaissance automatique de la parole de manière conjointe à l'analyse syntaxique.

Notre analyseur est composé de deux modules (figure 1a) : (i) un modèle acoustique qui prédit les transcriptions et la segmentation du signal en mots et (ii) un module d'analyse syntaxique qui utilise cette segmentation pour construire des représentations vectorielles de mots et prédit les arbres. Ces

1. Dans le reste de l'article, nous utiliserons le terme *phrase* pour garder la terminologie issue de la littérature en analyse syntaxique automatique, même si ce terme est sujet à débat lorsqu'il s'agit de langue parlée.



(a) Les deux modèles basés sur des représentations de l'audio. La flèche bleue désigne la configuration **AUDIO**, et la flèche rouge **ORACLE** (cf section 4).



(b) Les deux modèles de référence basés sur un modèle de langue pré-entraîné. La flèche bleue désigne la configuration **CASCADE**, et la flèche rouge **TEXTE**, (transcriptions manuelles de référence).

FIGURE 1 – Vue générale de l'architecture avec les 4 configurations décrites dans la section 4.

deux modules sont entraînés de manière simultanée.

Construire des vecteurs de mots directement à partir du signal de parole Pour extraire des représentations depuis le signal de parole, nous utilisons un modèle wav2vec2 (Baevski *et al.*, 2020) pré-entraîné sur 7000 heures de parole en français : `LeBenchmark7K`[2] (Parcollet *et al.*, 2024). Ce modèle wav2vec2 construit une suite de représentations vectorielles, chaque vecteur correspondant à un intervalle de 25ms dans le signal de parole. On appelle cette suite de vecteurs la trame du signal.

L'analyse syntaxique nécessitant des représentations vectorielles mot. Nous utilisons la méthodologie de Pupier *et al.* (2022) pour construire des représentations de mots à partir de ces représentations du signal. Nous utilisons l'algorithme de reconnaissance de la parole CTC (*connectionist temporal classification,* Graves *et al.*, 2006), qui a la particularité d'étiqueter chaque vecteur de la trame avec soit une lettre soit un caractère de séparation ou un caractère spécial 'vide' qui permet d'obtenir plusieurs fois la même lettre à la suite. Il est ainsi possible, en identifiant le caractère spécial de séparation (espace) de segmenter la trame du signal audio en mots. Ensuite, la méthode consiste à combiner les vecteurs de trames correspondant à un seul mot à l'aide d'un LSTM, afin d'obtenir une représentation vectorielle de taille fixe pour chaque mot. Intuitivement, cette procédure peut être considérée comme l'analogue des bi-LSTM de caractères classiquement utilisés pour le texte, mais où l'on utiliserait la forme acoustique d'un mot au lieu de sa forme orthographique.

Analyse syntaxique par graphe Nous utilisons les représentations de mot audio –dont la construction est décrite ci-dessus– comme entrée de notre implémentation d'un analyseur syntaxique par

2. https://huggingface.co/LeBenchmark/wav2vec2-FR-7K-large

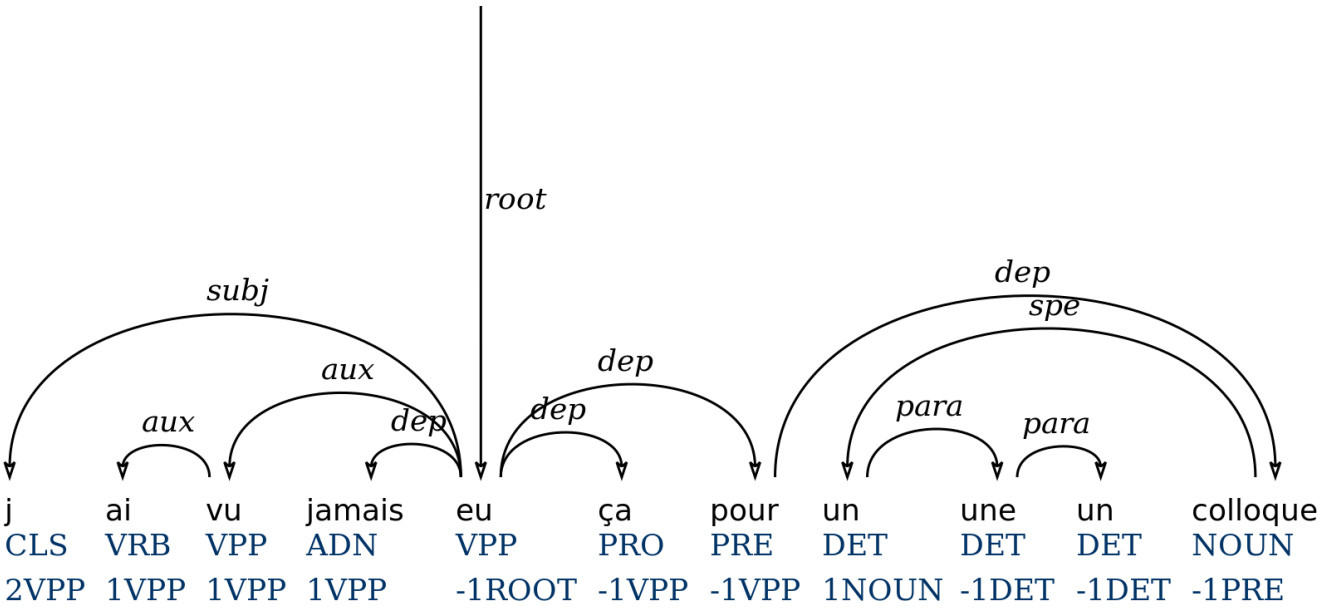


FIGURE 2 – Illustration d'une phrase extraite d'Orféo étiquetée en dep2label ainsi que l'arbre syntaxique en dépendance correspondant.

graphes classique (Dozat & Manning, 2016). Cet analyseur consiste à (i) calculer un score pour chaque arc possible à l'aide d'un classifieur biaffine puis (ii) trouver l'arbre de meilleur score (c'est-à-dire l'arbre maximisant les probabilités globales) en utilisant un algorithme d'arbre couvrant maximal.

Analyse syntaxique par étiquetage de séquence Nous utilisons le même analyseur syntaxique par étiquetage de séquence que Pupier *et al.* (2022). Cet analyseur est basé sur l'approche *dep2label* (Gómez-Rodríguez *et al.*, 2020; Strzyz *et al.*, 2020). Cette méthode réduit le problème de l'analyse syntaxique à un simple problème d'étiquetage de séquence. Elle nécessite un algorithme réversible pour coder un arbre en dépendances dans une séquence d'étiquettes. Nous utilisons l'algorithme d'encodage par position relative en prenant en compte les parties du discours (appelé *Relative POS-based* par Strzyz *et al.*, 2019). Le gouverneur de chaque mot est codé dans une étiquette de la forme ±Entier@POS. L'entier correspond à la distance relative du gouverneur en prenant uniquement compte les mots ayant la même partie de discours. Par exemple, `-3@NOM` signifie que le gouverneur du mot courant est le troisième nom avant le mot courant. Un exemple de cet étiquetage est présenté dans la figure 2. On note que cette méthode ne garantit pas que l'arbre prédit est un arbre en dépendances bien formé.

3 Jeu de données

Nous utilisons le corpus arboré CEFC-Orféo, un corpus du français annoté en dépendances et composé de plusieurs sous-corpus : CFPP (CLESTHIA, 2018), Clapi (ICAR, 2017) TCOF (ATILF, 2020), OFROM (Mathieu *et al.*, 2012 2020), Fleuron (André, 2016), French Oral Narrative (FON Carruthers, 2013), C-ORAL-ROM (Cresti *et al.*, 2004), Corpus de référence du français parlé (DELIC *et al.*, 2004), Valibel (Francard *et al.*, 2009), TUFS (Kawaguchi *et al.*, 2006), Corpus de réunions de travail (Husianycia, 2011), et mis à disposition avec les enregistrements audios. Ce corpus contient des types

Modèle	WER↓	CER↓	POS↑	UAS↑	LAS↑	Nombre de paramètres	Modèles préentraînés
AUDIO SEQ2LABEL	35.9	22.3	73.0	65.7	60.4	315M + 34.9M	Wav2vec2
AUDIO GRAPHE	**35.6**	**22.1**	**73.1**	**66.0**	**60.9**	315M + 34.9M	Wav2vec2
ORACLE SEQ2LABEL	36.3	22.2	75.6	68.7	62.7	315M + 34.9M	Wav2vec2
ORACLE GRAPHE	**35.6**	**22.2**	**77.4**	**73.3**	**67.5**	315M + 34.9M	Wav2vec2
CASCADE SEQ2LABEL	35.6	22.0	70.8	63.8	58.4	314M + 110M + 39.2M	Wav2vec2 + CamemBERT
CASCADE GRAPHE	35.6	22.0	69.3	60.5	53.1	314M + 110M + 41.4M	Wav2vec2 + CamemBERT
CASCADE HOPS	35.6	22.0	**72.4**	**65.8**	**61.0**	314M + 110M + 100M	Wav2vec2 + CamemBERT
TEXTE SEQ2LABEL	–	–	96.9	88.8	85.7	110M + 39.2M	CamemBERT
TEXTE GRAPHE	–	–	95.1	87.4	84.0	110M + 41.4M	CamemBERT
TEXTE HOPS	–	–	**98.2**	**90.3**	**87.7**	110M + 100M	CamemBERT

TABLE 1 – Évaluation sur le corpus de test d'Orféo avec les configurations décrites dans la section 4.

Modèle	WER↓	CER↓	POS↑	UAS↑	LAS↑	Nombre de paramètres	Modèles préentraînés
AUDIO SEQ2LABEL	31.0	18.4	**77.1**	70.2	65.2	315M + 34.9M	Wav2vec2
AUDIO GRAPHE	**30.6**	**18.2**	77.0	**70.9**	**66.2**	315M + 34.9M	Wav2vec2
ORACLE SEQ2LABEL	**30.9**	**18.6**	78.3	71.9	66.2	315M + 34.9M	Wav2vec2
ORACLE GRAPHE	31.4	19.2	**79.8**	**76.0**	**70.4**	315M + 34.9M	Wav2vec2
CASCADE SEQ2LABEL	30.5	18.2	74.7	67.7	62.4	314M + 110M + 39.2M	Wav2vec2 + CamemBERT
CASCADE GRAPHE	30.5	18.2	73.5	64.2	57.3	314M + 110M + 41.4M	Wav2vec2 + CamemBERT
CASCADE HOPS	30.5	18.2	**76.3**	**69.4**	**64.6**	314M + 110M + 100M	Wav2vec2 + CamemBERT
TEXTE SEQ2LABEL	–	–	94.5	86.7	83.1	110M + 39.2M	CamemBERT
TEXTE GRAPHE	–	–	96.8	88.3	84.5	110M + 41.4M	CamemBERT
TEXTE HOPS	–	–	**98.2**	**90.3**	**87.1**	110M + 100M	CamemBERT

TABLE 2 – Évaluation sur le corpus Valibel (un sous-corpus du jeu de test).

d'interactions variés, ayant toutes des spécificités de la parole spontanée, à l'exception de French Oral Narrative qui contient de la parole lue. Les situations de paroles sont également variées (ex : interactions commerciales, discussions informelles entre amis). Il s'agit du corpus de parole français le plus volumineux annoté en dépendances (196h, 2.5M tokens). Le schéma d'annotation (Benzitoun *et al.*, 2016) contient 20 parties du discours et 14 étiquettes de relations syntaxiques.

Les annotations syntaxiques d'Orféo ont été effectuées manuellement (données *gold* ou de référence) pour environ 5% du corpus, et automatiquement pour le reste (données *silver*). Les jeux d'entraînement/validation/test que nous utilisons dans cet article font en sorte que le jeu de validation ainsi que celui de test ne contiennent que des données annotées manuellement. Cependant, les sous-corpus avec les annotations de référence correspondent aussi à des enregistrements audios de qualité faible, ce qui rend la tâche difficile.

4 Expérimentations

Configurations expérimentales Nos expériences ont pour but de répondre aux questions suivantes :

— Est-ce qu'une approche bout-en-bout est préférable à une approche par cascade de systèmes ?
— Quel algorithme d'analyse (par graphe ou par étiquetage de séquence) est le plus approprié pour l'analyse syntaxique de la parole ?

	WER↓	CER↓	POS↑	UAS↑	LAS↑	Nombre de paramètres
Graph-tiny	35.7	22.3	73.0	65.9	60.8	314M + 11.7M
Graph-base	35.6	22.1	73.1	66.1	60.9	314M + 34.9M
Graph-large	35.6	22.0	73.2	66.0	60.7	314M + 67.6M

TABLE 3 – Comparaison des métriques d'analyse syntaxique avec l'architecture basée sur les graphes et différents nombres de paramètres.

— Est-ce qu'une approche est plus ou moins robuste au bruit inhérent du signal de parole ? Est-ce que les frontières de mot prédites sont un facteur important pour les différents systèmes ? Est-ce que les deux approches sont aussi robustes aux erreurs du module de RAP ?

Pour répondre à ces questions, nous comparons les configurations expérimentales suivantes, également illustrées par la figure 1 :

— **AUDIO** : Utilise uniquement **l'audio** comme entrée, le modèle crée des représentations vectorielles de mot depuis les représentations acoustiques comme décrit dans la section 2.
— **ORACLE** : Utilise **l'audio** ainsi que les **horodatages des mots** disponibles dans le corpus [3], ce qui rend la construction de vecteurs de mot plus facile et réduit l'impact de la qualité de la reconnaissance de la parole sur l'analyse syntaxique.
— **CASCADE** : Utilise seulement les **transcriptions prédites** par le modèle acoustique. Ensuite, un modèle de langue pré-entraîné calcule des vecteurs de mots contextualisés qui sont ensuite utilisés par le module d'analyse syntaxique. Dans cette configuration, nous modifions les arbres d'entraînement pour prendre en compte les suppressions ou insertions de mot. Cependant, comme pour l'approche basée sur l'audio, les insertions ou suppressions impactent le score global, car le score final est évalué sur les transcriptions manuelles et non sur une version modifiée. L'inconvénient de cette approche est que les informations prosodiques ne sont pas accessibles.
— **TEXTE** : Utilise les **transcriptions manuelles de référence** : cette configuration artificielle permet d'obtenir une borne supérieure pour les performances du modèle dans un cas idéal (RAP parfaite).

Les configurations **AUDIO** et **ORACLE** sont des modèles conjoints multitâches pour la RAP et l'analyse syntaxique. Comme les deux modules (RAP et analyse syntaxique) sont entraînés simultanément dans ces configurations, les résultats en RAP ne sont pas identiques d'une expérience à l'autre.

Les configurations **CASCADE** et **TEXTE** utilisent un modèle BERT du français : `camembert-base` [4] (Martin *et al.*, 2020) pour extraire des représentations vectorielles contextualisées. En plus de nos implémentations, nous utilisons également `hops` (Grobol & Crabbé, 2021), un analyseur syntaxique état-de-l'art. L'analyseur `hops` utilise également un bi-LSTM de caractères en plus de BERT pour produire des représentations vectorielles des mots, contrairement à nos implémentations qui sont conçues pour varier minimalement d'une configuration à l'autre.

Chaque méthode d'analyse pour chaque modalité est entraînée avec le même nombre d'époques, les même hyperparamètres (voir tables 4 et 5) et approximativement le même nombre de paramètres. Nos implémentations utilisent la bibliothèque `speechbrain` (Ravanelli *et al.*, 2021). La liste complète

3. Le corpus contient des horodatages construits automatiquement via un alignement forcé, c'est-à-dire que l'on dispose pour chaque mot des instants correspondant à son début et à sa fin.
4. https://huggingface.co/almanach/camembert-base

Analyseur	SEQ2LABEL	GRAPHE
Époques	30	30
Taille de *batch*	8	8
Paramètres d'optimisation		
Pas d'apprentissage	0.0001	0.0001
Algorithme d'optimisation	AdaDelta	AdaDelta
Model préentraîné	`LeBenchmark7K`	
Encodeur		
Nombre de couches	3	3
Dropout	0.15	0.15
Dimension de l'encodeur	1024	1024
Fonction d'activation	LeakyReLU	LeakyRelu
Fusion LSTM		
Nombre de couches	2	2
Dimension	500	500
Bidirectionnel	non	non
Biais	oui	oui
LSTM parser		
Nombre de couches	2	3
Dimensions	800	768
Bidirectionnel	oui	oui
Étiqueteur (SEQ2LABEL)		
Dimension	1600	
Nombre de couches	1	
Nombre d'étiquettes dep2label	846	
Nombre d'étiquettes POS	23	
Nombre d'étiquettes syntaxiques	19	
MLP pour les arcs (GRAPHE)		
Dimension		768
Nombre de couches		1
Taille de la couche de sortie		768
MLP pour les étiquettes syntaxiques (GRAPHE)		
Dimension		768
Nombre de couches		1
Taille de la couche de sortie		768
MLP pour les POS (GRAPHE)		
Dimension		768
Taille de la couche de sortie		24

TABLE 4 – Hyperparamètres pour les configurations **AUDIO** et **ORACLE** SEQ2LABEL et GRAPHE.

Analyseur	SEQ2LABEL	GRAPHE	HOPS
Époques	40	40	40
Taille de *batch*	32	32	32
Paramètres d'optimisation			
Pas d'apprentissage	0.001	0.001	0.00003
Algorithme d'optimisation	Adam	Adam	Adam
Modèle préentraîné		`camembert_base`	
Plongements lexicaux	dernière couche	dernière couche	Moy. de 12 couches
Taille des plongements	768	768	768
Bi-LSTM de caractères HOPS			
Taille des plongements			128
Plongements lexicaux HOPS			
Taille des plongements			256
LSTM parser			
Dimension	768	768	512
Nombre de couches	3	2	3
Bidirectionnel	oui	oui	oui
Étiqueteur (SEQ2LABEL)			
Dimension	1536		
Nombre de couches	1		
Nombre d'étiquettes dep2label	846		
Nombre d'étiquettes POS	23		
Nombre d'étiquettes syntaxiques	19		
MLP pour les arcs (GRAPHE and HOPS)			
Dimension		768	1024
Nombre de couches		1	2
Taille de la couche de sortie		768	768
MLP pour les étiquettes syntaxiques (GRAPHE)			
Dimension		768	1024
Nombre de couches		1	2
Taille de la couche de sortie		768	768
POS MLP (GRAPHE)			
Dimension		768	1024
Taille de la couche de sortie		24	24

TABLE 5 – Hyperparamètres pour les configurations **CASCADE** et **TEXTE** SEQ2LABEL, GRAPHE et **CASCADE**.

des hyperparamètres que nous utilisons est présentée dans les tableaux 4 et 5.

Métriques d'évaluation Nous utilisons les métriques d'évaluation classiques : *taux d'erreur mot* (`WER`) et *taux d'erreur caractère* (`CER`) pour la RAP, *exactitude des parties de discours* (`POS`), *Score d'attachement non étiqueté* (`UAS`), et *Score d'attachement étiqueté* (`LAS`) pour l'analyse en dépendances.

Les résultats sont présentés dans le tableau 1 pour tout le corpus, et dans le tableau 2 pour un sous-corpus du jeu de test (Valibel) pour lequel la RAP est moins difficile.

Résultats : effet de la reconnaissance de la parole sur la qualité de l'analyse syntaxique Dans le tableau 1, nous observons que les approches par graphe ou seq2label donnent des résultats similaires quand aucune information additionnelle n'est fournie, ce qui montre que le facteur limitant du modèle est la reconnaissance de la parole plutôt que l'analyse syntaxique.

Il est important de noter qu'étant donnée la nature du corpus (conversation spontanée), le `WER` est plus haut que ce que l'on pourrait typiquement attendre en reconnaissance de la parole sur des corpus classiques de cette tâche (qui contiennent plutôt de la parole lue). Par exemple, le module RAP de notre modèle atteint environ 8 `WER` quand il est entraîné et évalué sur CommonVoice5.1 (Ardila *et al.*,

2020). D'autres indices de cela sont montrés dans le tableau 3 : changer le nombre de paramètres du modèle basé sur les graphes ne modifie pas significativement les performances. De plus, dans le tableau 2, nous observons une claire amélioration des résultats quand on évalue sur un corpus avec des meilleures performances au niveau de la RAP. La qualité du module de RAP affecte directement le nombre de mots reconnus (certains peuvent être supprimés ou ajoutés), ce qui affecte la qualité de l'analyse.

Résultats : différence entre étiquetage de séquence et approche par graphe Il est relativement surprenant que sur la modalité textuelle (**CASCADE**), l'approche par étiquetage obtient de meilleures performances que l'approche par graphe étant donné que ce n'est pas le cas sur d'autres modalités (**AUDIO**). Ainsi, l'approche GRAPHE semble être plus sensible au bruit provenant de l'approche par **CASCADE** que l'approche SEQ2LABEL. Cependant, elle ne parvient pas à surpasser un modèle graphe plus large et plus complexe utilisant un bi-LSTM de caractères comme `hops`. Ce bi-LSTM de caractères réduit probablement l'impact des mots hors vocabulaire produits à cause des erreurs orthographiques de la RAP.

Sur les configurations **AUDIO** et **ORACLE**, le modèle basé sur les graphes obtient de meilleures performances, mais l'écart entre la configuration **AUDIO** et **ORACLE** est plus important que pour le modèle SEQ2LABEL. En effet, cet écart sur le LAS est de 7 points pour le modèle GRAPHE alors qu'il est de seulement 2 points pour SEQ2LABEL. Cela suggère que le modèle GRAPHE est plus efficace pour extraire des informations syntaxique directement depuis le signal, mais qu'il est plus sensible au bruit présent dans le signal de parole. On peut également faire l'hypothèse que le décodage global du système GRAPHE est plus adapté pour extraire les arbres syntaxiques. Le plus grand écart entre les deux approches a lieu quand plus d'informations à propos de la segmentation du signal audio sont données au modèle (**ORACLE**), réduisant l'influence de la RAP sur la qualité de l'analyse. On peut supposer que l'approche GRAPHE est plus adaptée pour l'analyse syntaxique directement depuis le signal de parole, particulièrement quand le `WER` baisse.

Transcrire puis analyser ou analyser directement ? L'approche **CASCADE** avec `hops` atteint des performances similaires à la version **AUDIO** avec notre analyseur par graphe. Cependant, le modèle cascade utilisant `hops` est un modèle plus complexe et n'est pas complètement comparable à notre analyseur. De plus, il possède environ 50% de paramètres en plus que notre modèle, nécessite 2 modèles pré-entraînés ; il est donc plus coûteux à entraîner.

Finalement, le tableau 2 montre que l'approche **AUDIO** surpasse l'approche **CASCADE** quand la qualité de la RAP augmente. Ce résultat suggère que l'analyse syntaxique profite de l'audio dès lors que la RAP atteint une qualité raisonnable.

Limitations Nous évaluons seulement nos analyseurs sur le français, grâce à la disponibilité d'un grand corpus arboré, ainsi nos conclusions doivent être interprétées dans ce cadre restreint. Nous comptons étendre ce travail à d'autres langues dans le futur.

Nous n'avons pas effectué une recherche poussée d'hyperparamètres, à cause du coût computationnel, mais nous avons dédié approximativement le même budget computationnel pour chaque modèle dans une configuration donnée.

5 Conclusion

Nous avons présenté un analyseur syntaxique par graphe prenant uniquement le signal audio en tant qu'entrée et avons évalué ses performances dans des expériences variées et avec plusieurs expériences contrôles. Nous montrons qu'un simple analyseur syntaxique basé sur les graphes avec wav2vec2 obtient des résultats similaires ou surpasse des modèles plus complexe nécessitant deux modèles pré-entraînés. Avec l'expérience de contrôle (**ORACLE**), nous montrons qu'obtenir des représentations vectorielles de mot de bonne qualité est le défi principal pour l'analyse syntaxique de la parole. Nous supposons également que l'approche par GRAPHE est plus adapté pour l'analyse syntaxique directement depuis le signal de parole qu'une approche par étiquetage de séquence. Nous concentrerons nos efforts futurs sur l'amélioration de la qualité de la segmentation du signal audio pour pallier ces problèmes.

Remerciements

Nous remercions les relecteurices anonymes pour leur remarques et suggestions. Ce travail a bénéficié d'un financement du Fond National Suisse (No. 197864) et de l'Agence Nationale de la Recherche, via les projet PROPICTO (ANR-20-CE93-0005) et SynPaX (ANR-23-CE23-0017). Ce travail a également bénéficié d'un accès aux moyens de calcul de l'IDRIS via l'allocation de ressources AD011013463R1 attribuée par GENCI.

Références

ANDRÉ V. (2016). Fleuron : Français langue Étrangère universitaire–ressources et outils numériques.

ARDILA R., BRANSON M., DAVIS K., KOHLER M., MEYER J., HENRETTY M., MORAIS R., SAUNDERS L., TYERS F. & WEBER G. (2020). Common voice : A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4218–4222, Marseille, France : European Language Resources Association.

ATILF (2020). Tcof : Traitement de corpus oraux en français. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, **33**, 12449–12460.

BENZITOUN C., DEBAISIEUX J.-M. & DEULOFEU H.-J. (2016). Le projet orféo : un corpus d'étude pour le français contemporain. *Corpus*, (15).

CARRUTHERS J. (2013). French oral narrative corpus. Commissioning Body / Publisher : Oxford Text Archive.

CHARNIAK E. & JOHNSON M. (2001). Edit detection and parsing for transcribed speech. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

CHRUPAŁA G. (2023). Putting natural in natural language processing. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Findings of the Association for Computational Linguistics : ACL 2023*, p. 7820–7827, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.495.

CLESTHIA (2018). Cfpp2000. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

CRESTI E., DO NASCIMENTO F. B., SANDOVAL A. M., VERONIS J., MARTIN P. & CHOUKRI K. (2004). The c-oral-rom corpus. a multilingual resource of spontaneous speech for romance languages. p. 26–28.

DELIC E., TESTON-BONNARD S. & VÉRONIS J. (2004). Présentation du corpus de référence du français parlé. *Recherches sur le français parlé*, **18**, 11–42. Equipe DELIC, HAL : halshs-01388193.

DOZAT T. & MANNING C. D. (2016). Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

FRANCARD M., HAMBYE P., SIMON A.-C. & DISTER A. (2009). Du corpus à la banque de données. : Du son, des textes et des métadonnées. l'évolution de banque de données textuelles orales valibel (1989-2009). *Cahiers de l'Institut de linguistique de Louvain-CILL*, **33**(2), 113.

GÓMEZ-RODRÍGUEZ C., STRZYZ M. & VILARES D. (2020). A unifying theory of transition-based and sequence labeling parsing. In D. SCOTT, N. BEL & C. ZONG, Éds., *Proceedings of the 28th International Conference on Computational Linguistics*, p. 3776–3793, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.336.

GRAVES A., FERNÁNDEZ S., GOMEZ F. & SCHMIDHUBER J. (2006). Connectionist temporal classification : Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, p. 369–376, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1143844.1143891.

GROBOL L. & CRABBÉ B. (2021). Analyse en dépendances du français avec des plongements contextualisés (French dependency parsing with contextualized embeddings). In P. DENIS, N. GRABAR, A. FRAISSE, R. CARDON, B. JACQUEMIN, E. KERGOSIEN & A. BALVET, Éds., *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 106–114, Lille, France : ATALA.

HONNIBAL M. & JOHNSON M. (2014). Joint incremental disfluency detection and dependency parsing. *Transactions of the Association for Computational Linguistics*, **2**, 131–142. DOI : 10.1162/tacl_a_00171.

HUSIANYCIA M. (2011). *Caractérisation de types de discours dans des situations de travail*. Theses, Université Nancy 2. HAL : tel-01749085.

ICAR (2017). Clapi. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

JAMSHID LOU P., WANG Y. & JOHNSON M. (2019). Neural constituency parsing of speech transcripts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 2756–2765, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1282.

JOHNSON M. & CHARNIAK E. (2004). A TAG-based noisy-channel model of speech repairs. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, p. 33–39, Barcelona, Spain. DOI : 10.3115/1218955.1218960.

KAWAGUCHI Y., ZAIMA S. & TAKAGAKI T., Éds. (2006). *Spoken Language Corpus and Linguistic Informatics*. John Benjamins.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MATHIEU A., MARIE-JOSÉ B., GILLES C., FEDERICA D. & ANNE J. L. ((2012-2020)). Corpus ofrom – corpus oral de français de suisse romande. Université de Neuchâtel.

PARCOLLET T., NGUYEN H., EVAIN S., ZANON BOITO M., PUPIER A., MDHAFFAR S., LE H., ALISAMIR S., TOMASHENKO N., DINARELLI M., ZHANG S., ALLAUZEN A., COAVOUX M., ESTÈVE Y., ROUVIER M., GOULIAN J., LECOUTEUX B., PORTET F., ROSSATO S., RINGEVAL F., SCHWAB D. & BESACIER L. (2024). Lebenchmark 2.0 : A standardized, replicable and enhanced framework for self-supervised representations of french speech. *Computer Speech Language*, **86**, 101622. DOI : https ://doi.org/10.1016/j.csl.2024.101622.

PRICE P. J., OSTENDORF M., SHATTUCK-HUFNAGEL S. & FONG C. (1991). The use of prosody in syntactic disambiguation. *the Journal of the Acoustical Society of America*, **90**(6), 2956–2970.

PUPIER A., COAVOUX M., LECOUTEUX B. & GOULIAN J. (2022). End-to-End Dependency Parsing of Spoken French. In *Proc. Interspeech 2022*, p. 1816–1820. DOI : 10.21437/Interspeech.2022-381.

RASOOLI M. S. & TETREAULT J. (2013). Joint parsing and disfluency detection in linear time. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 124–129, Seattle, Washington, USA : Association for Computational Linguistics.

RAVANELLI M., PARCOLLET T., PLANTINGA P., ROUHE A., CORNELL S., LUGOSCH L., SUBAKAN C., DAWALATABAD N., HEBA A., ZHONG J., CHOU J.-C., YEH S.-L., FU S.-W., LIAO C.-F., RASTORGUEVA E., GRONDIN F., ARIS W., NA H., GAO Y., MORI R. D. & BENGIO Y. (2021). SpeechBrain : A general-purpose speech toolkit. arXiv :2106.04624.

STRZYZ M., VILARES D. & GÓMEZ-RODRÍGUEZ C. (2019). Viable dependency parsing as sequence labeling. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 717–723, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1077.

STRZYZ M., VILARES D. & GÓMEZ-RODRÍGUEZ C. (2020). Bracketing encodings for 2-planar dependency parsing. In D. SCOTT, N. BEL & C. ZONG, Éds., *Proceedings of the 28th International Conference on Computational Linguistics*, p. 2472–2484, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.223.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Vers la traduction automatique des néologismes scientifiques

Paul Lerner François Yvon
Sorbonne Université, CNRS, ISIR, 75005, Paris, France
lerner@isir.upmc.fr, yvon@isir.upmc.fr

RÉSUMÉ

La recherche scientifique découvre et invente continuellement de nouveaux concepts qui sont alors désignés par de nouveaux termes, des néologismes, ou *néonymes* dans ce contexte. Puisque les publications se font très majoritairement en anglais, diffuser ces nouvelles connaissances en français demande souvent de traduire ces termes, afin d'éviter de multiplier les anglicismes qui sont moins facilement compréhensibles pour le grand public. Nous proposons d'explorer cette tâche à partir de deux thésaurus en exploitant la définition du terme afin de le traduire plus fidèlement. Pour ce faire, nous explorons les capacités de deux grands modèles de langue multilingues, BLOOM et CroissantLLM, qui parviennent à traduire des néologismes scientifiques dans une certaine mesure. Nous montrons notamment qu'ils utilisent souvent des procédés morphosyntaxiques appropriés mais sont limités par la segmentation en unités sous-lexicales et biaisés par la fréquence d'occurrences des termes ainsi que par des similarités de surface entre l'anglais et le français.

ABSTRACT

Towards Machine Translation of Scientific Neologisms

Scientific research continually discovers and invents new concepts, which are then referred to by new terms, neologisms, or *neonyms* in this context. As the vast majority of publications are written in English, disseminating this new knowledge in French often requires translating these terms, to avoid multiplying anglicisms that are less easily understood by the general public. We propose to explore this task using two thesauri, exploiting the definition of the term to translate it more accurately. To this end, we explore the capabilities of two large multilingual models, BLOOM and CroissantLLM, which can translate scientific terms to some extent. In particular, we show that they often use appropriate morphological procedures, but are limited by the segmentation into sub-lexical units. They are also biased by the frequency of term occurrences and surface similarities between English and French.

MOTS-CLÉS : néologisme, terminologie, morphologie, traduction automatique.

KEYWORDS: neologism, terminology, morphology, machine translation.

1 Introduction

De nouveaux concepts sont continuellement découverts et inventés par des chercheurs du monde entier, ce qui mène à une prolifération de néologismes. Cabré (1999) parle alors de *néonymes*, par opposition aux néologismes du langage courant (Cartier *et al.*, 2018). L'immense majorité des publications scientifiques se font en anglais (Gordin, 2015; Larivière & Riddles, 2021) [1], ce qui pose

1. Il subsiste une part importante de publications en français en sciences humaines et sociales.

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 245–261.

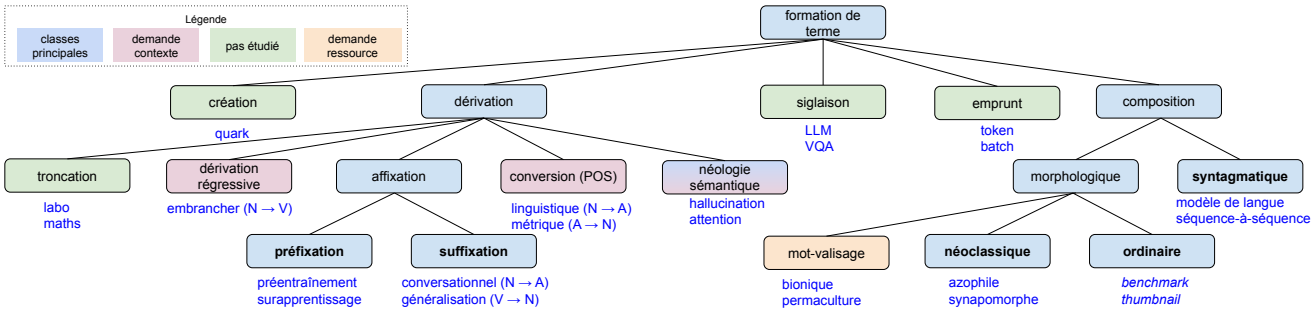


FIGURE 1 – Aperçu des procédés néologiques étudiés dans ce travail. D'après Daille (2017).

problème en particulier dans une optique de diffusion des connaissances dans la société [2]. Les termes scientifiques sont produits et construits en anglais, ce qui amène à multiplier les anglicismes pour communiquer et diffuser les savoirs vers des publics non-spécialistes. Par exemple, un enseignant de Traitement Automatique des Langues (TAL) pourrait préférer « le grand modèle de langue GPT-3 apprend une nouvelle tâche grâce au contexte sans ajuster ses paramètres » à « le *large language model* GPT-3 *learn* une nouvelle *task* grâce au *in-context learning* sans *fine-tuner* ses paramètres » afin de rendre son cours plus intelligible. Pour citer Liu *et al.* (2021) : « la définition précise de la terminologie est la première étape de la communication scientifique. » [3]

Notre principale question de recherche est : comment traduire automatiquement un néologisme scientifique ? Par définition, il n'existe pas de données parallèles où trouver des exemples de traduction. Nous proposons donc d'exploiter la définition des termes afin de les traduire plus fidèlement. Nous étudions alors comment exploiter cette définition et sa complémentarité au terme source. Enfin, puisque les termes peuvent être formés par divers procédés non-exclusifs (par ex. la préfixation, suffixation, composition néoclassique, ordinaire ou syntagmatique), qui peuvent différer de la langue source à la langue cible, nous analysons : (i) l'impact de cette différence sur la qualité des traductions ; (ii) les procédés utilisés lors de la génération.

À notre connaissance, le seul travail à envisager une approche similaire est (Zhang *et al.*, 2020), qui est limité au domaine très spécifique de la génétique, où un terme décrit la fonction du gène et est lié à plusieurs gènes selon sa fonction moléculaire, son processus biologique et sa composante cellulaire. Dans notre étude, nous expérimentons avec des thésaurus qui fournissent directement la définition du terme, en laissant l'extraction automatique des définitions à partir des publications (Jin *et al.*, 2013; Head *et al.*, 2021; August *et al.*, 2022; Huang *et al.*, 2022) pour des travaux futurs. Une telle extraction serait d'autant plus nécessaire que certains termes ne sont utilisés que dans l'article les définissant ou les quelques articles s'y référant. Ce nouveau problème se rapproche de l'extraction de termes multilingues (Laroche & Langlais, 2010; Delpech *et al.*, 2012; Rigouts Terryn *et al.*, 2020), à l'importante exception près que nous ne supposons pas que le terme cible apparaisse, même une seule fois, dans un corpus comparable. En effet, nous verrons qu'une part importante des termes étudiés n'apparaissent pas une seule fois, même dans le gigantesque corpus OSCAR (Abadji *et al.*, 2022).

Nous explorons ce nouveau problème en testant les capacités de deux grands modèles de langues (*Large Language Model* ; LLM) multilingues, BLOOM (BigScience & et al., 2023) et CroissantLLM (Faysse *et al.*, 2024). Nous montrons que ces modèles sont capables, dans une certaine mesure, de traduire des termes isolés de l'anglais vers le français, mais surtout, de générer un terme à partir

2. Cf. l'initiative d'Helsinki : https://www.helsinki-initiative.org/
3. « *Precisely defining the terminology is the first step in scientific communication.* »

de sa définition et de combiner les deux sources d'information. Toutefois, ces modèles traduisent mieux les termes français proches des termes anglais. Nous étudions deux types de similarité : morphosyntaxique et surfacique (distance d'édition). La similarité surfacique révèle souvent des cognats ou des emprunts, pour lesquels la traduction se rapproche davantage d'une translittération (*exocytosis* → exocytose ; Claveau & Zweigenbaum, 2005). Par ailleurs, nos résultats suggèrent une corrélation négative entre la fertilité des termes et la performance des modèles, notamment pour les termes préfixés qui sont sursegmentés par le tokeniseur BPE (Gage, 1994). Enfin, il est souvent difficile de classer objectivement un terme comme néologique ou lexicalisé (Lombard & Huyghe, 2020). Certains termes de thésaurus sont plus fréquents en corpus et sont alors mieux traduits par les modèles.

Cette étude ouvre donc un nouveau défi pour le TAL, sur une thématique importante pour la diffusion des savoirs en français, et sur laquelle il reste beaucoup à faire, comme nous l'évoquons en conclusion. Notre code est disponible à l'adresse : `https://github.com/PaulLerner/neott`.

2 Procédés néologiques et morphologiques

Notre typologie des néologismes, d'après Lieber (2010) et Daille (2017), repose sur des traits morphosyntaxiques qui peuvent facilement être détectés automatiquement. D'autres typologies existent, voir par exemple (Lombard & Huyghe, 2020). Nous avons retenu les cinq procédés suivants : (i) la **préfixation**, où un affixe est concaténé au début d'un mot pour en former un nouveau (*pré+entraînement* = *préentraînement*) ; (ii) la **suffixation**, où l'affixation se fait à la fin du mot (*généraliser+tion* = *généralisation*) ; (iii) la **composition ordinaire** (*native compounding*), qui compose deux mots indépendants, est plus fréquente en anglais (*bench+mark* = *benchmark*) qu'en français (où elle se développe cependant de plus en plus ; Arnaud, 2003) ; (iv) la **composition néoclassique** (ou savante), qui compose uniquement des morphèmes liés (*bound morphemes*), c'est-à-dire qui ne peuvent agir comme mots indépendants (*azo+phile* = *azophile*)[4] ; (v) enfin, la **composition syntagmatique**, où des syntagmes qui suivent les règles syntaxiques de la langue se lexicalisent et donnent lieu à des termes, souvent non-compositionnels. Par exemple, *modèle de langue* a pris un sens bien plus spécifique que la simple somme du sens individuel de chacun de ses composants. De plus, remarquons que l'insertion n'est pas possible à l'intérieur d'un terme : on dira « modèle de langue *préentraîné* » et non pas « *modèle *préentraîné* de langue ». D'autre part, ces figements empêchent souvent une traduction compositionnelle (littéralement, mot-à-mot), par exemple *low-resource language* → langue peu dotée. Également, l'anglais modifie fréquemment les noms avec d'autres noms pour former des syntagmes (*language* modifie *model* dans *language model* ; Biber *et al.*, 2010). Ces formes sont alors souvent traduites N P N (*language model* → modèle de langue ; Isabelle *et al.*, 2017) ou N A (*machine translation* → traduction automatique) en français[5].

Remarquons qu'il y a souvent une adaptation phonologique des morphèmes à leur jonction, et non pas une simple concaténation. Remarquons également que plusieurs procédés peuvent être cumulés, ainsi *surapprentissage* est une préfixation (*sur-*) d'une suffixation (*-age*). Ces procédés sont représentés à la figure 1, de laquelle sont exclus les flexions, qui ne créent pas de nouveaux lexèmes.

4. Remarquons que, contrairement à la grammaire française, l'élément recteur est à droite (*azophile* se dit d'un composé qui présente une affinité pour un atome d'azote, et pas d'un atome d'azote amoureux ; Namer, 2003 ; Amiot & Dal, 2008).

5. Dans les cas plus rares où le français conserve la formation N N, l'ordre est inversé pour garder la tête à gauche (*source language* → langue source).

Cette figure inclut également des procédés qui ne sont pas pris en compte par notre analyse [6] : (i) La **néologie sémantique**, où une unité lexicale est associée à un nouveau concept, créant ainsi un homonyme, souvent par transfert métaphorique d'un domaine source à un domaine cible. Par exemple, *hallucination* est désormais utilisé en TAL pour désigner des générations infondées de modèles, par analogie avec les hallucinations humaines [7]. Certains changements sémantiques suivent une régularité métaphorique (Lombard *et al.*, 2023). Par exemple, les parties du corps humain sont souvent utilisées pour désigner une partie d'un objet selon sa position (*tête d'attention*). Nous étudierons ce phénomène, non pas selon la forme du terme, puisqu'elle ne change pas, ni par son contexte, dont nous ne disposons pas, mais par la fréquence du terme dans un corpus. (ii) La **conversion**, ou changement de catégorie morphosyntaxique (Tribout, 2010), résultant également en un homonyme. Par exemple, « une métrique neuronale » où *métrique* est utilisé comme un nom et non comme un adjectif. Nous ne pouvons étudier ce phénomène faute de contexte pour les termes. (iii) La **dérivation régressive** (*back-affixation*) qui demande une perspective diachronique pour la différencier des autres affixations (*embranchement - ment = embrancher*). (iv) Le **mot-valisage**, qui compose deux lexèmes tronqués (*biologie + électronique = bionique*). Nous ne disposons pas de ressource pour ce procédé par ailleurs relativement rare (Cartier *et al.*, 2018).

Nous n'étudions pas les quatre procédés suivants, bien qu'ils soient fréquents en anglais et en français : (i) les **emprunts**, que nous cherchons justement à éviter (*token* est calqué tel quel depuis l'anglais) [8] ; (ii) les **troncations**, qui sont plus souvent utilisées dans un langage courant voire familier, mais sont moins présentes dans les publications scientifiques (*labo = laboratoire*) ; (iii) les **créations** (*coinage*), très rares et qui ont allure de nom propre, donc ne doivent pas être traduites (*quark* qui provient de Joyce ; Gell-Mann, 1964) ; (iv) les **sigles et acronymes**, pour la même raison (à l'exception d'un éventuel réordonnancement de leurs lettres).

Nous renvoyons finalement vers (Dal, 2003b), (Lieber, 2010) ou (Corbin, 2012) pour une introduction plus complète à la morphologie [9], couvrant d'autres langues que l'anglais et le français, et donc, d'autres procédés (par exemple les *transfixes* dans les langues sémitiques).

3 Méthodes

Nous étudions trois approches pour traduire des néologismes scientifiques, sachant que la langue cible est toujours le français : (i) traduire le terme anglais isolé, ce qui n'est pas notre intérêt premier mais sert de point de référence ; (ii) générer le terme à partir de sa définition (en français également), la principale nouveauté que nous proposons ; (iii) générer le terme à partir du terme anglais et de sa définition en français, c'est-à-dire en combinant les deux sources d'information.

Fort heureusement, nous pouvons traiter ces trois sous-tâches de la même manière dans un cadre de génération de texte (Raffel *et al.*, 2020; Brown *et al.*, 2020). Nous adoptons l'approche maintenant standard qui consiste à laisser un LLM compléter une amorce (*prompt*). L'amorce contiendra donc : (i) le terme anglais ; (ii) la définition du terme ; (iii) les deux. Remarquons que ces trois approches sont

6. La troncation et le mot-valisage, de par leur irrégularité, ne sont habituellement pas classés comme dérivation et composition, respectivement. Toutefois, nous préférons organiser les procédés de façon hiérarchique.

7. L'analogie est souvent considérée de façon orthogonale aux autres procédés néologiques (Dal, 2003a; Mattiello, 2017).

8. Les traductions littérales sont habituellement considérés comme des emprunts mais, étant donné l'écrasante majorité de publication scientifique anglophone, donc de création et de définition de terme en anglais, nous acceptons les traductions littérales et évitons seulement d'utiliser des mots anglais.

9. Voir aussi (Aronoff, 1976) et (Fradin, 2015) pour une approche lexématique de la morphologie.

Langue	Formulation	Patron de l'amorce
EN	VERSION	`If the original version says {src_term} then the French version should say :`
EN	TERM	`The term {src_term} can be translated in French as :`
EN	TATOEBA_MT	`Translate the following term from English to French {src_term} :`
FR	TERM	`Le terme anglais {src_term} peut se traduire en français par :`
FR	DEF	`{src_def} définit le terme :`
FR	DEF+TERM	`{src_def} définit le terme anglais {src_term} qui peut se traduire en français par :`
FR	TATOEBA_MT	`Traduis le terme anglais suivant en français {src_term} :`

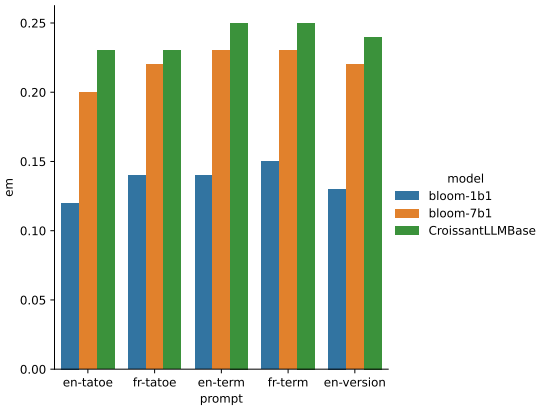
TABLE 1 – Les différentes amorces utilisées pour les modèles décodeurs.

donc : (i) translingue (traduction classique) ; (ii) monolingue (français seulement) ; (iii) multilingue (anglais et français mélangés). Ces différents scénarios impliquent d'utiliser des modèles multilingues.

Un autre point de comparaison pour (i) repose sur un scénario de traduction classique utilisant ici un modèle de traduction neuronal séquence-à-séquence dérivé de mBART50-One-to-Many (610M de paramètres ; Tang *et al.*, 2021). Ce modèle est ajusté avec 1,1 million de phrases parallèles (EN-FR) extraites du corpus SciPar (Roussis *et al.*, 2022) afin de le rendre robuste au vocabulaire scientifique. Il atteint 37,3 BLEU sur un jeu d'évaluation de 3 000 phrases (voir Peng *et al.* (2024) pour une présentation détaillée). Contrairement aux LLM, ce modèle ne prend pas en charge des entrées bilingues qui associeraient terme (en anglais) et définition (en français).

Implémentation Nous expérimentons avec deux modèles multilingues : BLOOM (BigScience & et al., 2023) et CroissantLLM (Faysse *et al.*, 2024). BLOOM fut le premier modèle multilingue librement disponible à franchir le seuil des milliards de paramètres. C'est un modèle entraîné sur de nombreuses langues, dont l'anglais et le français, qui est très efficace pour la traduction et pour diverses tâches de TAL en français (Bawden & Yvon, 2023 ; Bawden *et al.*, 2024). Nous expérimenterons avec deux versions de BLOOM, comprenant respectivement 1,1G et 7,1G de paramètres. CroissantLLM est un modèle bilingue (anglais-français) ouvert, entraîné sur une quantité égale de données dans les deux langues. Plus compact (1,3G de paramètres), il a été conçu pour être efficace à l'inférence pour amortir son coûteux pré-entraînement, dans la lignée de (Liu *et al.*, 2019) et (Hoffmann *et al.*, 2022). Les différentes formulations d'amorce sont listées dans le tableau 1. L'amorce peut être formulée en anglais et en français, mais le terme source est toujours en anglais et la définition est toujours en français. Les formulations VERSION et TATOEBA_MT ont été adaptées d'après (Bawden & Yvon, 2023) et (Muennighoff *et al.*, 2023), respectivement. Les autres formulations ont été inventées pour le besoin de ce travail. Ces variantes seront validées sur le jeu de validation mais nous verrons que la formulation a peu d'impact sur les performances puisque nous utilisons systématiquement cinq exemples aléatoires dans l'amorce comme contexte (*in-context learning* ; ICL). Remarquons que les formulations (i) TERM, (ii) DEF et (iii) DEF+TERM ont des entrées différentes qui correspondent aux trois approches proposées. Les exemples sont séparés par les trois caractères ###.

Évaluation Nous évaluons principalement les modèles avec des métriques standard en question-réponse (Rajpurkar *et al.*, 2016) : (i) la correspondance exacte (*exact match* ; EM) entre les chaînes de caractères prédite et attendue ; (ii) le score F1 au niveau du token ; après un prétraitement simpliste (insensible à la casse, filtrage des mots vides et de la ponctuation). Nous évaluons également la capacité des modèles à prédire une forme néologique adéquate.



(a)

Modèle	Entrée	FranceTerme		TERMIUM	
		EM	F1	EM	F1
mBART	TERM	26,3	41,3	36,8	49,4
BLOOM-1,1G	TERM	15,9	31,3	22,7	35,0
BLOOM-1,1G	DEF	1,1	11,3	3,8	9,8
BLOOM-1,1G	DEF+TERM	17,8	34,9	25,3	38,2
BLOOM-7,1G	TERM	23,7	40,3	38,8	50,9
BLOOM-7,1G	DEF	10,0	24,7	13,6	23,6
BLOOM-7,1G	DEF+TERM	**27,1**	**44,6**	**41,8**	**54,6**
CroissantLLM	TERM	25,6	42,2	41,6	53,9
CroissantLLM	DEF	4,6	19,8	7,2	16,5
CroissantLLM	DEF+TERM	25,3	42,9	38,7	51,6

(b)

FIGURE 2 – (a) Correspondance exacte (EM) selon les formulations de l'amorce sur le jeu de validation de FranceTerme (gauche). (b) Résultats de tous les modèles selon les différentes entrées sur les jeux de test de FranceTerme et TERMIUM. Les meilleurs résultats sont en gras, et les meilleurs pour chaque type d'entrée sont soulignés.

4 Résultats

Jeux de données Nous exploitons deux thésaurus bilingues (anglais / français) dans ce travail : FranceTerme [10] et TERMIUM [11], fournis par le gouvernement français et canadien, respectivement. Pour TERMIUM, nous limitons ici notre analyse au sous-domaine "symptômes" (biomédical). Afin de filtrer les emprunts (cf. section 2), nous filtrons les termes qui sont identiques en anglais et en français (insensible à la casse). Pour filtrer les sigles et acronymes, nous filtrons les termes comprenant plus de deux lettres majuscules successives. Nous conservons uniquement les entrées auxquelles sont associées une version anglaise, française et une définition en français. Après filtrage, FranceTerme est réduit à 6 623 termes que nous divisons aléatoirement en jeu de validation et de test de taille égale. TERMIUM-Symptômes ne contient que 1 608 termes donc nous l'utilisons seulement pour le test (sans ajuster aucun paramètre ou hyperparamètre).

Différentes amorces pour des résultats similaires Nous commençons par valider les différentes amorces sur le jeu de validation de FranceTerme (figure 2a). Pour tous les modèles, la meilleure amorce est TERM, formulée en français, que nous utilisons dans la suite des expériences. De façon générale, toutes les amorces donnent des résultats proches, ce que nous attribuons à l'utilisation de plusieurs exemples en contexte (Bawden & Yvon, 2023). Les préférences entre les formulations sont plutôt intuitives : le français est préféré à l'anglais et l'amorce TERM, d'un style « test de complétion » (*cloze test*) est préféré à TATOEBA_MT, d'un style « instruction » et ce pour tous les modèles.

Performances générales Les résultats principaux sont présentés à la figure 2b. Sur tous les jeux de données, les modèles qui prennent uniquement le terme en entrée surpassent ceux qui prennent seulement la définition. Nous trouvons cependant que la performance de ces modèles semble limitée, mBART, BLOOM-7,1G et CroissantLLM obtenant tous trois des résultats similaires. Sur tous les jeux

10. https://www.culture.fr/franceterme, version du 17 novembre 2023.
11. https://www.btb.termiumplus.gc.ca/, version du 6 février 2023.

Modèle	Ordinaire	Néo.	Pré.	Suff.	Synt.
anglais A	5,8	26,7	52,5	67,4	87,4
TERM	13,5	57,3	71,3	85,6	87,2
DEF	19,6	36,6	59,1	81,9	77,5
DEF+TERM	14,9	55,9	73,5	87,0	87,5

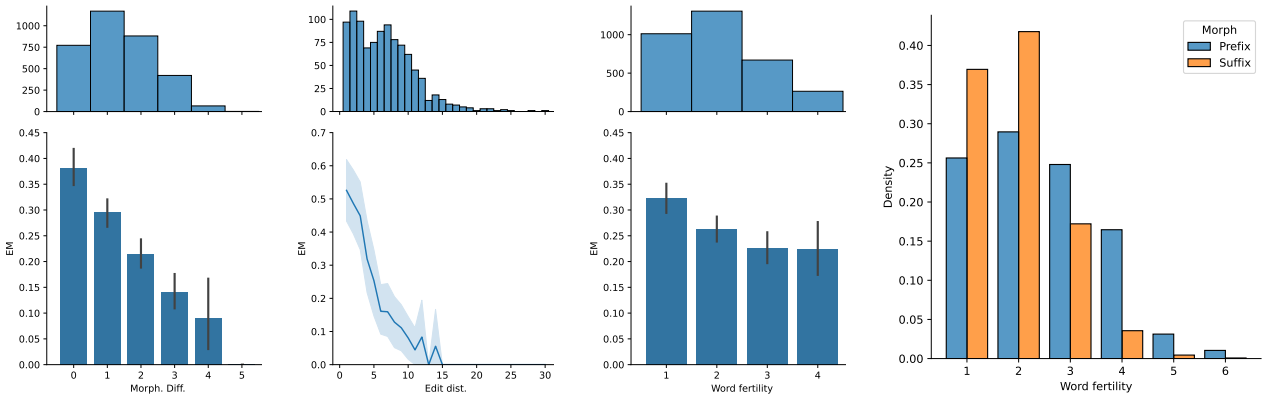
TABLE 2 – Pouvoir de prédiction (F1) des procédés morphosyntaxiques du terme français F selon les modèles de BLOOM-7,1G (et comparé aux procédés morphosyntaxiques du terme anglais A) sur le jeu de validation de FranceTerme.

de données nous trouvons que BLOOM-7,1G parvient à combiner les informations provenant du terme anglais et de la définition française et surpasse largement la version TERM. BLOOM-7,1G parvient même à légèrement surpasser une fusion oracle tardive des modèles TERM et DEF, ce qui suggère une interaction positive entre ces deux informations. Par exemple, BLOOM-7,1G DEF+TERM parvient à correctement prédire *capteur de mission* pour *mission sensor* « capteur réalisant des mesures qui font partie de l'objet de la mission d'un engin spatial », contrairement à TERM qui prédit *mission de reconnaissance* et DEF qui prédit *instrument de mesure*. Nous avons vérifié que la performance de ce modèle était stable en faisant varier les exemples de l'amorce aléatoirement. Comme souvent, la taille du modèle semble enfin importante pour cette tâche, notamment pour traiter les définitions, puisque les performances de BLOOM-1,1G (DEF+TERM), comme celles de CroissantLLM dépassent à peine, voire sont moindres que pour les versions TERM seul.

Classification morphosyntaxique Nous construisons un classifieur multi-étiquettes pour quatre des cinq classes définies à la section 2 : préfixation, suffixation, composition néoclassique ou ordinaire. Pour la cinquième (composition syntagmatique), nous nous reposons sur la simple heuristique du nombre de mots segmentés par spaCy (Honnibal *et al.*, 2020). S'il y a plusieurs mots, nous considérons que le terme est un syntagme (possiblement lexicalisé). Pour détecter ces quatre procédés morphologiques, nous utilisons l'architecture de FastText (Joulin *et al.*, 2017). Dans notre utilisation, ce classifieur est entraîné en mode « un contre tous » (*one versus all*), équivalent à un classifieur binaire pour chacune des classes identifiées supra. Le classifieur est entraîné sur les bases étymologiques MorphyNet (Batsuren *et al.*, 2021) et de celle utilisée pour la *shared task* SIGMORPHON 2022 (Batsuren *et al.*, 2022), toutes deux extraites depuis le Wiktionnaire anglais [12]. Nous vérifions que ce classifieur est efficace sur un ensemble d'évaluation où il arrive à 92,5 de F1 en anglais et 95,8 en français. Ce classifieur est décrit plus précisément à l'annexe A.

Impact sur la génération Pour mesurer l'impact de la morphosyntaxe des termes sur nos modèles génératifs, nous mesurons la différence symétrique Δ entre les procédés morphosyntaxiques du terme anglais A et français F (prédites par notre classifieur multi-étiquettes) : $\Delta = |(A \setminus F) \cup (F \setminus A)|$. La figure 3a montre que les prédictions de BLOOM-7,1G (DEF+TERM) sont bien plus souvent correctes lorsque les termes anglais et français ont des morphosyntaxes proches (identiques ou différant seulement d'un procédé). Même lorsque les morphosyntaxes sont différentes, nous observons que les modèles prédisent souvent le bon procédé morphosyntaxique, cf. le tableau 2. Par exemple, BLOOM prédit bien un composé néoclassique avec *polyactif* alors qu'il devrait produire *pluriactif* (EN : *slasher*). Notons qu'il n'est pas surprenant que les composés ordinaires anglais aident peu à prédire

12. https://en.wiktionary.org/



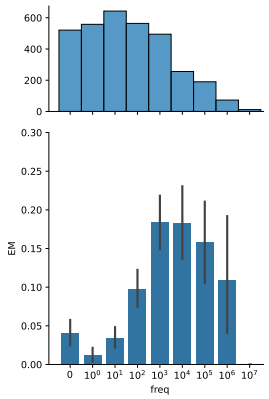
(a) (b) (c) (d)

FIGURE 3 – (a) Exactitude de la prédiction de BLOOM-7,1G (DEF+TERM) selon la différence symétrique Δ entre les procédés morphosyntaxiques du terme anglais A et français F. (b) Exactitude de la prédiction de BLOOM-7,1G (TERM) selon la distance d'édition entre les termes anglais et français monolexicaux. (c) exactitude de la prédiction de BLOOM-7,1G (DEF+TERM) sur tous les termes selon la fertilité des mots. (d) fertilité des mots selon que le terme soit préfixé ou suffixé. Les distributions sont normalisées séparément (comme s'il y avait autant de préfixes que de suffixes) afin de faciliter la visualisation. Tous les résultats proviennent du jeu de validation de FranceTerme.

leurs équivalents français (rappel de 51% mais précision de 3%) puisque ce procédé est très rare en français.

Traduction ou translittération ? Nous avons vu plus haut que les modèles parvenaient beaucoup mieux à générer le terme français à partir du terme anglais que de sa définition. Ce résultat est dû pour partie à la similarité de surface entre un grand nombre de termes anglais et français, pour lesquels la traduction s'apparente à une translittération. Pour quantifier ce phénomène, nous étudions la distance d'édition entre les termes anglais et français monolexicaux (l'ordre des mots étant rarement le même en anglais et en français, la distance d'édition n'est pas fiable pour les termes polylexicaux). La figure 3b montre que le modèle BLOOM-7,1G (TERM) traduit beaucoup mieux les termes qui diffèrent de trois caractères ou moins (par ex., *mycotoxin* $\rightarrow$ mycotoxine, *exocytosis* $\rightarrow$ exocytose, *iconomatic* $\rightarrow$ iconomatique). Le modèle qui prédit à partir de la définition du terme ne montre pas cette tendance. Ce phénomène explique, au moins partiellement, la différence de performances entre FranceTerme et TERMIUM-Symptômes (figure 2b), où les termes anglais et français sont souvent proches avec une distance d'édition médiane de 2 pour les termes monolexicaux contre 6 pour FranceTerme. La distance minimale est de 1 puisque nous avons préalablement filtré les termes identiques.

BPE et fertilité des termes, en relation avec la préfixation BLOOM et CroissantLLM utilisent tous deux la tokenisation BPE, comme la majorité des LLM (Gage, 1994; Sennrich *et al.*, 2016). Cette méthode permet de décomposer en unités sous-lexicales les mots inconnus ou trop rares pour bénéficier d'une représentation dédiée. Toutefois, elle n'est pas fondée morphologiquement mais statistiquement, s'appuyant sur des co-occurrences de n-grammes de caractères. Cette méthode différencie les tokens en début et en milieu de mot. Par conséquent, les préfixations et suffixations subissent des sorts différents (Hofmann *et al.*, 2020). Par exemple, le terme suffixé *collisionneur* est raisonnablement segmenté en `_collision neur` (ou _ indique le début de mot) et partagera



(a)

Décile	Terme	Occurrences
min	classification semi-dirigée	0
0,1	moment d'exécution	0
0,2	stellarateur	2
0,3	horloge à fontaine atomique	7
0,4	sondage au limbe	22
0,5	sauvetage côtier sportif	74
0,6	planche nautique	273
0,7	effet de rebond	1 052
0,8	embardée	4 327
0,9	clonage	45 680
max	pas	232 506 256

(b)

FIGURE 4 – (a) Exactitude de la prédiction de BLOOM-7,1G (DEF) sur le jeu de validation de FranceTerme selon le nombre d'occurrences des termes dans OSCAR-fr et ROOTS-fr. L'abscisse est logarithmique. (b) On montre un exemple aléatoire de terme pour chaque décile.

donc la même représentation que sa racine `_collision`. En revanche, le terme préfixé *précollision* sera segmenté `_préc oll ision`. Outre le fait que ni *oll* ni *préc* ne sont des morphèmes du français, nous observons que le modèle ne partage pas de représentation entre ces tokens et la racine `_collision` (il en irait de même si le terme avait été segmenté en `_pré collision`).

Nous quantifions ce phénomène à la figure 3d où l'on voit que les mots préfixés ont une plus grande fertilité que les mots suffixés. Pour les termes polylexicaux, nous définissons la fertilité d'un terme comme le nombre de segments maximum pour chacun de ses mots. D'autre part, nous montrons dans la figure 3c que BLOOM a plus de difficulté à prédire correctement les termes les plus fertiles. Ainsi, BLOOM prédit *consultation à distance* plutôt que le compact *téléconsultation* (référence) qui est segmenté `_tél éc ons ult ation` (contrairement à *consultation* qui a un token dédié). Ces résultats encouragent de futurs travaux sur une segmentation morphologique des termes complexes.

Fréquence et changement sémantique Faute d'une méthode objective pour classer un terme français comme néologisme ou lexicalisé, nous étudions à quelle fréquence les termes apparaissent dans deux grands corpus. Le premier, ROOTS-fr-open (Laurençon *et al.*, 2022), est un sous-ensemble du corpus d'entrainement du modèle BLOOM (BigScience & et al., 2023), restreint aux documents français disponibles sous licence *Creative Commons* [13] : il comprend environ 4 milliards de mots (20 Go), principalement extraits de contenus Wikimedia. Le second, OSCAR-fr 22.01 (Abadji *et al.*, 2022), est un extrait "nettoyé" du *Common Crawl*, dont une partie a également servi à l'entraînement de BLOOM. Il comprend 42 milliards de mots (382 Go).

Les résultats sont présentés à la figure 4a. Nous trouvons que 15,8% des termes de FranceTerme n'apparaissent aucune fois, même dans cet immense corpus. Nous estimons que les exemples aléatoires de la figure 4b, pour chaque décile, montrent bien une progression du sentiment néologique. À partir du septième décile, soit environ 1 000 occurrences, l'effet néologique est moins fort. C'est en effet à partir de ce seuil où BLOOM-7,1G (DEF) prédit bien mieux les termes.

D'autre part, nous remarquons que les termes très fréquents sont bien des néologismes mais sont

13. https://huggingface.co/bigscience-data

employés dans un sens différent dans FranceTerme : il s'agit donc de néologismes sémantiques (cf. section 2). Par exemple, *pas*, le terme le plus fréquent, provient du domaine électronique et est défini comme la « distance séparant deux lignes d'interconnexion voisines dans un circuit intégré ou sur un circuit imprimé nu », et non pas dans le sens du pas de la marche ou de l'adverbe de négation, dans lequel il apparaît vraisemblablement le plus souvent. Parmi les termes les plus fréquents, on peut citer d'autres exemples de néologismes sémantiques dans différents domaines : cœur (nucléaire), entrée (spatiologie), bois (sports). Encore une fois, nos observations se reflètent dans les performances de BLOOM-7,1G (DEF) qui prédit beaucoup moins précisément les termes à partir de 10^5 occurrences et ne fait aucune prédiction correcte après 10^7. Par exemple, pour *pression* « marquage serré de l'adversaire en possession du ballon », le modèle génère *marquage individuel*, ou bien pour *pont* « dispositif destiné à assurer entre deux réseaux locaux l'échange des trames de données sans les modifier, tout en détectant et en corrigeant les erreurs », le modèle génère *réseau local sans fil*.

Ces métaphores ne sont pas bien traitées par le modèle et il ne serait pas trivial de lui apprendre. Une première étape serait de mieux les identifier grâce à un corpus diachronique (Ryskina *et al.*, 2020) ou en étudiant le contexte d'utilisation des termes.

5 Conclusions et perspectives

Nous avons présenté une nouvelle approche pour traduire des néologismes scientifiques en exploitant leurs définitions. Nos expériences sur les thésaurus FranceTerme et TERMIUM montrent que les grands modèles de langues BLOOM et CroissantLLM sont capables d'utiliser cette information pour traduire le terme plus fidèlement, en particulier BLOOM qui dispose d'une plus grande expressivité. Nous avons également montré que ces modèles prédisent souvent une forme de néologisme adéquate mais qu'ils sont pénalisés lorsque la forme diffère entre l'anglais et le français.

Nous avons également mis en évidence plusieurs limites de ces modèles, qui traduisent mieux les termes lorsque la source et la cible sont surfaciquement proches (des emprunts ou des cognats) ou lorsque la cible apparaît assez fréquemment en corpus. Dans ce dernier cas, nous estimons qu'il ne s'agit pas d'un néologisme mais d'un terme lexicalisé. Ce phénomène est un travers de nos données d'évaluation : dès lors qu'un terme est présent dans un lexique ou thésaurus, il est institutionnalisé [14], contrairement aux néologismes que l'on peut rencontrer dans une nouvelle publication scientifique. Cette analyse devrait être approfondie car nous avons également remarqué que les termes les plus fréquents étaient créés par glissement sémantique, et sont très difficiles à traduire depuis leur définition donc, puisque leur sens n'est pas reflété par leurs composants (Temmerman, 2010).

Par ailleurs, nos futurs efforts porteront sur une modélisation morphologique des termes et de leurs définitions. En effet, nous avons mis en évidence les limites de la tokenisation BPE, en particulier pour les termes préfixés. Une segmentation morphologique pourrait être effectuée par un modèle dédié (Smit *et al.*, 2014; Batsuren *et al.*, 2022) ou apprise implicitement, directement à partir des caractères (Cherry *et al.*, 2018; Wang *et al.*, 2024). D'un point de vue applicatif, notre méthode pourrait être intégrée dans un système de traduction intégrant des lexiques (voir (Yvon & Abdul Rauf, 2020), pour un état de l'art ou encore (Semenov *et al.*, 2023)) ou pourrait servir à suggérer de nouvelles traductions aux lexicographes et traducteurs (pour enrichir FranceTerme, par exemple).

14. On peut distinguer l'institutionnalisation d'un terme de sa *lexicalisation*, qui supposerait un changement phonologique, syntaxique ou sémantique (Hohenhaus, 2005).

Remerciements

Nous remercions les membres du comité de programme pour leurs précieux commentaires. Nos remerciements s'adressent également à Natalie Kübler, Mathilde Huguin et Alexandra Mestivier pour leurs retours sur une première version de l'article, avec nos excuses pour les entorses aux théories linguistiques, dans l'intérêt de concilier TAL, morphologie et terminologie. Nous remercions enfin Ziqian Peng pour les expériences avec mBART et Felix Herron pour ses premiers travaux sur le sujet.

Ce projet a reçu un soutien de l'Agence Nationale de la Recherche (convention ANR-22-CE23-0033).

Références

ABADJI J., ORTIZ SUAREZ P., ROMARY L. & SAGOT B. (2022). Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 4344–4355, Marseille, France : European Language Resources Association.

AMIOT D. & DAL G. (2008). La composition néoclassique en français et l'ordre des constituants. *La composition dans une perspective typologique. Arras : Artois Presses Université*, p. 89–113.

ARNAUD P. J. (2003). *Les composés timbre-poste*. Presses Universitaires Lyon.

ARONOFF M. (1976). Word formation in generative grammar. *Linguistic Inquiry Monographs Cambridge, Mass*, (1), 1–134.

AUGUST T., REINECKE K. & SMITH N. A. (2022). Generating Scientific Definitions with Controllable Complexity. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 8298–8317, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.569.

BATSUREN K., BELLA G., ARORA A., MARTINOVIC V., GORMAN K., ŽABOKRTSKÝ Z., GANBOLD A., DOHNALOVÁ S., SEVČÍKOVÁ M., PELEGRINOVÁ K., GIUNCHIGLIA F., COTTERELL R. & VYLOMOVA E. (2022). The SIGMORPHON 2022 Shared Task on Morpheme Segmentation. In G. NICOLAI & E. CHODROFF, Éds., *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, p. 103–116, Seattle, Washington : Association for Computational Linguistics. DOI : 10.18653/v1/2022.sigmorphon-1.11.

BATSUREN K., BELLA G. & GIUNCHIGLIA F. (2021). Morphynet : a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th sigmorphon workshop on computational research in phonetics, phonology, and morphology*, p. 39–48.

BAWDEN R., BOURFOUNE H., CABOT B., CASSEREAU N., CORNETTE P., NAGUIB M., NÉVÉOL A. & YVON F. (2024). Les modèles Bloom pour le traitement automatique de la langue française. working paper or preprint, HAL : hal-04435371.

BAWDEN R. & YVON F. (2023). Investigating the Translation Performance of a Large Multilingual Language Model : the Case of BLOOM. DOI : 10.48550/ARXIV.2303.01911.

BIBER D., GRIEVE J. & IBERRI-SHEA G. (2010). Noun phrase modification.

BIGSCIENCE & ET AL. (2023). BLOOM : A 176B-Parameter Open-Access Multilingual Language Model. arXiv :2211.05100 [cs], DOI : 10.48550/arXiv.2211.05100.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G.,

HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I. & AMODEI D. (2020). Language Models are Few-Shot Learners. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. F. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 1877–1901 : Curran Associates, Inc.

CABRÉ M. T. (1999). *Terminology : Theory, methods, and applications*, volume 1. John Benjamins Publishing.

CARTIER E., SABLAYROLLES J.-F., BOUTMGHARINE N., HUMBLEY J., BERTOCCI M., JACQUET-PFAU C., KÜBLER N. & TALLARICO G. (2018). Détection automatique, description linguistique et suivi des néologismes en corpus : point d'étape sur les tendances du français contemporain. In *6e Congrès Mondial de Linguistique Française-Université de Mons, Belgique, 9-13 juillet 2018*, volume 46, p. 1–20. EDP Sciences.

CHERRY C., FOSTER G., BAPNA A., FIRAT O. & MACHEREY W. (2018). Revisiting Character-Based Neural Machine Translation with Capacity and Compression. In E. RILOFF, D. CHIANG, J. HOCKENMAIER & J. TSUJII, Éds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 4295–4305, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1461.

CLAVEAU V. & ZWEIGENBAUM P. (2005). Translating Biomedical Terms by Inferring Transducers. In S. MIKSCH, J. HUNTER & E. T. KERAVNOU, Éds., *Artificial Intelligence in Medicine*, Lecture Notes in Computer Science, p. 236–240, Berlin, Heidelberg : Springer. DOI : 10.1007/11527770_34.

CORBIN D. (2012). *Morphologie dérivationnelle et structuration du lexique*, volume 193. Walter de Gruyter.

DAILLE B. (2017). *Term Variation in Specialised Corpora : Characterisation, automatic discovery and applications*, volume 19 de *Terminology and Lexicography Research and Practice*. Amsterdam : John Benjamins Publishing Company. DOI : 10.1075/tlrp.19.

DAL G. (2003a). Analogie et lexique construit : quelles preuves ? Publisher : Toulouse : Université de Toulouse-le-Mirail, 1979-2006.

DAL G. (2003b). Productivité morphologique : définitions et notions connexes. *Langue française*, p. 3–23.

DELPECH E., DAILLE B., MORIN E. & LEMAIRE C. (2012). Extraction of Domain-Specific Bilingual Lexicon from Comparable Corpora : Compositional Translation and Ranking. In M. KAY & C. BOITET, Éds., *Proceedings of COLING 2012*, p. 745–762, Mumbai, India : The COLING 2012 Organizing Committee.

FAYSSE M., FERNANDES P., GUERREIRO N., LOISON A., ALVES D., CORRO C., BOIZARD N., ALVES J., REI R., MARTINS P., CASADEMUNT A. B., YVON F., MARTINS A., VIAUD G., HUDELOT C. & COLOMBO P. (2024). CroissantLLM : A Truly Bilingual French-English Language Model. arXiv :2402.00786 [cs], DOI : 10.48550/arXiv.2402.00786.

FRADIN B. (2015). *Nouvelles approches en morphologie*. PUF.

GAGE P. (1994). A New Algorithm for Data Compression. *Computer Users Journal*, **12**(2), 23–38. Place : USA Publisher : R & D Publications, Inc.

GELL-MANN M. (1964). A schematic model of baryons and mesons. *Physics Letters*, **8**(3), 214–215. DOI : https ://doi.org/10.1016/S0031-9163(64)92001-3.

GORDIN M. D. (2015). *Scientific Babel : How Science Was Done Before and After Global English*. University of Chicago Press. Google-Books-ID : UrnnBgAAQBAJ.

HEAD A., LO K., KANG D., FOK R., SKJONSBERG S., WELD D. S. & HEARST M. A. (2021). Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and

Symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, p. 1–18, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3411764.3445648.

HOFFMANN J., BORGEAUD S., MENSCH A., BUCHATSKAYA E., CAI T., RUTHERFORD E., CASAS D. D. L., HENDRICKS L. A., WELBL J., CLARK A., HENNIGAN T., NOLAND E., MILLICAN K., DRIESSCHE G. V. D., DAMOC B., GUY A., OSINDERO S., SIMONYAN K., ELSEN E., RAE J. W., VINYALS O. & SIFRE L. (2022). Training Compute-Optimal Large Language Models. arXiv :2203.15556 [cs], DOI : 10.48550/arXiv.2203.15556.

HOFMANN V., PIERREHUMBERT J. & SCHÜTZE H. (2020). DagoBERT : Generating Derivational Morphology with a Pretrained Language Model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 3848–3861, Online : Association for Computational Linguistics.

HOHENHAUS P. (2005). Lexicalization and institutionalization. In *Handbook of word-formation*, p. 353–373. Springer.

HONNIBAL M., MONTANI I., VAN LANDEGHEM S. & BOYD A. (2020). spaCy : Industrial-strength Natural Language Processing in Python. DOI : 10.5281/zenodo.1212303.

HUANG J., SHAO H., CHANG K. C.-C., XIONG J. & HWU W.-M. (2022). Understanding Jargon : Combining Extraction and Generation for Definition Modeling. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 3994–4004, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.266.

ISABELLE P., CHERRY C. & FOSTER G. (2017). A Challenge Set Approach to Evaluating Machine Translation. In M. PALMER, R. HWA & S. RIEDEL, Éds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 2486–2496, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1263.

JIN Y., KAN M.-Y., NG J. P. & HE X. (2013). Mining scientific terms and their definitions : A study of the ACL anthology. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 780–790.

JOULIN A., GRAVE E., BOJANOWSKI P. & MIKOLOV T. (2017). Bag of Tricks for Efficient Text Classification. In M. LAPATA, P. BLUNSOM & A. KOLLER, Éds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 427–431, Valencia, Spain : Association for Computational Linguistics.

LARIVIÈRE V. & RIDDLES A. (2021). Langues de diffusion des connaissances : quelle place reste-t-il pour le français. *Magazine de l'Acfas*.

LAROCHE A. & LANGLAIS P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)*, p. 617–625.

LAURENÇON H., SAULNIER L., WANG T., AKIKI C., VILLANOVA DEL MORAL A., LE SCAO T., VON WERRA L., MOU C., GONZÁLEZ PONFERRADA E. & NGUYEN H. (2022). The bigscience roots corpus : A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, **35**, 31809–31826.

LIEBER R. (2010). *Introducing morphology*. Cambridge : Cambridge University Press. OCLC : 650278652.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). RoBERTa : A Robustly Optimized BERT Pretraining Approach.

LIU Z., WANG S., GU Y., ZHANG R., ZHANG M. & WANG S. (2021). Graphine : A Dataset for Graph-aware Terminology Definition Generation. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 3453–3463, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.278.

LOMBARD A. & HUYGHE R. (2020). Catégorisation comme néologisme et sentiment des locuteurs. *Langue française*, **207**(3), 123–138. Place : Paris Publisher : Armand Colin, DOI : 10.3917/lf.207.0123.

LOMBARD A., HUYGHE R., BARQUE L. & GRAS D. (2023). Regular polysemy and novel word-sense identification. *The Mental Lexicon*, **18**(1), 94–119. DOI : 10.1075/ml.21002.lom.

MATTIELLO E. (2017). *Analogy in word-formation : A study of English neologisms and occasionalisms*, volume 309. Walter de Gruyter GmbH & Co KG.

MUENNIGHOFF N., WANG T., SUTAWIKA L., ROBERTS A., BIDERMAN S., SCAO T. L., BARI M. S., SHEN S., YONG Z.-X., SCHOELKOPF H., TANG X., RADEV D., AJI A. F., ALMUBARAK K., ALBANIE S., ALYAFEAI Z., WEBSON A., RAFF E. & RAFFEL C. (2023). Crosslingual Generalization through Multitask Finetuning. arXiv :2211.01786 [cs], DOI : 10.48550/arXiv.2211.01786.

NAMER F. (2003). Automatiser l'analyse morpho-sémantique non affixale : le système DériF. *Cahiers de grammaire*, **28**, 31–48.

PENG Z., BAWDEN R. & YVON F. (2024). À propos des difficultés de traduire automatiquement de longs documents. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2024*, Toulouse, France.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, **21**, 1–67. https ://github.com/google-research/text-to-text-transfer-transformer.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). Squad : 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 2383–2392.

RIGOUTS TERRYN A., HOSTE V. & LEFEVER E. (2020). In no uncertain terms : a dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Language Resources and Evaluation*, **54**(2), 385–418. DOI : 10.1007/s10579-019-09453-9.

ROUSSIS D., PAPAVASSILIOU V., PROKOPIDIS P., PIPERIDIS S. & KATSOUROS V. (2022). SciPar : A collection of parallel corpora from scientific abstracts. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 2652–2657, Marseille, France : European Language Resources Association.

RYSKINA M., RABINOVICH E., BERG-KIRKPATRICK T., MORTENSEN D. R. & TSVETKOV Y. (2020). Where New Words Are Born : Distributional Semantic Analysis of Neologisms and Their Semantic Neighborhoods. In *Proceedings of the Society for Computation in Linguistics 2020*, p. 367–376.

SEMENOV K., ZOUHAR V., KOCMI T., ZHANG D., ZHOU W. & JIANG Y. E. (2023). Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, p. 663–671, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.wmt-1.54.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Compu-*

tational Linguistics (Volume 1 : Long Papers), p. 1715–1725, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1162.

SMIT P., VIRPIOJA S., GRÖNROOS S.-A. & KURIMO M. (2014). Morfessor 2.0 : Toolkit for statistical morphological segmentation. In S. WINTNER, M. TADIĆ & B. BABYCH, Éds., *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, p. 21–24, Gothenburg, Sweden : Association for Computational Linguistics. DOI : 10.3115/v1/E14-2006.

TANG Y., TRAN C., LI X., CHEN P.-J., GOYAL N., CHAUDHARY V., GU J. & FAN A. (2021). Multilingual translation from denoising pre-training. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 3450–3466, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.304.

TEMMERMAN R. (2010). Why special language translators need insight into the mechanisms of metaphorical models and figurative denominations. *Meaning in Translation*, **19**, 347–365.

TRIBOUT D. (2010). *Les conversions de nom à verbe et de verbe à nom en français*. Thèse de doctorat, Université Paris Diderot (Paris 7).

WANG J., GANGAVARAPU T., YAN J. N. & RUSH A. M. (2024). MambaByte : Token-free Selective State Space Model. arXiv :2401.13660 [cs], DOI : 10.48550/arXiv.2401.13660.

YVON F. & ABDUL RAUF S. (2020). *Utilisation de ressources lexicales et terminologiques en traduction neuronale*. Research Report 2020-001, LIMSI-CNRS.

ZHANG Y., CHEN Q., ZHANG Y., WEI Z., GAO Y., PENG J., HUANG Z., SUN W. & HUANG X.-J. (2020). Automatic term name generation for gene ontology : task and dataset. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 4705–4710.

A Classification morphosyntaxique

Nous construisons un classifieur multi-étiquettes pour quatre des cinq classes définies à la section 2 : préfixation, suffixation, composition néoclassique ou ordinaire. Pour la cinquième (composition syntagmatique), nous nous reposons sur la simple heuristique du nombre de mots segmentés par spaCy. S'il y a plusieurs mots, nous considérons que le terme est un syntagme.

Pour détecter ces quatre procédés morphologiques, nous utilisons l'architecture de FastText (Joulin *et al.*, 2017), qui fournit un classifieur linéaire pour des séquences de caractères, représentées par l'ensemble des mots et des n-grammes de caractères qui y sont trouvées. Dans notre utilisation, ce classifieur est entraîné en mode « un contre tous » (*one versus all*), équivalent à un classifieur binaire pour chacune des classes identifiées supra.

Dans cette section, nous décrivons plus précisément les données utilisées pour entraîner et évaluer ce classifieur.

A.1 MorphyNet et SIGMORPHON

Nous construisons un jeu d'entraînement et d'évaluation à partir des bases étymologiques MorphyNet (Batsuren *et al.*, 2021) et de celle utilisée pour la *shared task* SIGMORPHON 2022 (Batsuren *et al.*, 2022), toutes deux extraites depuis le Wiktionnaire anglais [15]. Nous combinons les deux bases car

15. https://en.wiktionary.org/

Procédé	Occurrences EN	Occurrences FR
Ordinaire	45 463	2 854
Néoclassique	32 766	7 583
Préfixation	190 305	96 721
Suffixation	217 404	155 169

TABLE 3 – Nombre de mots dans nos corpus de classification morphologique anglais et français pour chaque procédé indépendamment

elles contiennent des informations complémentaires : SIGMORPHON contient des compositions ordinaires mais fournit seulement la segmentation morphologique, tandis que MorphyNet permet de retrouver la racine de tous les mots, même complexes, et différencie préfixation et suffixation.

Ces deux bases partagent toutefois le même défaut : elles ne considèrent pas les compositions néoclassiques, qui se trouvent mêlées aux affixations. Pour les différencier, nous usons d'une simple heuristique : si tous les morphèmes d'un mot sont classés comme affixes par MorphyNet, alors aucun n'est libre, il s'agit donc d'un composé néoclassique.

Notre algorithme est récursif pour décomposer les termes complexes (avec plus de deux morphèmes). Par exemple, *prétraitement* sera décomposé en *pré+traitement* (préfixation) et *traitement* sera à son tour décomposé en *traiter+ment* (suffixation). *Prétraitement* héritera donc de ces deux étiquettes.

A.2 Implémentation

Les statistiques des lexiques anglais et français sont dans le tableau 3, qui confirment que les composés ordinaires sont bien plus rares en français. Nous remarquons également que les composés néoclassiques sont moins systématiquement annotés en français qu'en anglais, peut-être parce que MorphyNet et SIGMORPHON proviennent du Wiktionnaire anglais. Nous montrons également comment les différents procédés se combinent dans le tableau 5. Il est fréquent que des termes dérivés soient à la fois préfixés et suffixés, ce qui est en revanche impossible, par construction, pour les composés néoclassiques.

Ces lexiques sont divisés aléatoirement en ensemble d'entraînement (80%), de validation (10%) et de test (10%). Nous entraînons un modèle pour chaque langue. Les monomorphèmes (fléchis ou non) sont conservés et servent d'exemple négatifs pour toutes les classes pendant l'entraînement.

Les hyperparamètres de FastText sont déterminés automatiquement sur le jeu de validation grâce à la bibliothèque python fastText. Pour les deux langues, nous trouvons notamment qu'il est optimal d'utiliser des n-grammes de caractères pour $n \in [\![3, 6]\!]$.

A.3 Résultats

Les résultats sur le jeu de test sont dans le tableau 4. Le classifieur est très précis et a un très bon rappel, à l'exception des composés ordinaires en français qui sont sous-représentés, de par leur rareté, et dont le rappel est modeste. Dans une moindre mesure, le rappel pour les composés néoclassiques est moins élevé en français qu'en anglais à cause de leur sous-représentation dans SIGMORPHON,

	Anglais			Français		
	Précision	Rappel	F1	Précision	Rappel	F1
Ordinaire	95.3	93.0	94.1	89.7	66.3	76.2
Néoclassique	93.4	91.4	92.4	92.2	87.2	89.6
Préfixation	91.5	91.3	91.4	93.8	93.5	93.6
Suffixation	93.2	93.3	93.2	97.4	98.0	97.7
Total	92.7	92.4	92.5	95.9	95.7	95.8

TABLE 4 – Résultats de la classification morphologique multi-étiquettes, en anglais et en français

Ordinaire	**Néo.**	**Pré.**	**Suff.**	**Occ. EN**	**Occ. FR**
				207 074	118 811
			X	109 353	90 646
		X		91 115	35 646
		X	X	88 349	60 307
	X			17 191	3 508
	X		X	9 677	3 640
	X	X		5 593	432
	X	X	X	0	0
X				34 425	2 162
X			X	5 552	353
X		X		808	115
X		X	X	4 373	221
X	X			138	1
X	X		X	100	2
X	X	X		67	0
X	X	X	X	0	0

TABLE 5 – Nombre de mots dans nos corpus de classification morphologique anglais et français pour chaque combinaison de procédé

comme évoqué plus haut.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

WikiFactDiff: Un Grand jeu de données Réaliste et Temporellement Adaptable pour la Mise à Jour Atomique des Connaissances Factuelles dans les Modèles de Langue Causaux

Hichem Ammar Khodja[1, 2] Frederic Bechet[2,3] Quentin Brabant[1]
Alexis Nasr[2] Gwénolé Lecorvé[1]
(1) Orange Innovation - *Lannion, France*
(2) Aix-Marseille Univ, CNRS, LIS, UMR 7020 - *Marseille, France*
(3) International Laboratory on Learning Systems (ILLS - IRL2020 CNRS)
hichem.ammarkhodja@orange.com, frederic.bechet@lis-lab.fr,
quentin.brabant@orange.com, alexis.nasr@lis-lab.fr,
gwenole.lecorve@orange.com

RÉSUMÉ

La factualité des modèles de langue se dégrade avec le temps puisque les événements postérieurs à leur entraînement leur sont inconnus. Une façon de maintenir ces modèles à jour pourrait être la mise à jour factuelle à l'échelle de faits atomiques. Pour étudier cette tâche, nous présentons WikiFactDiff, un jeu de données qui représente les changements survenus entre deux dates sous la forme d'un ensemble de faits simples, sous format RDF, divisés en trois catégories : les faits à apprendre, les faits à conserver et les faits obsolètes. Ces faits sont verbalisés afin de permettre l'exécution des algorithmes de mise à jour et leur évaluation, qui est présentée dans ce document. Contrairement aux jeux de données existants, WikiFactDiff représente un cadre de mise à jour réaliste qui implique divers scénarios, notamment les remplacements de faits, leur archivage et l'insertion de nouvelles entités.

ABSTRACT

WikiFactDiff : A Large, Realistic, and Temporally Adaptable Dataset for Atomic Factual Knowledge Update in Causal Language Models

The factuality of language models deteriorates over time since events subsequent to their training are unknown to them. One way to keep these models up to date could be factual update of atomic facts. To study this task, we present WikiFactDiff, a dataset that represents changes between two dates as a set of simple facts, in RDF format, divided into three categories : facts to learn, facts to keep and obsolete facts. Indeed, WikiFactDiff was built by comparing the state of Wikidata on January 4, 2021 and February 27, 2023. These facts are verbalized in order to allow the execution of update algorithms and their evaluation, which is presented in this document . Unlike existing datasets, WikiFactDiff represents a realistic editing framework that involves various scenarios, including replacements, archiving, and inserting new entities.

MOTS-CLÉS : Mise à jour des connaissances, Modèles de langue, Jeu de données.

KEYWORDS: Knowledge update, Language models, Dataset.

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 262–281.

1 Introduction

Les grands modèles de langue (GML) n'apprennent que les faits datant d'avant la date de collecte de leurs données d'entraînement (Lazaridou *et al.*, 2021). Avec le temps, ces modèles peuvent ainsi propager des informations obsolètes, ce qui peut avoir des implications concrètes dans des domaines tel que la santé ou la politique. Par conséquent, savoir mettre à jour les faits connus par ces modèles est crucial pour garantir leur utilité et leur pertinence, ainsi que la fiabilité globale de toutes les applications d'IA qui en découlent.

Si la notion de connaissance est large (incluant les connaissances sur les faits, la linguistique, les procédures, etc.), la mise à jour des connaissances factuelles constitue actuellement un domaine de recherche particulièrement actif. En effet, contrairement aux approches traditionnelles de mise à jour globale *via* un affinage, une approche récente propose de réaliser des mises à jour atomiques, c'est-à-dire en considérant des faits uniques pour la mise à jour. Ces faits sont représentés dans la littérature comme des triplets (sujet, relation, objet), tel que (Inde, chef de l'État, Ram Nath Kovind). Dans ce cadre, de nombreux scénarios de mise à jour peuvent se produire (p. ex., l'archivage de faits obsolètes, l'insertion de nouvelles entités). Cependant, à notre connaissance, les algorithmes et les jeux de données de mise à jour actuels se limitent au seul scénario de remplacement (p. ex., mettre à jour le président des USA) (Yao *et al.*, 2023). De plus, les mises à jour dans ces travaux sont irréalistes (p. ex., changer le domaine d'expertise d'Albert Einstein de la physique à la biologie), ce qui introduit des défis dans le maintien de la cohérence globale des connaissances. Cette situation ne reflète pas l'utilisation souhaitée pour des applications réelles.

Sujet	**Relation**	**Objet**	**Étiquette**
Japan	Population	125,96M	obsolète
		125,44M	nouveau
Cristiano Ronaldo	member of sports team	Portugal national association football team	inchangé
		Juventus F.C.	obsolète
		Al-Nassr	nouveau
USA	head of government	Donald Trump	obsolète
		Joe Biden	nouveau
Vyacheslav Geraschenko	coach of sports team	FC Smorgon	obsolète
		FC Dnepr Mogilev	nouveau
ChatGPT	instance of	language model	nouveau
		...	nouveau
	inception	30 November 2022	nouveau
	...	...	nouveau

TABLE 1 – Exemples de mises à jour tirées de WikiFactDiff.

Pour remédier à ces limites, nous introduisons WikiFactDiff, un vaste jeu de données pour la mise à jour des connaissances factuelles des GML avec un large éventail de scénarios pour un large éventail d'entités de popularité variable. Il se présente sous la forme d'un ensemble de 223K mises à jour reflétant l'évolution des connaissances entre deux instances de Wikidata à deux dates, T_{anc} et T_{nouv}. Comme illustré dans la table 1, chaque triplet est étiqueté par l'une des classes « nouveau », « obsolète » ou « inchangé » [1]. Ces triplets sont également verbalisés afin de permettre l'application des algorithmes de mise à jour actuels et la mesure des métriques d'évaluation du domaine.

1. Chaque triplet dont le sujet est "ChatGPT" est classé comme *nouveau* car cette entité est nouvelle par rapport à T_{anc}.

En pratique, WikiFactDiff couvre l'évolution des connaissances factuelles entre T_{anc} = *4 janvier 2021* et T_{nouv} = *27 février 2023*. Le choix de T_{anc} est tel que les nouveaux faits du jeu de données sont postérieurs à ceux du corpus Pile (Gao *et al.*, 2021), largement utilisé pour entraîner des GMLs (Wang & Komatsuzaki, 2021; Black *et al.*, 2022; Biderman *et al.*, 2023; Black *et al.*, 2021). Une autre force de WikiFactDiff est que le processus de création s'adapte à la période $[T_{anc}, T_{nouv}]$ de notre choix. Par conséquent, de nouvelles versions de WikiFactDiff peuvent être publiées pour s'aligner sur les dates de collecte de jeux de données autres que Pile. Pour illustrer l'utilisabilité du corpus, une évaluation des algorithmes de mise à jour existants est présentée. Cela fournit une base de référence à la communauté.

L'article est organisé comme suit : la section 2 présente le domaine et les travaux associés ; la section 3 donne un aperçu global de WikiFactDiff ; la section 4 décrie son processus de création ; et la section 5 présente les performances des algorithmes de mise à jour sur WikiFactDiff.

2 État de l'art

La famille d'algorithmes présentée ici permet la mise à jour du modèle en utilisant une phrase en langage naturel, dite d'**injection**, exprimant un fait (p. ex., *"The president of USA is Joe Biden"*). De même, l'évaluation de ces algorithmes repose également sur des faits verbalisés, où il est demandé au modèle de compléter des phrases à trou avec les informations correctes. En général, la qualité de la mise à jour est mesurée sur deux aspects : la généralisation du modèle sur des phrases sémantiquement équivalentes à la phrase d'injection (**généralisation**) et le maintien des performances sur des faits indépendants de celui mis à jour (**spécificité**).

Plusieurs algorithmes ont été proposés pour la mise à jour atomique des faits dans les GMLs, telles que le prompting (Si *et al.*, 2023), l'affinage partiel des paramètres, avec ou sans régularisation, noté **FT+L** et **FT** respectivement (Zhu *et al.*, 2020). D'un autre côté, De Cao *et al.* (2021) et Sinitsin *et al.* (2020) ont proposé des approches dites "hyper-réseaux", dont la plus avancée, **MEND** (Mitchell *et al.*, 2022) propose une solution rapide et qui passe à l'échelle.

Des progrès ont également été réalisés dans la localisation des connaissances dans les modèles de langue. En particulier, Devlin *et al.* (2019) ont sondé les connaissances de BERT et ont montré qu'un ensemble restreint de neurones joue un rôle crucial dans la prédiction correcte de faits dans un GML. Plus tard, Meng *et al.* (2022) ont procédé à une analyse causale (Vig *et al.*, 2020) sur GPT-2 (Radford *et al.*, 2018) qui a conduit à des résultats similaires et ont introduit un algorithme nommé **ROME** qui s'est démarqué comme le plus efficace. Enfin, inspiré de ROME, Meng *et al.* (2023) ont introduit **MEMIT**, un algorithme de mise à jour capable d'effectuer des milliers de mises à jour simultanées.

En termes de bancs d'essais, CounterFact (Meng *et al.*, 2022) et zsRE (Levy *et al.*, 2017) sont les jeux de données de référence pour évaluer les algorithmes de mise à jour. Cependant, leurs mises à jour se limitent au scénario de remplacement et ne contiennent pas de litéraux (p. ex., la mise à jour de la population d'un pays). De plus, elles sont irréalistes car les nouvelles valeurs d'un fait sont générées de manière aléatoire, donnant lieu à des mises à jour telles que le remplacement de la spécialité d'Albert Einstein, qui est la physique, par la biologie, ce qui est totalement irréaliste. Il convient de noter que les auteurs de zsRE ont sélectionné au hasard les faits utlisés pour évaluer la spécificité. Notamment, les recherches de Meng *et al.* (2022) ont révélé que l'évaluation de la spécificité sur ces faits sélectionnés au hasard n'est pas une mesure suffisamment sensible. En revanche, l'évaluation de

la spécificité sur des faits voisins s'est avérée plus sensible et met mieux en évidence les limites des algorithmes de mise à jour.

3 Présentation de WikiFactDiff

Dans ce travail, les faits sont représentés à l'aide de triplets (sujet, relation, objet) tels que (France, capitale, Paris) ou (Allemagne, partage une frontière, Suisse). Nous définissons alors le (s, r)**-groupe** comme la collection de triplets qui ont s comme sujet et r comme relation. Au besoin, nous utilisons parfois plus simplement le terme **groupe** pour désigner une collection de triplets partageant le même sujet et la même relation.

WikiFactDiff est collecté pour la période du 4 janvier 2021 au 27 février 2023 et contient 223K mises à jour. Chaque mise à jour concerne un (s, r)-groupe unique avec chacun de ses triplets étiquetés comme : **nouveau** lorsque le fait porté par ce triplet s'est produit après T_{anc}, c'est-à-dire qu'il n'était pas valide avant T_{anc} mais qu'il l'est à l'instant T_{nouv} ; **obsolète** lorsque le fait était valide jusqu'à T_{anc} mais ne l'est plus à l'instant T_{nouv} ; ou **inchangé** pour les faits qui sont restés valides entre T_{anc} et T_{nouv}.

	CounterFact	**WFD$_{\text{remp}}$**	**WikiFactDiff**
Triplets	43,838	20,988	349,441
Sujets	20,391	9,926	100,986
Relations	34	153	667
Objets	870	12,283	109,122
Objets "entité"	870	5,496	75,417
Objets "litéral"	0	6,787	33,705
Màj	21,919	10,494	223,140
RemplaceObjet	21,919	10,494	32,996
Archivage	0	0	2,756
AjoutObjet	0	0	1,494
AjoutRelation	0	0	50,541
AjoutEntité	0	0	132,844
Autre	0	0	2,509
Réaliste ?	✗	✓	✓
Adap. temp. ?	✗	✓	✓

TABLE 2 – Statistiques des jeux de données. "Adap. temp." signifie "Adaptabilité temporelle"

De ce cadre, nous décrivons plusieurs scénarios de mise à jour :

- **RemplaceObjet** : Le groupe mis à jour contient deux triplets : l'un *nouveau*, l'autre *obsolète*.
- **Archivage** : Tous les triplets du groupe sont *obsolètes*.
- **AjoutObjet** : Le groupe contient au moins un triplet *nouveau* et au moins un triplet *inchangé*, p. ex., ajouter un membre à une organisation existante.
- **AjoutRelation** : Le groupe ne contient que des triplets *nouveau* et s n'est pas une nouvelle entité, ce qui signifie ajouter une nouvelle propriété à une entité s existante. Un exemple serait d'ajouter une « date de décès ».
- **AjoutEntité** : Lorsque s est une nouvelle entité. Dans ce cas, tous les triplets du groupe sont nécessairement étiquetés *nouveau*.

— **Autre** : Le reste des mises à jour. Il s'agit d'autres situations, plus rares et variées.

Puisque les algorithmes actuels ne gèrent que le scénario de remplacement de faits, une version réduite de WikiFactDiff est également proposée dans le but de comparer ces algorithmes, notamment avec le corpus CounterFact. Cette version réduite s'appelle WFD_{remp} [2]. Une comparaison mettant en évidence les différences entre WikiFactDiff et WFD_{remp} par rapport à CounterFact est présentée dans la table 2.

4 Processus de création du corpus

La Figure 1 illustre le processus de création de WikiFactDiff. Ce processus est fondé sur 2 instances brutes de Wikidata aux instants T_{anc} et T_{nouv}, notées $\mathbf{W}_{anc}$ et $\mathbf{W}_{nouv}$. Un aperçu de haut niveau des étapes de création est présentée dans ce qui suit. Une description détaillée se trouve dans l'annexe A.

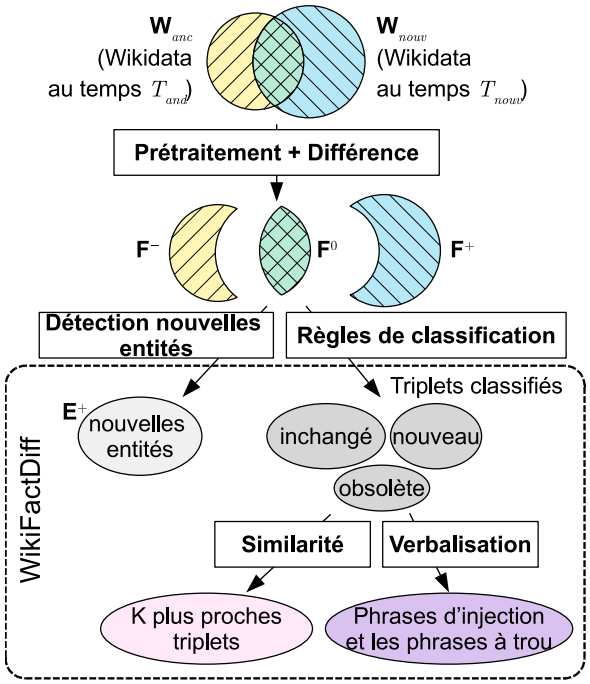


FIGURE 1 – Illustration étape par étape du processus de création du WikiFactDiff

Prétraitement. Les triplets dans les instances brutes de Wikidata sont nettoyés et filtrés. Cela inclut la suppression des informations non-pertinentes autour d'un triplet dans Wikidata (p. ex., des informations additionnelles, les références permettant de justifier le fait porté par ce triplet, etc.), la suppression des triplets dont l'information est incomplète ou peu fiable, le filtrage des triplets décrivant des méta-données de Wikidata. De plus, les triplets dont l'objet est non pertinent, tel que qu'un document PDF, une URL, une image ou une vidéo, sont aussi supprimés.

Afin d'assurer la pertinence des entités retenues, nous supprimons les triplets dont le sujet (une entité) ne possède pas un article Wikipédia dédié. Enfin, nous conservons seulement les faits à jour pour chaque version de Wikidata. En d'autres termes, chaque triplet dans la version prétraitée de $\mathbf{W}_{anc}$ (resp. $\mathbf{W}_{nouv}$), notée $\mathbf{W}_{anc}^{Pré}$ (resp. $\mathbf{W}_{nouv}^{Pré}$), est valide au temps T_{anc} (resp. T_{nouv}).

2. WFD_{remp} ne contient que les remplacements de WikiFactDiff avec un sous-échantillonage de la relation "population" d'un facteur 14 afin garder une diversité dans les relations évaluées.

Différence. L'intersection et la différence sur les ensembles de triplets provenant de $\mathbf{W}_{anc}^{Pré}$ et de $\mathbf{W}_{nouv}^{Pré}$ sont calculées pour produire les ensembles complémentaires : Les faits qui sont uniquement dans $\mathbf{W}_{anc}^{Pré}$, ceux qui sont uniquement dans $\mathbf{W}_{nouv}^{Pré}$, et ceux qui sont à la fois dans $\mathbf{W}_{anc}^{Pré}$ et $\mathbf{W}_{nouv}^{Pré}$.

Détection de nouvelles entités. Les nouvelles entités sont des objets tangibles ou intangibles qui n'existaient pas avant T_{anc}. Des exemples notables incluent *ChatGPT*, *L'invasion russe de l'Ukraine en 2022*, *Lilibet of Sussex*, entre autres (en supposant $T_{anc} = $ *4 janvier 2021*). Cet ensemble pourra constituer pour la communauté un support pour une tâche d'insertion d'entités dans les GMLs.

Les nouvelles entités sont toutes les entités e telles que : (i) e n'est présent que dans $\mathbf{W}_{nouv}^{Pré}$; (ii) il existe un triplet (e, r, d) où r est une relation désignant la date de création de e [3], et d est une date telle que $d > T_{anc}$. La condition (ii) est nécessaire car certains faits peuvent être antérieurs à T_{anc} mais le fait manquait dans $\mathbf{W}_{anc}$.

Règles de classification. Tous les triplets de $\mathbf{W}_{anc}^{Pré}$ et $\mathbf{W}_{nouv}^{Pré}$ sont filtrés à l'aide de règles élaborées manuellement pour les étiqueter avec « nouveau », « obsolète », ou « inchangé » (section 3). L'étape de filtrage permet également de supprimer les (s, r)-groupes où la nature des changements n'est pas tout à fait claire. Il s'agit de garantir que les changements factuels retenus reflètent des changements effectifs dans le monde réel (annexe A.3).

Recherche de faits voisins pour évaluer la spécificité. Lorsqu'un fait est mis à jour, la distribution de probabilité du modèle de langue est modifiée, ce qui peut dégrader sa précision sur d'autres faits. Ce phénomène est connu sous le nom de **débordement**. Pour permettre sa détection et sa mesure, WikiFactDiff est accompagné de faits voisins susceptibles d'être modifiés négativement lorsqu'un (s, r)-groupe donné est mis à jour.

Notre méthode des K plus proches triplets s'appuie sur une similarité entre entités. Cette similarité est calculée de la façon suivante : pour chaque entité s, nous identifions l'ensemble des triplets I_s dont elle est le sujet $(s, *, *)$. Ensuite, l'ensemble des relations, des objets (de type "entité" seulement), et des paires relation-objet mentionnés dans I_s sont organisés dans une liste L_s. Après cela, des représentations TF-IDF de L_s sont calculées pour chaque entité s. La similarité entre deux entités est le cosinus entre leurs vecteurs TF-IDF respectifs.

Afin de trouver la liste (notée P) des K triplets les plus proches d'un (s, r)-groupe donné, les entités les plus proches de s sont listées. Pour chaque entité, un unique triplet de la forme (s', r, o') est ajouté à P s'il existe. Ce processus est maintenu jusqu'à ce que P possède K triplets. Chaque (s, r)-groupe de WikiFactDiff est accompagné de ses 10 triplets les plus proches ($K = 10$). Les détails de notre méthode se trouvent dans l'annexe A.4.

Verbalisation. A l'aide de ChatGPT combiné à une procédure de post-traitement, des patrons sont générés pour chaque relation afin de verbaliser tout triplet de WikiFactDiff (détails et exemples en annexe A.5). Des phrases à trou sont générées à partir de ces patrons pour effectuer les mises à jour factuelles applicables par les algorithmes existants et évaluer leurs performances.

En conséquence, WikiFactDiff inclut tous les triplets filtrés et étiquetés, l'ensemble des entités nouvelles, les triplets les plus proches et les verbalisations pour chaque triplet. La chaîne de création sera publié sur GitHub.

3. En pratique, ces relations sont 'inception', 'date of birth', 'start time' , 'date of discovery or invention', 'date of official opening', 'announcement date', 'point in time' et 'publication date'.

Efficacité : différence	$\mathbb{P}^*[o^*\|\phi_m] - \mathbb{P}^*[o\|\phi_m]$
Efficacité : succès	$\mathbb{1}_{\mathbb{P}^*[o^*\|\phi_m]>\mathbb{P}^*[o\|\phi_m]}$
Généralisation : différence	$\frac{1}{\|\Phi\|}\sum_{\phi\in\Phi}\mathbb{P}^*[o^*\|\phi] - \mathbb{P}^*[o\|\phi]$
Généralisation : succès	$\frac{1}{\|\Phi\|}\sum_{\phi\in\Phi}\mathbb{1}_{\mathbb{P}^*[o^*\|\phi]>\mathbb{P}^*[o\|\phi]}$
Débordement	$-\frac{1}{\|N\|}\sum_{(\phi',o')\in N} min(\mathbb{P}^*[o'\|\phi'] - \mathbb{P}[o'\|\phi'], 0)$
Fluidité	$\frac{1}{\|\Phi\|}\sum_{\phi\in\Phi}\frac{2}{3}H_2(G(\phi)) + \frac{4}{3}H_3(G(\phi))$

TABLE 3 – Définition des métriques d'évaluation. $\mathbb{P}$ et $\mathbb{P}^*$ sont la fonction de probabilité du modèle de langue normalisée par la longueur de l'objet (en utilisant la moyenne géométrique), respectivement avant et après la mise à jour. $H_n(x)$ est l'entropie n-gramme pondérée sur le texte x. $G(\phi)$ est la fonction de génération de texte (gloutonne) du modèle étant donné l'amorce ϕ.

5 Expérimentations

Cette section évalue les algorithmes de mise à jour atomique existants sur le sous-ensemble WFD_{remp}, la version restreinte de WikiFactDiff pour le seul scénario de mise à jour de remplacement d'objet (voir la section 3). Les algorithmes sont ROME, MEMIT, MEND, FT et FT+L, tels qu'implémentés par Meng [4]. Le modèle à mettre à jour est GPT-J configuré en précision `bfloat16`. Ces mises à jour sont effectuées à l'aide d'une RTX3090 (24 Go de VRAM).

Une mise à jour m de WFD_{remp} consiste en le remplacement d'un fait (s, r, o) par un fait (s, r, o^*) ; par exemple, (*Japon*, *population*, *125.96M*) par (*Japon*, *population*, *125.44M*). La phrase d'injection est alors produite à partir d'un patron de phrase ϕ_m (*"The population of Japan is __"*) et instanciée sur o^* en comblant le trou (par exemple, *"The population of Japan is 125.44M"*). Nous désignons par $\phi_m + o^*$ la phrase d'injection ainsi construite.

Une fois la mise à jour effectuée, quatre aspects sont évalués : l'efficacité, la généralisation, la spécificité et la fluidité. L'efficacité est atteinte si le modèle préfère le nouvel objet o^* à l'ancien objet o étant donné l'amorce ϕ_m. La généralisation est obtenue si le modèle préfère o^* à o sur des phrases à trou alternatives à partir d'un ensemble Φ (dans notre expérience, Φ contient 4 phrases à trou). La spécificité est obtenue en minimisant le débordement, qui est une dégradation de la capacité du modèle à prédire les bons objets des phrases à trou correspondantes à des faits indépendants qui

4. github.com/kmeng01/memit

Algo.	Efficacité-D ↑	Efficacité-S ↑	Gén.-D ↑	Gén.-S ↑	Débordement ↓		Fluidité ↑	Temps↓
					Aléatoire	K-plus-proche		sec/màj
GPT-J	-1.4 ± 0.2	44.6 ± 1.0	-1.3 ± 0.2	44.4 ± 0.9	0.0 ± 0.0	0.0 ± 0.0	5.2 ± 0.0	0.0 ± 0.0
FT	45.9 ± 0.5	99.6 ± 0.1	45.7 ± 0.5	**99.5 ± 0.1**	**3.3 ± 0.1**	**5.6 ± 0.2**	**0.6 ± 0.0**	1.4 ± 0.0
FT+L	**12.9 ± 0.6**	**72.9 ± 0.9**	**1.1 ± 0.2**	**53.6 ± 0.8**	0.1 ± 0.0	**0.3 ± 0.0**	5.1 ± 0.0	2.4 ± 0.0
MEND	64.5 ± 0.6	99.4 ± 0.1	28.8 ± 0.5	96.5 ± 0.3	**0.0 ± 0.0**	1.0 ± 0.1	4.9 ± 0.0	1.1 ± 0.0
ROME	**95.5 ± 0.2**	**99.7 ± 0.1**	**59.5 ± 0.6**	98.0 ± 0.2	**0.0 ± 0.0**	0.6 ± 0.1	**5.2 ± 0.0**	4.9 ± 0.0
MEMIT‡	87.4 ± 0.3	99.5 ± 0.1	42.1 ± 0.6	94.4 ± 0.3	**0.0 ± 0.0**	**0.3 ± 0.0**	**5.2 ± 0.0**	**41.4 ± 0.2**
PROMPT	58.6 ± 0.5	98.9 ± 0.2	30.8 ± 0.5	93.3 ± 0.4	**1.1 ± 0.0**	**0.3 ± 0.0**	4.4 ± 0.0	**0.0 ± 0.0**

TABLE 4 – Résultats numériques des algorithmes de mise à jour sur WFD_{remp} avec leurs intervalles de confiance respectifs à 95%. **D** et **S** signifient respectivement *différence* et *succès*. Les valeurs **soulignées en vert** représentent les maxima par colonne et les valeurs en **rouge** indiquent un échec évident d'un algorithme sur une métrique. ‡ indique les algorithmes conçus pour les mises à jour par lots.

n'ont pas été mis à jour. Pour mesurer le débordement, nous nous appuyons sur un ensemble de triplets $\{(s_i, r, o_i)\}_i$ sélectionnés soit de manière aléatoire [5], ou en utilisant la méthode de recherche des faits voisins. Pour chaque triplet (s_i, r, o_i), par ex. (*Chine*, *population*, *1.412B*), nous choisissons au hasard un patron sur r et remplissons l'emplacement du sujet avec s_i pour créer une phrase à trou ϕ_i, par ex. *"The population of China is __"*. Enfin, la fluidité (Zhang *et al.*, 2018) est la capacité du modèle à produire des phrases fluides ; elle ne devrait pas diminuer après la mise à jour. La définition exacte de ces métriques est disponible dans le tableau 3. Les performances moyennes de chaque algorithme sur WFD_{remp} sont présentées dans le tableau 4.

En plus des méthodes de mise à jour mentionnées dans la section 2, nous évaluons l'algorithme PROMPT, qui consiste à influencer les connaissances du modèle au moment de l'inférence en préfixant chaque phrase à trou avec $\phi_m + o^*$. Par exemple, si nous voulons mettre à jour le président des États-Unis en *Joe Biden*, nous préfixons le modèle avec *"The president of USA is Joe Biden."*. C'est une opportunité de comparer deux catégories de méthodes : celles qui modifient les paramètres du modèle (comme évoqué précédemment) et celles qui injectent des connaissances *via* du prompting. Cette dernière approche, largement utilisée dans les méthodes de génération augmentée par récupération (GAR) (Lewis *et al.*, 2020), évite le défi de la mise à jour des connaissances en insufflant directement les connaissances requises dans le préfixe de génération. Il est pertinent de noter que PROMPT ajoute une surcharge de calcul qui croit quadratiquement avec la taille du préfixe utilisé, qui de plus, est limité par taille du contexte du modèle de langue.

5.1 Résultats généraux

Nos résultats (*cf.* table 4) sont principalement en accord avec ceux produits avec CounterFact (Meng *et al.*, 2023). Ceci valide la qualité des triplets et verbalisations dans notre corpus. Sur un autre plan, il est intéressant de noter que cette similitude tend à démontrer que le réalisme des mises à jours (comme dans WikiFactDiff par rapport à CounterFact) n'est peut-être pas important pour comparer globalement des approches entre elles. Plus en détails, la méthode FT généralise bien mais ne parvient pas complètement à maintenir la spécificité et la fluidité. En revanche, FT+L ne

5. Les faits aléatoires sont échantillonnés uniformément à partir de l'union des faits voisins de toutes les instances dans WFD_{remp}.

provoque pas de débordement mais ne parvient pas à généraliser sur les phrases alternatives. Bien que ROME soit globalement le meilleur algorithme, l'écart avec MEND n'est pas aussi prononcé que dans CounterFact, notamment en termes de spécificité et de généralisation. Nous notons également que l'efficacité de FT+L n'est pas aussi élevée que dans CounterFact.

Enfin, PROMPT est compétitif avec l'état de l'art sur toutes les métriques, à l'exception du débordement sur des triplets aléatoires (cette particularité est commentée plus en détail dans la section 5.2). Étant donné que la solicitation de faits sans rapport avec le préfixe utilisé est rarement effectuée dans la pratique, le débordement de PROMPT sur des faits aléatoires ne constitue pas une faiblesse critique de la méthode. Cependant, il faut garder à l'esprit que l'insertion de connaissances dans un GML à l'aide de préfixes est limitée par la taille du contexte, contrairement aux méthodes qui mettent à jour les paramètres du modèle.

5.2 Efficacité de notre recherche des faits voisins pour la détection de débordement

Pour toutes les méthodes sauf PROMPT, nous observons que les scores de débordement sont plus élevés sur les K triplets les plus proches que sur les triplets aléatoires. Le débordement plus élevé de PROMPT sur des voisins aléatoires pourrait s'expliquer comme suit : préfixer la phrase à trou par un fait totalement indépendant crée un contexte inhabituel au regard des données d'entraînement du modèle (p. ex., *"My Mister has a duration of 90 min. Langenbach has a population of ___"*), ce qui perturbe son comportement à l'inférence.

Le fait que tous les algorithmes reposant sur la mise à jour des paramètres GML aient un débordement significativement plus élevé sur les K voisins les plus proches confirme la pertinence de cette approche. Cependant, le fait que PROMPT soit sujet à des débordements sur des voisins aléatoires suggère que cette dernière métrique pourrait être une mesure complémentaire utile.

Enfin, il existe des améliorations possibles à notre méthode des plus proches triplets. En effet, les connaissances des modèles de langue sont biaisées vers des sujets populaires (Kandpal *et al.*, 2023), nous pouvons donc soupçonner que la popularité d'une entité a une certaine influence sur la magnitude du débordement. Nous avons mesuré cette influence en calculant le débordement moyen sur les triplets, en fonction de la popularité de leur sujet et de leur similarité avec le triplet mis à jour (*cf.* figure 2). Il apparaît que les deux facteurs ont une influence positive sur la probabilité de débordement. Plus un sujet est populaire et similaire au sujet édité, plus il est probable que des débordements se produiront. Cependant, il existe des cas où la similarité est faible mais où des débordements se produisent lorsque le sujet est populaire. Pour les recherches futures, notre méthode de sélection de voisins pour la détection des débordements pourrait ainsi être améliorée en intégrant la popularité du sujet.

6 Conclusion et perspectives

Dans cet article, nous avons présenté WikiFactDiff, un nouveau jeu de données de mise à jour des connaissances contenant des changements de connaissances factuelles sur une période donnée. Il élargit considérablement la gamme des faits considérés et les scénarios précédemment proposés dans la littérature (présence de littéraux ; réalisme ; insertion d'entités ; etc.). Notre jeu de données est accessible avec tout le matériel nécessaire pour exécuter et évaluer les algorithmes de mise à jour. De

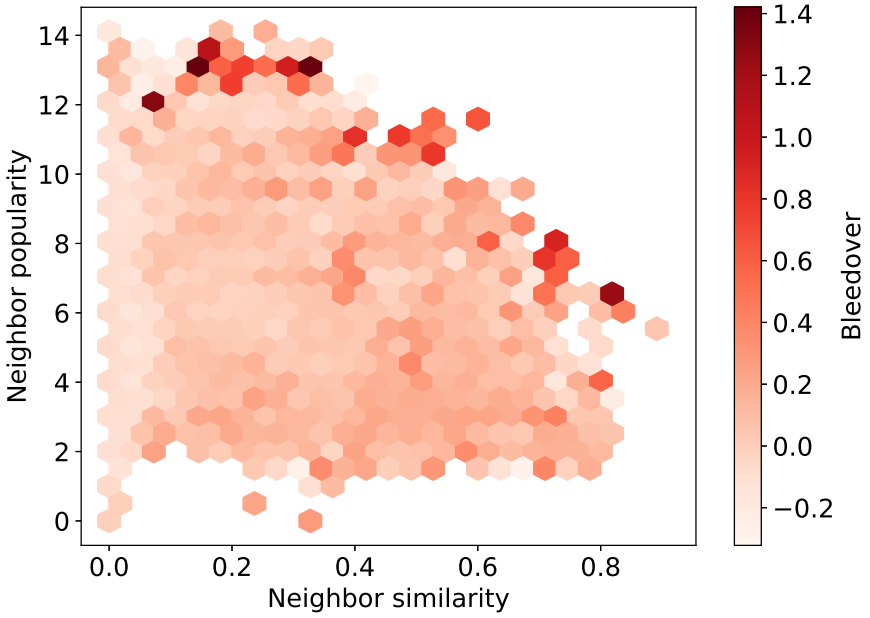


FIGURE 2 – Impact moyen de la popularité et de la similarité des voisins sur le débordement. Cette métrique est normalisée pour chaque algorithme afin d'atténuer la variance entre eux.

plus, le processus de création de corpus est adaptable à de nouvelles périodes.

WikiFactDiff introduit ainsi de multiples pistes de recherche pour le futur. Tout d'abord, les méthodes de mise à jour actuelles ne considérant que le scénario de remplacement (pour lequel des résultats ont été fournis dans cet article), développer des méthodes pour les autres scénarios de mise à jour (comme proposé par WikiFactDiff) ou évaluer comment celles existantes généralisent est une question majeure. Une autre piste est la mise à jour simultanée de multiples connaissances. Les meilleurs algorithmes connus sont efficaces jusqu'à quelques milliers de mises à jour avant de rencontrer des problèmes. Comme WikiFactDiff compte 224 000 mises à jour, il s'agit d'un terrain d'expérimentation propice. Enfin, comme les faits de WikiFactDiff sont dérivés de la réalité, il est naturel de s'attendre à ce que le modèle se souvienne des faits passés. Cette problématique de recherche est encore (à notre connaissance) négligée dans la communauté, et à cet égard, WikiFactDiff introduit d'autres scénarios complexes pour des travaux futurs.

Références

BIDERMAN S., SCHOELKOPF H., ANTHONY Q. G., BRADLEY H., O'BRIEN K., HALLAHAN E., KHAN M. A., PUROHIT S., PRASHANTH U. S., RAFF E., SKOWRON A., SUTAWIKA L. & VAN DER WAL O. (2023). Pythia : A suite for analyzing large language models across training and scaling. In A. KRAUSE, E. BRUNSKILL, K. CHO, B. ENGELHARDT, S. SABATO & J. SCARLETT, Éds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 de *Proceedings of Machine Learning Research*, p. 2397–2430 : PMLR.

BLACK S., BIDERMAN S., HALLAHAN E., ANTHONY Q., GAO L., GOLDING L., HE H., LEAHY C., MCDONELL K., PHANG J., PIELER M., PRASHANTH U. S., PUROHIT S., REYNOLDS L.,

TOW J., WANG B. & WEINBACH S. (2022). GPT-NeoX-20B : An open-source autoregressive language model. In A. FAN, S. ILIC, T. WOLF & M. GALLÉ, Éds., *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, p. 95–136, virtual+Dublin : Association for Computational Linguistics. DOI : 10.18653/v1/2022.bigscience-1.9.

BLACK S., GAO L., WANG P., LEAHY C. & BIDERMAN S. (2021). GPT-Neo : Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. DOI : 10.5281/zenodo.5297715.

DAI D., DONG L., HAO Y., SUI Z., CHANG B. & WEI F. (2022). Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 8493–8502, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.581.

DE CAO N., AZIZ W. & TITOV I. (2021). Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 6491–6506, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.522.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DHINGRA B., COLE J. R., EISENSCHLOS J. M., GILLICK D., EISENSTEIN J. & COHEN W. W. (2022). Time-aware language models as temporal knowledge bases. *Trans. Assoc. Comput. Linguistics*, **10**, 257–273. DOI : 10.1162/tacl_a_00459.

DONG Q., DAI D., SONG Y., XU J., SUI Z. & LI L. (2022). Calibrating factual knowledge in pretrained language models. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, p. 5937–5947 : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-emnlp.438.

GAO L., BIDERMAN S., BLACK S., GOLDING L., HOPPE T., FOSTER C., PHANG J., HE H., THITE A., NABESHIMA N., PRESSER S. & LEAHY C. (2021). The pile : An 800gb dataset of diverse text for language modeling. *CoRR*, **abs/2101.00027**.

JANG J., YE S., LEE C., YANG S., SHIN J., HAN J., KIM G. & SEO M. (2022a). TemporalWiki : A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 6237–6250, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.418.

JANG J., YE S., YANG S., SHIN J., HAN J., KIM G., CHOI S. J. & SEO M. (2022b). Towards continual knowledge learning of language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022* : OpenReview.net.

KANDPAL N., DENG H., ROBERTS A., WALLACE E. & RAFFEL C. (2023). Large language models struggle to learn long-tail knowledge. In A. KRAUSE, E. BRUNSKILL, K. CHO, B. ENGELHARDT, S. SABATO & J. SCARLETT, Éds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 de *Proceedings of Machine Learning Research*, p. 15696–15707 : PMLR.

LAZARIDOU A., KUNCORO A., GRIBOVSKAYA E., AGRAWAL D., LISKA A., TERZI T., GIMENEZ M., DE MASSON D'AUTUME C., KOCISKÝ T., RUDER S., YOGATAMA D., CAO K., YOUNG S. & BLUNSOM P. (2021). Mind the gap : Assessing temporal generalization in neural language models. In M. RANZATO, A. BEYGELZIMER, Y. N. DAUPHIN, P. LIANG & J. W. VAUGHAN, Éds., *Advances in Neural Information Processing Systems 34 : Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, p. 29348–29363.

LEVY O., SEO M., CHOI E. & ZETTLEMOYER L. (2017). Zero-shot relation extraction via reading comprehension. In R. LEVY & L. SPECIA, Éds., *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, p. 333–342 : Association for Computational Linguistics. DOI : 10.18653/v1/K17-1034.

LEWIS P. S. H., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W., ROCKTÄSCHEL T., RIEDEL S. & KIELA D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems 33 : Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

LIVSKA A., KOVCISK'Y T., GRIBOVSKAYA E., TERZI T., SEZENER E., AGRAWAL D., DE MASSON D'AUTUME C., SCHOLTES T., ZAHEER M., YOUNG S., GILSENAN-MCMAHON E., AUSTIN S., BLUNSOM P. & LAZARIDOU A. (2022). Streamingqa : A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*.

MENG K., BAU D., ANDONIAN A. & BELINKOV Y. (2022). Locating and editing factual associations in GPT. In *NeurIPS*.

MENG K., SHARMA A. S., ANDONIAN A. J., BELINKOV Y. & BAU D. (2023). Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023* : OpenReview.net.

MITCHELL E., LIN C., BOSSELUT A., FINN C. & MANNING C. D. (2022). Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022* : OpenReview.net.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D. & SUTSKEVER I. (2018). Language models are unsupervised multitask learners.

SI C., GAN Z., YANG Z., WANG S., WANG J., BOYD-GRABER J. L. & WANG L. (2023). Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023* : OpenReview.net.

SINITSIN A., PLOKHOTNYUK V., PYRKIN D. V., POPOV S. & BABENKO A. (2020). Editable neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020* : OpenReview.net.

VIG J., GEHRMANN S., BELINKOV Y., QIAN S., NEVO D., SINGER Y. & SHIEBER S. M. (2020). Investigating gender bias in language models using causal mediation analysis. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems 33 : Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

WANG B. & KOMATSUZAKI A. (2021). GPT-J-6B : A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

WANG R., TANG D., DUAN N., WEI Z., HUANG X., JI J., CAO G., JIANG D. & ZHOU M. (2021). K-Adapter : Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 1405–1418, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.121.

YAO Y., WANG P., TIAN B., CHENG S., LI Z., DENG S., CHEN H. & ZHANG N. (2023). Editing large language models : Problems, methods, and opportunities. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 10222–10240, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.632.

ZHANG Y., GALLEY M., GAO J., GAN Z., LI X., BROCKETT C. & DOLAN B. (2018). Generating informative and diverse conversational responses via adversarial information maximization. In S. BENGIO, H. M. WALLACH, H. LAROCHELLE, K. GRAUMAN, N. CESA-BIANCHI & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 31 : Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, p. 1815–1825.

ZHU C., RAWAT A. S., ZAHEER M., BHOJANAPALLI S., LI D., YU F. X. & KUMAR S. (2020). Modifying memories in transformer models. *CoRR*, **abs/2012.00363**.

A Processus détaillé de la création de WikiFactDiff

La Figure 1 illustre le processus de création de WikiFactDict. Basé sur 2 instances brutes de Wikidata aux instants T_{anc} et T_{nouv}, notés $\mathbf{W}_{anc}$ et $\mathbf{W}_{nouv}$, les étapes clés suivantes sont traitées :

1. **Prétraitement et différence** : les triplets dans les dumps Wikidata sont nettoyés et filtrés pour éliminer les informations et métadonnées non pertinentes, et une différence naïve est calculée entre tous les groupes (s, r) dans ces deux instantanés prétraités. Cela entraîne un partitionnement de tous les triplets en $\mathbf{F}^-$, $\mathbf{F}^+$ et $\mathbf{F}^0$, qu'ils appartiennent à l'ancien ou au nouveau dump, ou aux deux.

2. **Détection de nouvelles entités** Pour déterminer les scénarios de mise à jour, les nouvelles entités apparues pendant la période $[T_{anc}, T_{nouv}]$ sont repérées à partir de $\mathbf{F}^+$.

3. **Règles de classification** : Tous les triplets de $\mathbf{F}^-$ et $\mathbf{F}^+$ sont filtrés à l'aide de règles élaborées manuellement pour les étiqueter avec « nouveau », « obsolète », ou « inchangé » (section 3). Cela permet également de supprimer les groupes (s, r) où la nature des changements n'est pas tout à fait claire. Il s'agit de garantir que les changements factuels retenus reflètent des changements dans le monde réel.

4. **Recherche de faits voisins** : Pour tous les triplets retenus après tout le filtrage, des triplets sémantiquement proches sont identifiés. Cette étape est cruciale pour évaluer les métriques de spécificité des algorithmes de mise à jour.

5. **Verbalisation** : Enfin, des phrases d'injection et des phrases à trou doivent être générées pour effectuer les mises à jour factuelles applicables par les algorithmes existants et évaluer leurs performances sur WikiFactDiff.

A.1 Prétraitement et différence entre les instantanés Wikidata

Le prétraitement de $\mathbf{W}_{anc}$ et $\mathbf{W}_{nouv}$ se compose de plusieurs étapes, chaque étape filtrant une partie des données d'origine.

Triplets vs informations supplémentaires. Nous divisons les instances Wikidata en deux parties : les faits de base formalisés sous forme de triplets tels que (*Elizabeth II*, *position occupée*, *Chef du Commonwealth*), et toutes les informations supplémentaires entourant ces faits, appelées qualificatifs.

Dans Wikidata, les qualificatifs permettent de développer, d'annoter ou de contextualiser des triplets représentant des faits simples. Par exemple, les qualificatifs "*start time*" et "*end time*" permettent de préciser la période de validité d'un fait. Par exemple, (*Elizabeth II*, *position occupée*, *Chef du Commonwealth*) est un triplet Wikidata avec les qualificatifs *start time* et *end time* avec les valeurs « 6 février 1952 » et « 8 septembre 2022 », respectivement. Trois qualificatifs temporels sont pris en compte lors de la création de l'ensemble de données : *start time*, *end time* et *point in time*.

Nous incluons uniquement les faits de base dans WikiFactDiff. Les informations supplémentaires sont toutefois utilisées à certaines étapes de la création du jeu de données.

Triplets restreints. Les qualificatifs temporels sont des exemples de *qualificatifs restrictifs* [6], c'est-à-dire des *"qualificatifs qui restreignent ou modifient le référent du sujet, sans quoi la déclaration peut être inexacte ou dénuée de sens"*. Par exemple, le triplet (Se7en, *review score*, 83%) est incomplet sans la qualification : *reviewed by : Rotten Tomatoes*. Nous supprimons tous les triplets avec un qualificatif restrictif autre que *point dans le temps*, *start time* ou *end time*.

Triplets peu fiables. Dans Wikidata, un rang [7] (*preferred*, *normal* ou *deprecated*) est joint à chaque triplet pour évaluer sa pertinence. Nous filtrons tous les triplets dont le rang est *deprecated*.

Méta-triplets. Certaines relations Wikidata sont du type *'Wikidata property about Wikimedia entities'* : ce sont des méta-relations, utilisées pour la gestion des projets Wikimédia. Par conséquent, nous supprimons les triplets qui ont une méta-relation.

Valeurs d'objet non pertinentes ou inexploitables. Nous filtrons les triplets dont l'objet est une URL ou un identifiant externe identifiant une entité dans une autre base de connaissances. En règle générale, les URL sont des liens vers des sites Web externes ou vers divers fichiers multimédias communs (par exemple, images, vidéos, documents, fichiers audio, etc.).

Les triplets dont l'objet sont les coordonnées du globe sont également filtrés, car ils ont provoqué des divergences lors du calcul de la différence entre $\mathbf{W}_{anc}$ et $\mathbf{W}_{nouv}$: des écarts mineurs résultant de la précision en virgule flottante conduisent souvent à une classification erronée de coordonnées égales comme étant distinctes.

Enfin, les triplets dont l'objet est *'some value'* ou *'no value'* sont également filtrés.

Entités non pertinentes. Notre objectif est de conserver uniquement les triplets qui concernent des entités réelles du monde. Pour identifier de telles entités, nous nous appuyons sur Wikipédia : nous supposons qu'une entité n'est pertinente que si elle dispose d'un article Wikipédia dédié (ou rarement, d'une partie d'un article Wikipédia). De plus, cette page ne doit pas être une liste, une catégorie, un modèle ou une page d'homonymie. Les entités qui ne remplissent pas cette condition sont filtrées ; tous les triplets contenant une de ces entités sont également supprimés.

6. https://www.wikidata.org/wiki/Q61719275
7. https://www.wikidata.org/wiki/Help:Ranking

Valeurs obsolètes dans les relations fonctionnelles temporelles. Nous appelons une relation r **fonctionnelle temporelle** si pour chaque sujet s dans le graphe de connaissances, le groupe (s, r) ne peut posséder qu'un seul élément lorsqu'il est contextualisé dans un certain point temporel. Par exemple, la relation « population » est une relation fonctionnelle temporelle, car il ne peut y avoir qu'une seule valeur pour la population dans un lieu à un moment donné. D'autres relations fonctionnelles temporelles sont : l'espérance de vie, la capitale d'un pays, le chef de l'Etat, etc. D'après cette définition, si une relation est fonctionnelle, alors elle est fonctionnelle temporelle.

Étape de filtrage : Si la relation r d'un (s, r)-groupe est fonctionnelle temporelle avec un qualificatif 'point in time', on garde le triplet le plus à jour dans ce groupe et nous supprimons le reste. De cette manière, nous conservons les informations les plus à jour pour chaque version de Wikidata. Si un triplet du groupe ne contient pas de qualificatif *point in time*, on ne garde que le triplet de rang *'preferred'* s'il existe.

Nous associons à chaque entité un indicateur de popularité, basé sur le nombre de visites humaines sur son article Wikipédia dans les mois précédant T_{nouv}. L'idée est de permettre à la communauté d'étudier comment les performances des algorithmes varient en fonction de cet indicateur. Dans l'ensemble de données, les (s, r)-groupes sont triés par ordre décroissant en fonction de la popularité de leur sujet s.

Tous les faits dans $\mathbf{W}_{anc}^{Pré}$ et $\mathbf{W}_{nouv}^{Pré}$ sont des triplets fiables (s, r, o) avec des informations temporelles facultatives $[t_{start}, t_{end}]$. Si t_{start} ou t_{end} n'est pas défini, les valeurs $-\infty$ et $+\infty$ leur sont respectivement affectées. Pour les faits avec des informations ponctuelles dans le temps t (p. ex., populations), l'intervalle de temps est fixé à $[t, +\infty]$.

Enfin, l'intersection et la différence sur les ensembles de triplets provenant de $\mathbf{W}_{anc}^{Pré}$ et de $\mathbf{W}_{nouv}^{Pré}$ sont calculées pour produire les ensembles complémentaires $\mathbf{F}^-$ (qui sont uniquement en $\mathbf{W}_{anc}^{Pré}$), $\mathbf{F}^+$ (qui sont uniquement en $\mathbf{W}_{nouv}^{Pré}$), et $\mathbf{F}^0$ (qui sont à la fois en $\mathbf{W}_{anc}^{Pré}$ et $\mathbf{W}_{nouveau}^{Pré}$).

A.2 Détection de nouvelles entités

Les nouvelles entités sont des objets tangibles ou intangibles qui n'existaient pas avant T_{anc}. Des exemples notables incluent *ChatGPT*, *L'invasion russe de l'Ukraine en 2022*, *Lilibet of Sussex*, entre autres. Cet ensemble pourrait constituer une référence pour l'insertion d'entités dans les modèles de langage. Les nouvelles entités sont toutes les entités e telles que : (i) e n'est présent que dans $\mathbf{F}^+$; (ii) il existe un triplet (e, r, d) où r est une relation désignant la date de création [8] de e, et d est une date telle que $d > T_{anc}$. La condition (ii) est nécessaire car certains faits peuvent être antérieurs à T_{anc} mais le fait manquait dans $\mathbf{W}_{anc}$. Si elles ne sont pas rejetées, les expériences de mise à jour des connaissances peuvent être biaisées car le fait pourrait apparaître dans les données d'entraînement du GML à mettre à jour.

8. En pratique, ces relations sont 'inception', 'date of birth', 'start time' , 'date of discovery or invention', 'date of official opening', 'announcement date', 'point in time' et 'publication date'.

A.3 Règles de classification

Tous les triplets de $\mathbf{F}^-$, $\mathbf{F}^0$ et $\mathbf{F}^+$ sont classés à l'aide de règles afin de spécifier leurs étiquettes. En plus des étiquettes « nouveau », « obsolète » et « garder » définis dans la section 3, deux autres étiquettes techniques sont introduites :

— 'ignorer' s'applique aux faits qui ne sont ni corrects au moment T_{anc}, ni T_{nouv}. C'est généralement le cas pour les faits avec un intervalle de validité $[t_{start}, t_{end}] \subset [T_{anc}, T_{nouv}]$.

— 'inconnu' est une étiquette par défaut, attribuée lorsqu'aucune autre étiquette ne peut être attribuée sur la base de nos règles d'étiquetage.

Ces étiquettes sont des informations clés car leur distribution au sein d'un (s, r)-groupe donné détermine le scénario de mise à jour. Par exemple, un groupe de 2 triplets, l'un étiqueté « obsolète » et l'autre étiqueté « nouveau », correspond à un scénario de remplacement, similaire au (s, r)-groupe (États-Unis, chef du gouvernement) de la table 1.

La table 3 répertorie les variables et prédicats utilisés dans les règles de classification de chaque triplet (s, r, o), la table 5 décrit ces règles. Pour un triplet (s, r, o) donné, les règles sont testées dans l'ordre dans lequel elles apparaissent dans la liste. Dès qu'une règle est évaluée comme vraie, le triplet se voit attribuer la classe correspondante. Si aucune règle ne peut être appliquée, le triplet reste dans la classe « inconnu ».

Caractéristique	**Description**
t_{start}	Qualificatif "start time"
t_{end}	Qualificatif "end time"
$e \in \mathbf{E}^+$	L'entité e est-elle nouvelle ?
$e \notin \mathbf{F}^x$	e apparait-elle dans un triplet de $\mathbf{F}^x$?
r is death	r est-elle la relation 'date of death' ou 'date of burial or cremation' ?
r is temporal	r est-elle temporelle fonctionnelle ?
n^-, n^0, n^+	Nombre de triplets dans le (s, r)-group qui sont dans $\mathbf{F}^-$, $\mathbf{F}^0$, and $\mathbf{F}^+$, resp.
n	Total number of triplets of the (s, r)-group

FIGURE 3 – Caractéristiques du triplet (s, r, o)

Enfin, une étape supplémentaire est effectuée sur chaque triplet des (s, r)-groupes de taille 2 ($n = 2$) où r est une relation fonctionnelle temporelle. Étant donné la paire de triplets (s, r, o_1) et (s, r, o_2), si l'un d'eux est étiqueté comme « nouveau » et l'autre est dans $\mathbf{F}^-$, cet autre se voit attribuer la classe « obsolète ».

À la fin de cette procédure, tous les groupes (s, r) avec au moins un triplet étiqueté comme « inconnu » sont écartés pour ne prendre en compte que les mises à jour des connaissances où le changement est parfaitement compris. Ensuite, tous les triplets étiquetés « ignorer » sont supprimés des groupes restants. Enfin, les groupes dont tous les triplets sont étiquetés avec la classe « inchangé » sont également filtrés. Le résultat est une collection de (s, r)-groupes où au moins un triplet est soit « nouveau », soit « obsolète ».

Condition	Class
$s \in \mathbf{E}^+$	nouveau
$e \notin \mathbf{F}^-$	inconnu
r is death $\wedge$ $(s,r,o) \in \mathbf{F}^+ \wedge n = 1 \wedge T_{anc} < o < T_{nouv}$	nouveau
r is death $\wedge$ $\neg((s,r,o) \in \mathbf{F}^+ \wedge n = 1 \wedge T_{anc} < o < T_{nouv})$	inconnu
$t_{start} > t_{end}$	inconnu
r is temporal $\wedge$ $n^- = 1 \wedge n^+ = 1 \wedge n^0 = 0 \wedge (s,r,o) \in \mathbf{F}^+ \wedge T_{anc} < t_{start} < T_{nouv}$	nouveau
r is temporal $\wedge$ $n^- = 1 \wedge n^+ = 1 \wedge n^0 = 0 \wedge (s,r,o) \in \mathbf{F}^+ \wedge T_{anc} < t_{start} < T_{nouv} \wedge (t_{end} = +\infty \vee t_{end} > T_{nouv})$	nouveau
$(s,r,o) \in \mathbf{F}^+ \wedge T_{anc} < t_{start} < T_{nouv}$	nouveau
$t_{end} < T_{anc}$	ignorer
$t_{end} = +\infty \wedge t_{start} < T_{anc}$	inchangé
$t_{end} = +\infty \wedge T_{anc} < t_{start} < T_{nouv}$	nouveau
$t_{start} > T_{anc}$	ignorer
$T_{anc} < t_{start} < T_{nouv} \wedge T_{anc} < t_{end} < T_{nouv}$	ignorer
$t_{start} < T_{anc} \wedge T_{anc} < t_{end} < T_{nouv}$	obsolète
$t_{start} < T_{anc} \wedge t_{end} > T_{nouv}$	inchangé
$T_{anc} < t_{end} < T_{nouv}$	obsolète
$t_{end} > T_{nouv}$	inchangé
$(s,r,o) \in \mathbf{F}^- \wedge t_{end} < T_{anc}$	ignorer
$(s,r,o) \in \mathbf{F}^+ \wedge o \in \mathbf{E}^+$	nouveau

TABLE 5 – Liste des règles de classification pour un triplet (s,r,o)

A.4 Recherche des faits voisins

Lorsqu'un fait est mis à jour, la distribution de probabilité du modèle de langue est modifiée, ce qui peut dégrader sa précision sur d'autres faits. Ce phénomène est connu sous le nom de **débordement**. Pour permettre sa détection et sa mesure, WikiFactDiff est accompagné de faits voisins susceptibles d'être modifiés négativement lorsqu'un (s,r)-groupe donné est mis à jour. Cette section explique comment cela est effectué.

Il a été montré dans Meng *et al.* (2022) que, étant donné une mise à jour dans un scénario de remplacement à partir d'un fait (s,r,o), les faits avec une relation et un objet similaires (s',r,o) (avec $s' \neq s$) sont plus susceptibles d'être impactés que les faits aléatoires (pour lesquels aucune altération significative n'a été signalée). Par exemple, la mise à jour de (Albert Einstein, spécialité, *Physique*) vers (Albert Einstein, spécialité, *Biologie*) peut dégrader l'exactitude du modèle sur le fait (Isaac Newton, spécialité, Physique). La motivation est que parce que s et s' partagent la même paire relation-objet, leurs représentations latentes sont proches et donc les propriétés de s' sont plus susceptibles au débordement.

Cette idée ne peut pas être appliquée dans notre configuration car il n'est pas garanti qu'un fait (s',r,o) existe pour tous les (s,r,o) possibles. La raison est que, dans WikiFactDiff, les objets des

triplets ne se limitent pas aux entités. Ils peuvent aussi être des littéraux. Par exemple, si vous mettez à jour *(Seattle, population, 733.92K)*, trouver une entité proche de *Seattle* en utilisant la méthode de Meng signifie trouver une autre entité avec exactement la même population, ce qui est très peu probable. De plus, même dans le cas d'objets entités, la spécificité du sujet peut être telle qu'aucun triplet adéquat n'existe.

Pour contourner ce problème, les triplets voisins d'un triplet (s, r, o) sont définis comme des triplets (s', r, o') où s' est une entité similaire à s. Intuitivement, cette stratégie assouplit la contrainte sur o mais renforce celle sur s.

Concrètement, la similarité entre deux entités est calculée comme une similarité cosinus entre les vecteurs TF-IDF représentant chaque entité. Soit $\mathbf{W}^E_{anc}$ (resp. $\mathbf{W}^E_{nouv}$) l'ensemble de tous les triplets de $\mathbf{W}_{anc}$ (ou $\mathbf{W}_{nouv}$) dont l'objet est une entité (pas un littéral). Pour chaque entité s, une liste de caractéristiques $I(s)$ est construite comme suit

$$[s] \oplus [o \mid (s, r, o) \in \mathbf{W}^E_{anc}] \oplus [(r, o) \mid (s, r, o)\ dans \mathbf{W}^E_{anc}]$$

où $\oplus$ désigne l'opérateur de concaténation sur les listes. Si s n'est pas présent dans $\mathbf{W}^E_{anc}$, $I(s)$ est récupéré de $\mathbf{W}^E_{nouv}$ à la place. Ensuite, des représentations TF-IDF sont calculées pour toutes les entités s de WikiFactDiff, en considérant chaque représentation $I(s)$ comme un document.

Pour un triplet (s, r, o) donné, les k triplets les plus proches sont collectés en parcourant les n entités les plus similaires à s (en utilisant la similarité cosinus). Un unique triplet de la forme (s', r, o') dans $\mathbf{W}_{anc}$ est sélectionné pour chaque s' de cette liste, de manière itérative jusqu'à atteindre k triplets sélectionnés. Dans WikiFactDiff, les k-triplets les plus proches publiés ont été obtenus avec $k = 10$ et $n = 500$.

A.5 Verbalisation

Pour utiliser les algorithmes de mise à jour existants et les évaluer, des phrases d'injection et des phrases à trou doivent être générées pour tous les triplets de WikiFactDiff. Par exemple, en considérant le triplet *(France, capitale, Paris)*, une phrase d'injection possible est *"The capital of France is Paris"*, et les phrases à trou d'évaluation pourraient être *"The capital of France is ___"*, *"France's official capital is ___"* ou *"The capital of France is no other than ___"*. Notez que, puisque les phrases à trou sont conçues pour des modèles autorégressifs, le blanc doit être à la fin.

Les phrases d'injection et les phrases à trou sont générées sur la base de patrons où l'objet et le sujet sont manquants, tels que *The capital of ___ is ___*. La phrase à trou pour un triplet (s, r, o) peut être produite de manière triviale à partir d'un patron en remplissant le premier emplacement avec s. De même, la phrase d'injection est obtenue en injectant respectivement s et o dans chaque emplacement. Ainsi, disposer de patrons pour chaque relation est suffisant.

Les patrons sont créés comme suit. Tout d'abord, nous échantillonnons aléatoirement des triplets dont le sujet est l'une des 100 000 entités les plus populaires de $\mathbf{W}_{anc}$. Ensuite, pour chaque triplet (s, r, o), 10 verbalisations en anglais sont générées à l'aide de ChatGPT [9] (p. ex., *"The capital of France is Paris."*). Seules les verbalisations qui (i) contiennent des s, et (ii) se terminent par o sont conservées. Par conséquent, les patrons pour la relation r sont obtenus en remplaçant s et o par des trous. Cependant, tous les patrons ne sont pas suffisamment génériques pour être applicables à tous

9. GPT3.5 version 2023-03-15-preview

les triplets ayant la relation r. Par exemple, "*Danish actress* ___ was born in ___" ne s'applique que lorsque s est une actrice danoise. Pour filtrer ces patrons, nous conservons uniquement les 5 patrons les plus fréquents pour chaque relation, en partant de l'idée que les modèles qui s'appliquent à tous les triplets avec la relation r ont tendance à être générés plus fréquemment. Des exemples de phrases à trou construites à partir de ces modèles sont présentées dans la table 6.

Pair sujet-relation	**Phrase à trou**
(India, head of state)	India's head of state is __
(Google, employees)	The number of employees at Google is __
(Ukraine, BTI Status Index)	The BTI Status Index rated Ukraine at __
(Lionel Messi, head coach)	Lionel Messi's head coach is __
(Amazon, chief executive officer)	The CEO of Amazon is __
(Japan, age of majority)	In Japan, adulthood is recognized at __

TABLE 6 – Échantillons de phrases à trou de WikiFactDiff.

B Comment les relations fonctionnelles temporelles sont-elles identifiées ?

Les relations fonctionnelles temporelles sont identifiées à l'aide de la section « *property constraint* » d'une relation dans Wikidata.

Si une relation contient une contrainte qui est soit une « *single-value constraint* », soit une « *best-single-value constraint* », alors la relation est fonctionnelle. Si, en plus de cela, cette contrainte a un qualificatif « *separator* » avec une des valeures suivantes : « *start time* », « *end time* » ou « *point in time* », alors cette relation est fonctionnelle temporelle.

Par exemple, « *head of state* » est une relation fonctionnelle temporelle (www.wikidata.org/wiki/Property:P35)

C Préfixe utilisé pour la génération de phrases à trou avec ChatGPT

Voici le préfixe système (en anglais, *system prompt*) de ChatGPT utilisé pour la génération de phrases à trou :

```
You are an advanced knowledge triple verbalization system.
You take as input a knowledge triple (subject, relation, object) and
generate a list of 10 linguistically diverse verbalizations of the
triple.
For example, the input could be : (France, capital, Paris) and one of
your verbalizations may be : "The capital of France is Paris".

The veracity of the knowledge triple does not affect the quality of
your generation.

Examples of correct verbalizations:
- (Matriak, instance of, university) --> "Matriak is a university."
- (Johnathan Smith, date of death, 11-05-2012) --> "Johnathan Smith
died in 11-05-2012."
- (Tranquility Base Hotel & Casino, follows, AM) --> "Tranquility
Base Hotel & Casino follows AM."
- (Paris, named after, Parisii) --> "Paris was named after Parisii."
```

Et voici le préfixe principal :

```
Here is the knowledge triple to verbalize: ([SUB], [REL],
[OBJ]).
Your sentences should be concise and end with the term [OBJ].

Due to the ambiguity that could arise from the provided labels,
here is their meaning:
- (subject) "[SUB]" : "[SUB_DEF]"
- (relation) "[REL]" : "[REL_DEF]"
- (object) "[OBJ]" : "[OBJ_DEF]"

Finally, here is an example where the relation "[REL]" is
employed : ([EXP_SUB], [REL], [EXP_OBJ]).
```

Nous avons utilisé une génération avide avec une température égale à 0, pas de pénalité de fréquence, pas de pénalité de présence, et avec un nombre maximum de tokens générés égal à 800.

Deuxième partie

Articles présentés en session poster

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Adaptation des modèles de langue à des domaines de spécialité par un masquage sélectif fondé sur le genre et les caractéristiques thématiques

Anas Belfathi Ygor Gallina
Nicolas Hernandez Laura Monceaux Richard Dufour
LS2N, UMR CNRS 6004, Nantes Université
{firstname.lastname}@univ-nantes.fr

RÉSUMÉ

Les modèles de langue pré-entraînés ont permis de réaliser des avancées significatives dans diverses tâches de Traitement Automatique du Langage Naturel (TALN). Une des caractéristiques des modèles reposant sur une architecture Transformeur concerne la stratégie de masquage utilisée pour capturer les relations syntaxiques et sémantiques inhérentes à une langue. Dans les architectures de type encodeur, comme BERT, les mots à masquer sont choisis aléatoirement. Cette stratégie ne tient néanmoins pas compte des caractéristiques linguistiques spécifiques à un domaine. Dans ce travail, nous proposons de réaliser un masquage sélectif des mots en fonction de leur saillance thématique dans les documents dans lesquels ils se produisent et de leur spécificité au genre de document. Les performances des modèles résultant d'un pré-entraînement continu dans le domaine juridique soulignent l'efficacité de notre approche sur la majorité des tâches de *benchmark* LexGLUE en langue anglaise.

ABSTRACT

Language Model Adaptation to Specialized Domains through Selective Masking based on Genre and Topical Characteristics

Pre-trained language models have made significant advances in a variety of natural language processing (NLP) tasks. One of the key components of these models using the transformers architecture is their training based on the masking task, where encoder models, such as BERT, randomly select the tokens to be masked. This masking approach does not take into account domain-specific linguistic features. Thus, we propose a new approach to selective word masking with the aim of adapting language models to specialty domains. Our approach weights words according to their thematic salience or document genre specificity, then uses this weight to select words for masking. The performance of the models resulting from continuous pre-training in the legal domain underlines the effectiveness of our strategy on the majority of tasks of the English-language LexGLUE benchmark.

MOTS-CLÉS : Modèle de langue ; stratégie de masquage ; BERT ; méta-discours ; tfidf.

KEYWORDS: Language model ; masking strategy ; BERT ; metadiscourse ; tfidf.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 283–294.

1 Introduction

Les modèles de langue pré-entraînés à grande échelle sont devenus indispensables pour modéliser le langage humain, améliorant considérablement les performances dans diverses tâches de Traitement Automatique du Langage Naturel (TALN) (Bao *et al.*, 2020; Guu *et al.*, 2020; Zhang *et al.*, 2022). Toutefois, le pré-entraînement de tels modèles pour des domaines de spécialité n'est pas toujours possible car il requiert une importante quantité de données qui n'est pas toujours disponible, mais aussi en raison du coût de calcul élevé qu'il représente. Une solution réside ainsi dans le pré-entraînement continu (*Continual Pre-Training*, CPT) pour spécialiser des modèles de fondation (Chalkidis *et al.*, 2020; Wu *et al.*, 2021; Ke *et al.*, 2022; Labrak *et al.*, 2023).

Classiquement, les modèles ainsi construits réutilisent l'algorithme d'apprentissage utilisé lors du pré-entraînement initial. Dans les modèles qui reposent sur une architecture de type encodeur à la BERT (Devlin *et al.*, 2019), l'algorithme utilisé est le Modèle de Langage Masqué (*Masked Language Modeling*, MLM). Dans cet algorithme, le modèle apprend à prédire un mot masqué aléatoirement dans une séquence. Les modèles résultants de cet apprentissage ont montré leur capacité à capturer les relations sémantiques complexes et les structures syntaxiques inhérentes au langage naturel. Certains travaux ont étendu le MLM, notamment pour affiner la capacité des modèles à capturer des expressions multimots (Sun *et al.*, 2019; Joshi *et al.*, 2020; Levine *et al.*, 2020; Li & Zhao, 2021). Cependant, aucune étude n'a envisagé d'étendre le MLM pour prendre en compte la spécificité du genre de document étudié.

Dans cet article, nous présentons une nouvelle approche de masquage sélectif pour l'adaptation de modèles de langue à des domaines de spécialité qui exploite les informations relatives à la spécificité au genre de document et à la topicalité. Notre approche pondère les mots en fonction de leur spécificité au genre de documents (on parle aussi de "caractère méta-discursif") et de leur saillance thématique. Nous utilisons un score TF×IDF pour mesurer la saillance et proposons une formule pour estimer la spécificité. En utilisant ces scores pour choisir les mots à masquer, nous forçons le modèle à s'adapter à la compréhension et à la prédiction des mots spécifiques à un domaine cible. Pour illustrer l'efficacité de notre approche, nous menons des expériences sur le pré-entraînement continu de modèles type BERT dans le domaine juridique, en comparant différentes stratégies de masquage des mots.

Nos contributions sont les suivantes :

— Nous proposons une nouvelle approche de masquage fondée sur la sélection de mots (méta-discours et TF×IDF) pour l'entraînement de modèles de langue ;
— Nous effectuons une analyse comparative de plusieurs manières de sélectionner les mots pendant le processus d'entraînement ;
— Nous partageons nos modèles et notre code source pour rendre plus facile le pré-entraînement de modèles de langue pour des domaines de spécialité, selon notre stratégie d'apprentissage [1] ;

2 Travaux connexes

L'approche de masquage classique utilisé dans BERT (Devlin *et al.*, 2019) consiste à sélectionner 15 % de tokens dans une séquence. Chacun de ces 15 % de tokens est ensuite substitué par le token spécial `[MASK]` (80 % de probabilité), remplacé par un mot tiré aléatoirement (10 %) ou bien laissé

1. github.com/ygorg/legal-masking

intact (10 %). L'objectif du modèle est de prédire les mots masqués originaux.

Dans le but d'enrichir les capacités de représentation des MLMs, ERNIE (Sun *et al.*, 2019) et SpanBERT (Joshi *et al.*, 2020) affinent la stratégie de masquage classique (aléatoire) utilisée par BERT. Ces modèles masquent respectivement des mots entiers et des parties contiguës de texte de manière aléatoire. Ces approches améliorent les performances des tâches de réponse aux questions et de résolution de coréférence. Yang *et al.* (2023) observent qu'à partir d'un moment de l'entraînement, les modèles cessent d'apprendre à partir de types de mots spécifiques, identifiés par des étiquettes morphosyntaxiques (*POS tags*), en fonction de la stabilité de la fonction de coût de modèle. Ainsi, ils introduisent une approche de masquage qui varie dans le temps, cette approche diffère des méthodes statiques qui maintiennent un contenu inchangeable tout au long de l'apprentissage. Cette stratégie améliore les performances sur le *benchmark* GLUE (Wang *et al.*, 2018).

D'autres méthodes s'intéressent à adapter les modèles à des domaines spécifiques et choisissent donc les mots à masquer en tenant compte des spécificités du domaine des documents. Dans le domaine des brevets, Althammer *et al.* (2021) masque les mots qui apparaissent fréquemment dans des groupes nominaux. Cette approche montre son efficacité dans les tâches de classification. Dans le domaine clinique, EntityBERT (Lin *et al.*, 2021) masque des tokens identifiés comme "entité" (symptôme, médicament, date) par un modèle pré-entraîné de détection d'entités. L'application de cette stratégie repose néanmoins sur la disponibilité de modèles de détection d'entités du domaine cible.

Moins générique, certaines approches adaptent les modèles à une tâche spécifique avant d'effectuer un affinage. Par exemple, pour des tâches de classification de document, l'approche proposée par Golchin *et al.* (2023) utilise KeyBERT (Grootendorst, 2020) pour masquer uniquement les mots identifiés comme mots-clés. L'approche *Difference-Masking* (Wilf *et al.*, 2023) quant à elle, masque les mots les plus similaires (en termes de plongement de mots) à des mots "ancres". Les ancres sont des mots ayant une fréquence plus élevée dans le corpus que dans un corpus général. Bien que ces techniques soient similaires à celles faisant de l'adaptation au domaine, les résultats de ces techniques d'adaptation à la tâche ne sont pas directement comparables car moins généralisables.

Le choix du taux de masquage influence de manière significative la tâche de Modèle de Langage Masqué et donc le pré-entraînement. L'étude menée par Wettig *et al.* (2023) indique que le taux de masquage optimal dépend de la taille du modèle à entraîner, 40% étant optimal pour les modèles larges et 20% pour les modèles de base, comme l'ont montré les évaluations des *benchmarks* GLUE et SQuAD.

L'approche présentée ici s'intéresse à sélectionner les mots à masquer en fonction de leur saillance thématique et de leur spécificité au genre de document. Elle s'inscrit dans les travaux d'adaptation au domaine et ainsi est indépendante des tâches en aval.

3 Stratégie de masquage fondée sur le genre et la topicalité

Contrairement à l'approche originale de BERT, qui sélectionne aléatoirement les tokens à masquer (Devlin *et al.*, 2019), notre approche se concentre sur le masquage au niveau des mots, les choisissant en fonction de leur spécificité par rapport au genre du texte ou bien à leur saillance thématique dans un document. Notre approche fonctionne en deux étapes. Tout d'abord, nous attribuons un *score de spécificité au genre* et un *score de saillance thématique* à chaque mot à partir de notre corpus spécifique à un domaine (Section 3.1). Ensuite, nous utilisons ces scores pour hiérarchiser et

sélectionner les mots à masquer (Section 3.2).

3.1 Pondération des mots

Nous proposons deux manières de pondérer les mots à partir d'un corpus de documents de spécialité. La première approche, le *score de topicalité* (TF×IDF), quantifie la saillance thématique d'un mot dans un document donné. Pour cela, nous utilisons la mesure classique TF×IDF (Jones, 1972), qui pondère un mot en fonction de son nombre d'occurrences dans un document particulier et du nombre de documents dans lequel il apparaît dans le corpus.

La seconde approche, le *score de spécificité au genre de texte* (`MetaDis`), évalue dans quelle mesure un mot est caractéristique d'un genre de documents. Un genre de documents est caractérisé par une structure commune (Biber & Conrad, 2019; Hyland, 1998). Par exemple, les jurisprudences (domaine juridique) présentent successivement des faits, puis des arguments et enfin un raisonnement pour parvenir à une décision. Chacune de ces parties utilise un lexique particulier à ce genre, nous désignons ce lexique par le terme de *méta-discours*. Bien que Hernandez & Grau (2003) ait utilisé le score de fréquence inverse de document pour évaluer la spécificité, leur mesure ne tient pas compte de la distribution des occurrences dans les documents. Nous supposons qu'un indicateur de méta-discours est présent dans une proportion constante dans les documents du même genre. Pour capturer de telles propriétés et calculer un score de méta-discours, nous proposons la formule décrite dans l'équation 1 :

$$s_t = \frac{df_t}{tf_t} * \left(1 - \frac{std(tf_{d,t})}{max(tf_{d,t})}\right) * \frac{df_t}{N} \quad (1)$$

Ici, df_t représente le nombre de documents dans lequel le mot t apparaît, tf_t le nombre d'occurrences de t dans le corpus, $tf_{d,t}$ le nombre d'occurrences de t dans le document d et N le nombre de documents dans le corpus. L'intuition derrière le premier terme est de donner un haut score aux mots apparaissant dans de nombreux documents. Le second mesure la constance d'apparition du mot et donne un haut score à ceux qui ont la même fréquence dans tous les documents (faible $std(tf_{d,t})$), le dénominateur normalise par la fréquence maximale du mot. Enfin, le troisième terme donne un haut score aux mots apparaissant dans de nombreux documents.

3.2 Stratégie de sélection des mots à masquer

Nous proposons deux manières de sélectionner les mots à masquer qui utilisent les scores TF×IDF ou `Metadis`. Notre première méthode, `TopN`, sélectionne les mots ayant les scores les plus élevés. Cette méthode est déterministe, pour un document les mots masqués seront toujours les mêmes.

La seconde méthode, `Samp` (échantillonnage, *sample* en anglais), vise à améliorer la robustesse du modèle en évitant le masquage systématique des mêmes mots. Cette méthode s'inspire du masquage dynamique utilisé dans RoBERTa (Liu *et al.*, 2019) et introduit un niveau d'aléa pondéré qui change à chaque itération de modèle. En pratique, nous échantillonnons aléatoirement des mots (sans remise) sur la base de la distribution des scores calculés.

2. Pour la stratégie `TopN` la fonction Max est utilisée.

Algorithm 1 Masquage sélectif

function MASK($tokens$)
  $\mathcal{M} \leftarrow \{\}$
  $W \leftarrow MotsEntiers(tokens)$
  $S \leftarrow CalculeScore(W)$
  while $|\mathcal{M}| < 0.15 * |tokens|$ **do**
    $i \leftarrow$ Echantillone(S) [2]
    Supprime $W[i]$ and $S[i]$
    if $|\mathcal{M}| + |W[i]| \leq 0.15 * |tokens|$ **then**
      $\mathcal{M} \leftarrow \mathcal{M} + w$
    end if
  end while
  return $\mathcal{M}$
end function

Une fois les mots sélectionnés, nous choisissons les tokens à masquer suivant l'Algorithme 1 pour masquer effectivement 15 % (Devlin *et al.*, 2019) des tokens de la séquence. Dans l'algorithme, M dénote les mots à masquer, W les mots segmentés et S les scores associés à chaque mot.

4 Paramètres expérimentaux

Nous utilisons comme base les modèles pré-entraînés BERT (Devlin *et al.*, 2019) et LegalBERT (Chalkidis *et al.*, 2020). L'efficacité de notre approche de masquage est évaluée par un pré-entraînement continu sur ces modèles, en se concentrant sur l'adaptation au domaine juridique. Dans cette section, nous présentons d'abord les données utilisées pour le pré-entraînement continu (Section 4.1). Puis, les tâches d'évaluation (Section 4.2) et enfin les détails expérimentaux (Section 4.3).

4.1 Corpus de pré-entraînement

Pour le pré-entraînement continu et la sélection du masquage de mots, nous avons choisi de nous concentrer sur le domaine juridique en utilisant un sous-ensemble du corpus LexFiles (Chalkidis *et al.*, 2023) représentatif du *benchmark* LexGLUE (Chalkidis *et al.*, 2022). Les documents ont été sélectionnés pour offrir une collection équilibrée et diversifiée, englobant les nuances linguistiques (voir la Table 1).

Sous-Corpus	**# Doc.**	**# Tokens**
EU Case Law	29,8K	178,5M (29%)
ECtHR Case Law	12,5K	78,5M (13%)
U.S. Case Law	104,7K	235,5M (39%)
Indian Case Law	34,8K	111,6M (19%)
Total	**181,8K**	**604,1M**

TABLE 1 – Détails de l'ensemble de données utilisé pour le pré-entraînement continu.

4.2 Tâches d'évaluation

Nous évaluons la performance de nos modèles à l'aide de LexGLUE (Chalkidis *et al.*, 2022) un *benchmark* conçu spécifiquement pour le domaine juridique. LexGLUE englobe une gamme variée de tâches provenant des systèmes juridiques de l'Europe, des États-Unis et du Canada. Une telle configuration permet de tester rigoureusement la capacité de notre stratégie dans un spectre de tâches complexes. Voici un aperçu de chacune de ces tâches : [3]

— **ECtHR A & B** consistent à déterminer quels articles de lois sont enfreints (ECtHR A), ou prétendument enfreints (ECtHR B), par une liste de faits. Les articles de loi sont 11 articles de la Cour Européenne des Droits de l'Homme.
— **SCOTUS** consiste à classer les opinions de la Cour suprême des États-Unis (SCOTUS) parmi 14 thèmes, tels que la procédure pénale, les droits civils, l'activité économique…
— **EUR-LEX** concerne la législation de l'Union européenne publiée sur le portail EUR-Lex. L'objectif est d'assigner aux textes de loi les concepts du thésaurus EUROVOC pertinents (parmi les 100 concepts les plus fréquents).
— **LEDGAR**, Labeled EDGAR, consiste à trouver le thème principal (parmi 100) de chaque paragraphe des contrats de la base de données EDGAR.
— **UNFAIR-ToS** cherche à détecter les clauses de conditions de services qui enfreignent potentiellement le droit des consommateurs européens. Chaque phrase est classée parmi 9 types de clauses contractuelles abusives.

4.3 Détails expérimentaux

Pour le pré-entraînement continu, nous avons mené des sessions d'entraînement totalisant plus de 20 heures en utilisant 16 GPU V100 sur le supercalculateur Jean Zay. Suivant la méthode de Labrak *et al.* (2023), nous avons adopté une taille de batch de 16 et configuré les étapes d'accumulation de gradient à 16, ce qui donne une taille de batch effective de 4 096. Pour déterminer les scores de performance de la tâche, nous avons calculé la moyenne des scores de trois exécutions indépendantes. Les métriques utilisées sont les mesures : micro (μF_1) et macro (m-F1) F-mesure.

5 Résultats et discussions

Les résultats sont détaillés dans la Table 2, chaque colonne correspondant à une tâche du *benchmark* LexGLUE.

Effets du pré-entraînement continu Nos résultats montrent l'efficacité du pré-entraînement continu dans toutes les tâches pour les modèles BERT et LegalBERT. Plus précisément, la macro F1 (m-F1) de la configuration de référence BERT+CPT augmente sensiblement de 60,39 à 64,69 dans la tâche ECtHR (B). De manière similaire, LegalBERT+CPT montre des améliorations substantielles dans la tâche EUR-LEX au niveau de la m-F1. Ces améliorations, même sans modification de la stratégie de masquage, suggèrent que le corpus utilisé contient de nouvelles caractéristiques spécifiques au domaine permettant d'enrichir les connaissances des modèles de langue.

3. Un problème dans le code de la tâche `case_hold` empêche son exécution, nous ne rapportons donc pas les résultats.

Method	ECtHR (A)		ECtHR (B)		SCOTUS		EUR-LEX		LEDGAR		UNFAIR-ToS	
	μF_1	m-F1	μF_1	m-F1	μF_1	m-F1	μF_1	m-F1	μF_1	m-F1	μF_1	m-F1
BERT	62,12	52,66	69,59	60,39	69,61	58,65	71,70	54,87	87,85	**82,30**	**95,66**	80,97
+ CPT (référence)	63,12	54,13	71,06	**64,69**	70,57	**60,38**	71,86	56,18	**87,90**	82,02	95,56	81,46
+ MetaDis - Samp	62,55	54,88	70,45	63,10	70,26	59,12	71,66	56,00	87,68	82,25	95,38	79,18
+ MetaDis - TopN	62,17	53,35	70,29	62,29	69,92	60,08	71,67	56,95	87,78	82,11	95,57	**81,51**
+ TF×IDF - Samp	**63,36**	**56,60**	71,32	64,58	69,69	59,10	**71,93**	55,82	87,69	82,11	95,50	78,63
+ TF×IDF - TopN	62,66	56,46	**71,50**	63,58	**70,71**	60,06	71,73	**57,73**	87,67	81,89	95,49	79,45
LegalBERT	63,41	53,19	72,10	63,68	73,61	61,50	71,93	55,47	87,91	81,67	95,81	81,27
+ CPT (référence)	63,64	**58,73**	72,60	64,95	**74,64**	**63,13**	72,01	55,12	**88,41**	**82,92**	**95,82**	79,70
+ MetaDis - Samp	63,39	56,39	73,08	65,76	74,21	62,97	**72,03**	54,76	88,38	82,58	95,20	80,26
+ MetaDis - TopN	**64,07**	58,56	72,53	**66,83**	73,88	62,57	71,96	55,01	88,32	82,16	94,80	73,67
+ TF×IDF - Samp	63,38	56,78	72,21	65,67	73,71	62,85	71,78	**55,82**	88,19	82,36	95,80	**82,12**
+ TF×IDF - TopN	62,89	53,58	**73,26**	65,86	74,38	63,10	71,90	55,08	88,27	82,65	95,59	81,20

TABLE 2 – Performances des modèles BERT et LegalBERT de base, après pré-entraînement continu avec masquage aléatoire (+CPT) ou masquage sélectif (+MetaDis, +TF×IDF) sur le *benchmark* LexGLUE. La meilleure valeur de chaque colonne est indiquée en gras, la deuxième meilleure est soulignée. Les fonds Verts et Oranges représentent respectivement les scores MetaDis et TF×IDF, les fonds foncés montrent les améliorations par rapport à la référence.

Masquage sélectif vs classique Pour évaluer l'efficacité de nos masquages (MetaDis, TF×IDF), nous les comparons au masquage aléatoire classique de BERT (+CPT (référence)) après pré-entraînement continu. Les résultats indiquent que, quelle que soit la stratégie de masquage utilisée, au moins un de nos modèles apporte des améliorations par rapport à la référence. Avec BERT, des améliorations notables sont observées dans les tâches ECtHR(A) et LEDGAR, obtenant respectivement des scores de 54,88 et 82,25 de m-F1 pour MetaDis - Samp. Cela peut être attribué à la capacité de nos modèles à exploiter les informations relatives au genre (MetaDis) et à la thématique (TF×IDF) propres au domaine juridique. Pour les modèles LegalBERT, des améliorations ont été observées dans les tâches ECtHR(B) et UNFAIR-ToS avec les deux pondérations. Ces résultats soulignent les avantages d'un masquage sélectif des mots, en particulier avec des modèles déjà adaptés. Par rapport aux résultats rapportés dans Chalkidis *et al.* (2022), notre stratégie obtient, sur la tâche SCOTUS, des résultats supérieurs aux modèles hiérarchiques, tout en requérant moins de paramètres et une architecture moins complexe. Cela souligne l'intérêt de développer des techniques de pré-entraînement qui se concentrent sur des caractéristiques linguistiques spécifiques à un domaine. Cela permet ainsi de se passer de modèles complexes requérant des paramètres supplémentaires ou bien des temps d'entraînement plus longs.

Méta-discours vs. TF×IDF En comparant les stratégies de masquage pondérant le genre (+MetaDis) et la topicalité (+TF×IDF) sur les modèles BERT et LegalBERT, nous avons observé des tendances différentes. Pour les modèles BERT, le méta-discours a démontré son efficacité dans les tâches où les caractéristiques linguistiques spécifiques au genre jouent un rôle important, comme dans les tâches ECtHR (A), LEDGAR et UNFAIR-ToS. Au contraire, TF×IDF montre des améliorations pour les tâches qui concernent la pertinence thématique, comme les tâches ECtHR (B), EURLEX et SCOTUS. Par exemple, la tâche EURLEX se concentre sur les textes législatifs de l'Union européenne, marqués par une diversité de concepts issus d'EuroVoc [4], le thésaurus multilingue, soulignant

4. eur-lex.europa.eu/browse/eurovoc.html

ainsi leur riche contenu thématique.

Pour le modèle LegalBERT, les deux stratégies présentent des performances similaires pour les tâches ECtHR (B) et UNFAIR-ToS. En revanche, pour la tâche ECtHR (A) la pondération basée sur le meta-discours obtient de meilleures performances, pour EURLEX c'est la pondération TF×IDF. Ces résultats suggèrent que la pertinence thématique est un facteur déterminant pour la tâche EURLEX, quel que soit le modèle de base. Cependant, pour la tâche ECtHR (A) les aspects liés au genre (MetaDis) sont particulièrement pertinents, soulignant l'importance de la structure des documents pour cette tâche.

Samp vs. TopN La comparaison des configurations BERT sur les deux pondérations indique que la stratégie Samp (échantillonnage) permet d'obtenir de meilleures performances dans les tâches ECtHR (A) et LEDGAR. La stratégie TopN, quant à elle, apporte une amélioration pour la tâche EURLEX. Avec la pondération TF×IDF, la stratégie TopN obtient de meilleures performances pour les tâches ECtHR (B) et SCOTUS. Ainsi, TopN semble plus efficace lorsque la topicalité est un facteur déterminant.

En ce qui concerne les modèles LegalBERT utilisant la pondération MetaDis, la stratégie Samp permet d'améliorer les performances dans les tâches ECtHR (B), EURLEX et UNFAIR-ToS. La stratégie TopN, quant à elle, permet de réaliser des progrès notables dans les tâches ECtHR (A) et (B), ce qui souligne son intérêt pour les tâches nécessitant une compréhension thématique des documents. Toujours en partant de LegalBERT et en utilisant la pondération TF×IDF, des améliorations sont observées dans les tâches ECtHR (B) et UNFAIR-ToS avec les deux stratégies TopN et Samp. Notons que la stratégie Samp obtient les meilleurs résultats sur la tâche EURLEX.

6 Conclusion et perspectives

Nos expériences montrent que nos approches de masquage sélectif, qui intègrent les caractéristiques du genre et du thème du document, jouent un rôle crucial dans l'adaptation des modèles à un domaine de spécialité. Nous observons des améliorations dans chacune les tâches du *benchmark* LexGLUE axé sur le domaine juridique. En particulier, des améliorations importantes ont été obtenues sur les tâches ECtHR et EUR-LEX pour les modèles BERT et LegalBERT. Néanmoins, les améliorations obtenues ne sont pas consistantes pour chaque modèle et chaque tâche et indiquent que le choix de la pondération semble dépendant de la tâche. Ainsi, plusieurs axes de recherche émergent : tout d'abord, mesurer la capacité de notre approche à généraliser à d'autres domaines, tels que le domaine clinique ou scientifique. Ensuite, étudier l'impact de notre approche dans le cadre de l'adaptation à la tâche.

Considérations éthiques

Concernant les risques et les biais potentiels inhérents aux modèles de langue entraînés sur des ensembles de données juridiques, les corpus peuvent comprendre des textes de qualité et de représentativité variables. L'utilisation de modèles entraînés sur des textes juridiques, tels que BERT, pourrait introduire des biais liés à la justice, à l'utilisation d'un langage genré, à la représentation de groupes de minorités et à la nature dynamique des normes juridiques au fil du temps. Il est impératif

que ces biais soient évalués et atténués de manière approfondie afin de garantir des performances équitables entre les différents groupes démographiques et de rester en phase avec l'évolution des normes juridiques.

Limitations

Notre travail propose une nouvelle façon d'aborder le pré-entraînement continu des modèles de langue dans le domaine juridique avec un masquage sélectif, mais présente néanmoins certaines limitations. En particulier, l'étude actuelle se concentre uniquement sur l'architecture BERT, limitant notre capacité à étudier une gamme plus large de modèles de langue. Ces derniers pourraient en effet présenter des comportements distincts et différentes sensibilités à notre stratégie de pré-entraînement continu et de masquage. Les études futures devraient explorer d'autres modèles, tels que DrBERT (Labrak *et al.*, 2023) et RoBERTa (Liu *et al.*, 2019), afin de fournir une compréhension plus complète des effets de notre stratégie. En outre, notre étude manque d'une comparaison directe avec un modèle entraîné à partir de zéro (*from scratch*) utilisant le masquage sélectif. Une telle comparaison constituerait un point de référence précieux permettant de déterminer d'autres avantages à notre méthode. Enfin, un réglage plus poussé des hyper-paramètres pourrait conduire à une amélioration des performances du modèle.

Remerciements

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2023-AD011014882 attribuée par GENCI.

Ce travail a été financé, en totalité ou en partie, par l'Agence Nationale de la Recherche (ANR), projet NR-22-CE38-0004.

Références

ALTHAMMER S., BUCKLEY M., HOFSTÄTTER S. & HANBURY A. (2021). Linguistically informed masking for representation learning in the patent domain. In *2nd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech2021)*, New York, NY, USA : Association for Computing Machinery.

BAO H., DONG L., WEI F., WANG W., YANG N., LIU X., WANG Y., PIAO S., GAO J., ZHOU M. & HON H.-W. (2020). Unilmv2 : Pseudo-masked language models for unified language model pre-training.

BIBER D. & CONRAD S. (2019). *Register, Genre, and Style*. Cambridge University Press. Google-Books-ID : x7OQDwAAQBAJ.

CHALKIDIS I., FERGADIOTIS M., MALAKASIOTIS P., ALETRAS N. & ANDROUTSOPOULOS I. (2020). LEGAL-BERT : The muppets straight out of law school. In T. COHN, Y. HE & Y. LIU, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 2898–2904, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.261.

CHALKIDIS I., GARNEAU N., GOANTA C., KATZ D. & SØGAARD A. (2023). LeXFiles and LegalLAMA : Facilitating English Multinational Legal Language Model Development. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 15513–15535, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.865.

CHALKIDIS I., JANA A., HARTUNG D., BOMMARITO M., ANDROUTSOPOULOS I., KATZ D. & ALETRAS N. (2022). LexGLUE : A Benchmark Dataset for Legal Language Understanding in English. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 4310–4330, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.297.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

GOLCHIN S., SURDEANU M., TAVABI N. & KIAPOUR A. (2023). Do not mask randomly : Effective domain-adaptive pre-training by masking in-domain keywords. In B. CAN, M. MOZES, S. CAHYAWIJAYA, N. SAPHRA, N. KASSNER, S. RAVFOGEL, A. RAVICHANDER, C. ZHAO, I. AUGENSTEIN, A. ROGERS, K. CHO, E. GREFENSTETTE & L. VOITA, Éds., *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, p. 13–21, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.repl4nlp-1.2.

GROOTENDORST M. (2020). Keybert : Minimal keyword extraction with bert. DOI : 10.5281/zenodo.4461265.

GUU K., LEE K., TUNG Z., PASUPAT P. & CHANG M.-W. (2020). Realm : Retrieval-augmented language model pre-training.

HERNANDEZ N. & GRAU B. (2003). Automatic extraction of meta-descriptors for text description. In *International Conference on Recent Advances In Natural Language Processing (RANLP)*, Borovets, Bulgaria.

HYLAND K. (1998). Persuasion and context : The pragmatics of academic metadiscourse. *Journal of pragmatics*, **30**(4), 437–455.

JONES K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, **28**, 11–21. DOI : 10.1108/eb026526.

JOSHI M., CHEN D., LIU Y., WELD D. S., ZETTLEMOYER L. & LEVY O. (2020). SpanBERT : Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, **8**, 64–77. DOI : 10.1162/tacl_a_00300.

KE Z., SHAO Y., LIN H., KONISHI T., KIM G. & LIU B. (2022). Continual pre-training of language models. In *The Eleventh International Conference on Learning Representations*.

LABRAK Y., BAZOGE A., DUFOUR R., ROUVIER M., MORIN E., DAILLE B. & GOURRAUD P.-A. (2023). DrBERT : A robust pre-trained model in French for biomedical and clinical domains. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 16207–16221, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.896.

LEVINE Y., LENZ B., LIEBER O., ABEND O., LEYTON-BROWN K., TENNENHOLTZ M. & SHOHAM Y. (2020). Pmi-masking : Principled masking of correlated spans.

LI Y. & ZHAO H. (2021). Pre-training universal language representation. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 5122–5133, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.398.

LIN C., MILLER T., DLIGACH D., BETHARD S. & SAVOVA G. (2021). EntityBERT : Entity-centric masking strategy for model pretraining for the clinical domain. In D. DEMNER-FUSHMAN, K. B. COHEN, S. ANANIADOU & J. TSUJII, Éds., *Proceedings of the 20th Workshop on Biomedical Language Processing*, p. 191–201, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.bionlp-1.21.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized BERT pretraining approach. *CoRR*, **abs/1907.11692**.

SUN Y., WANG S., LI Y., FENG S., CHEN X., ZHANG H., TIAN X., ZHU D., TIAN H. & WU H. (2019). ERNIE : Enhanced Representation through Knowledge Integration. arXiv :1904.09223 [cs].

WANG A., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. (2018). GLUE : A multi-task benchmark and analysis platform for natural language understanding. In T. LINZEN, G. CHRUPAŁA & A. ALISHAHI, Éds., *Proceedings of the 2018 EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, p. 353–355, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5446.

WETTIG A., GAO T., ZHONG Z. & CHEN D. (2023). Should you mask 15% in masked language modeling ? In A. VLACHOS & I. AUGENSTEIN, Éds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 2985–3000, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.217.

WILF A., AKTER S., MATHUR L., LIANG P., MATHEW S., SHOU M., NYBERG E. & MORENCY L.-P. (2023). Difference-masking : Choosing what to mask in continued pretraining. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 13222–13234, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.881.

WU T., CACCIA M., LI Z., LI Y.-F., QI G. & HAFFARI G. (2021). Pretrained language model in continual learning : A comparative study. In *International Conference on Learning Representations*.

YANG D., ZHANG Z. & ZHAO H. (2023). Learning Better Masking for Better Language Model Pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 7255–7267, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.400.

ZHANG S., ROLLER S., GOYAL N., ARTETXE M., CHEN M., CHEN S., DEWAN C., DIAB M., LI X., LIN X. V., MIHAYLOV T., OTT M., SHLEIFER S., SHUSTER K., SIMIG D., KOURA P. S., SRIDHAR A., WANG T. & ZETTLEMOYER L. (2022). Opt : Open pre-trained transformer language models.

A Paramètres de pré-entraînement continu

Avant l'entraînement, les exemples ont été mélangés aléatoirement à trois reprises en utilisant la même graine. Nous entraînons chaque modèle en au moyen de la bibliothèque python `transformers`.

L'entraînement est réalisé sur 10 époques, représentant 4453 étapes pour BERT et 4396 étapes pour LegalBERT, les méthodes de segmentation en sous-mots de chaque modèle étant différentes.

Au total, nous estimons le temps de calcul total à $\simeq$ 4,100 h, à savoir 3,200 h d'entraînement, 380h pour l'évaluation des modèles et 520h pour le développement.

B Analyse des mots les plus masqués

Le score de TF×IDF a été calculé à l'aide du package python `scikit-learn`.

Pour mieux comprendre les différences de mots sélectionnés en fonction des deux scores d'importance, nous présentons dans la Table 3 les 50 mots les plus masqués pour 10 % du corpus d'entraînement.

TF×IDF
applicant, court, 2007, extradition, prosecutor, meshchanskiy, russian, dzhurayev, moscow, uzbekistan, tashkent, district, custody, government, convention, article, office, decision, §, detention, russia, ccp, 4, preventive, v, minsk, federation, 2, application, uzbek, proceedings, 1, 5, criminal, procedure, january, 38124, case, 29, merits, may, dismissed, law, rakhimovskiy, 466, request, decided, sobir, arrest, provisions
MetaDis
general, application, decision, january, september, decided, august, 4, 28, 9, 3, issued, request, rules, dismissed, 23, 29, indicated, basis, ordered, european, apply, be, 24, 17, date, 5, 30, held, final, december, 26, 6, 11, mentioned, applied, specified, 12, february, placed, 2, whether, remain, first, to, deliberated, represented, constitute, case, article

TABLE 3 – 50 mots les plus masqués avec les pondérations TF×IDF et `Metadis` ordonnés par fréquence

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Améliorer la traduction au niveau du document grâce au sur-échantillonnage négatif et au masquage ciblé

Gaëtan Caillaut[1] Mariam Nakhlé[1,2] Jingshu Liu[1] Raheel Qader[1]
(1) Lingua Custodia, Paris, France (2) Université Grenoble Alpes, CNRS, Grenoble INP, LIG, France
firstname.lastname@linguacustodia.com,
firstname.lastname@univ-grenoble-alpes.fr

RÉSUMÉ

Ces travaux visent à améliorer les capacités des systèmes de traduction automatique à tenir compte du contexte dans lequel se trouve la phrase source, et donc, ultimement, à améliorer les performances globales des systèmes de traduction automatique. L'approche que nous proposons repose uniquement sur les données et la manière dont elles sont fournies au modèle durant l'entraînement et est complètement agnostique de l'architecture du modèle. Nous montrons que les performances des modèles de traduction, sur la paire en-fr, peuvent être améliorées simplement en fournissant des données plus pertinentes vis-à-vis de la tâche cible, et ce sans modifier ni complexifier les architectures existantes, en particulier l'architecture Transformer couramment utilisée par les systèmes de TAL modernes. Pour ce faire, nous présentons deux stratégies d'augmentation de données (sur-échantillonnage négatif et masquage ciblé) conçues pour inciter le modèle à s'appuyer sur le contexte. Nous montrons, au travers de métriques appropriées, que ces méthodes permettent d'améliorer les performances des systèmes de traduction sans pour autant modifier ni l'architecture du modèle, ni le processus d'entraînement.

ABSTRACT

Improve Context-Aware Machine Translation with Negative Sampling and Focused Masking

This work aims at enhancing the context awareness of machine translation models without requiring modification on their architecture. Instead, we took a data-driven approach and explore different data augmentation strategies. We show that performance, on the en-fr pair, can be improved solely by improving the « relevance » of the train data according to the target task, instead of refining and/or complicating the transformer architecture, commonly used by modern machine translation systems. Hence, we propose two simple data augmentation strategies (Negative Sampling and Focused Masking) crafted in order to encourage the model to look at the context. We show through the use of appropriate test suite, as well as traditional BLEU, that these data augmentation strategies improve context-level machine translation performances without requiring change in the model architecture nor the training pipeline.

MOTS-CLÉS : traduction au niveau du document, traduction automatique, transformer.

KEYWORDS: context-level machine translation, transformer.

1 Introduction

Les moteurs de traduction traditionnels traduisent les phrases d'un même document indépendamment les unes des autres. Il est de notoriété publique que les prédictions effectuées par systèmes modernes

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 295–306.

de traduction automatique neuronale (Sutskever *et al.*, 2014; Bahdanau *et al.*, 2014; Luong *et al.*, 2015; Vaswani *et al.*, 2017) sont généralement très satisfaisantes. Cependant, ces systèmes bénéficieraient largement à davantage prendre en considération le contexte du texte en cours de traduction.

Nous cherchons à faire évoluer les systèmes de traduction centrés sur des phrases individuelles vers des systèmes en mesure de considérer un contexte plus large. En effet, la plupart des travaux autour de la traduction automatique se concentrent généralement sur le problème de la traduction de phrases individuelles (Bahdanau *et al.*, 2014). Toutefois, il semble communément admis que les informations contextuelles, souvent présentes dans les phrases précédentes, peuvent s'avérer décisives pour comprendre pleinement le sens d'une phrase, en particulier lorsqu'elle contient des références implicites comme des ellipses ou des deixis (Voita *et al.*, 2019). Plus généralement, il est courant qu'une phrase référence une précédente, notamment via l'usage de pronoms qui devront être traduits en genre et en nombre. Considérons un document contenant les deux phrases suivantes en anglais : « the girls are getting ready for their exam. they are a bit anxious. ». Si l'on traduit ces deux phrases vers le français de manière classique, c'est à dire indépendamment les unes des autres, on obtiendrait la traduction suivante : « les filles se préparent pour leur examen. Ils sont un peu anxieux. ». En revanche, la prise en compte de la première phrase lors de la traduction de la deuxième, permettrait de lever toutes ambiguïtés et d'obtenir une traduction correcte : « les filles se préparent pour leur examen. Elles sont un peu anxieuses. ». Bien que simple, cet exemple montre qu'il est véritablement crucial de tenir compte du contexte si l'on veut être en mesure de traduire un texte convenablement.

Par ailleurs, tenir compte du contexte ne vise pas uniquement à améliorer la qualité des traductions générées par le modèle, mais peut aussi permettre de réduire certains biais, tels que les biais de genre (Stanczak & Augenstein, 2021), puisque le modèle pourra s'appuyer sur le contexte afin de désambiguïser correctement les éléments critiques.

Plusieurs approches permettant d'incorporer ces informations contextuelles au sein de la représentation de la phrase source ont déjà été proposées. Toutefois, plusieurs travaux ont montré que, curieusement, cela n'améliore pas les systèmes de traduction automatique. Pire encore, ces informations contextuelles semblent être considérées comme du bruit (Kim *et al.*, 2019), ce qui, certes, semble améliorer la robustesse des modèles lorsqu'il s'agit de traduire des phrases individuelles, mais n'améliore pas nécessairement la qualité de la traduction lorsqu'il s'agit de traduire un paragraphe complet. Ces approches consistent généralement en une extension de l'architecture Transformer (Vaswani *et al.*, 2017) avec un encodeur dédié aux phrases contextuelles (Zheng *et al.*, 2020; Maruf *et al.*, 2019).

Les travaux présentés ici s'inspirent de ceux de Lupo *et al.* (2022), mais nous avons choisi d'explorer une piste différente. Les auteurs introduisent une évolution de l'architecture Transformer, censée être plus à même de tenir compte des informations contextuelles. Nous pensons que l'architecture originale est tout à fait capable de prendre le contexte en considération, et que le problème provient principalement de la manière dont les données sont fournies au modèle durant l'entraînement. Ainsi, nous proposons dans ces travaux de repenser le jeu de données et l'objectif ciblé durant la phase d'entraînement afin de pousser le modèle à utiliser de manière appropriée la phrase à traduire et (surtout) son contexte. À cet effet, nous proposons les deux stratégies **sur-échantillonnage négatif** (*negative sampling*) et **masquage ciblé** (*focused masking*). La première consiste à fournir au modèle des traductions erronées et à l'entraîner à ne pas les produire. La seconde consiste, durant l'entraînement, à masquer certains mots fortement dépendants du contexte afin de forcer le modèle à extraire cette information du contexte. Nous évaluons nos stratégies à l'aide du traditionnel score BLEU, mais aussi à l'aide de ContraPro (Lopes *et al.*, 2020) et GenPro (Post & Junczys-Dowmunt,

2023), deux méthodes centrées sur l'évaluation de la traduction des pronoms, et donc plus adaptées à notre objectif. Nous montrons que nos stratégies permettent d'améliorer la qualité des traductions lorsque le contexte est présent, tout en ne la dégradant pas lorsque celui-ci est absent.

2 Travaux connexes

La prise en compte du contexte est crucial afin de construire un système de traduction automatique fiable et robuste. Pourtant, il est encore difficile d'entraîner efficacement un modèle sur ce problème particulier. L'une des raison étant la faible disponibilité des ressources contextuelles : la plupart des jeux de données pour la traduction no sont disponibles que sous la forme de phrases parallèles, dépourvues de leurs contextes d'origine (Schwenk *et al.*, 2019a,b; Koehn, 2005). De plus, les systèmes actuels sont limités par la taille des phrases qu'ils sont capable de traiter, bien que ce problème est en passe d'être en partie résolu avec le développement des LLM de dernières génération, permettant de traiter plusieurs milliers de mots (Jiang *et al.*, 2024; AI@Meta, 2024). Toutefois, les statistiques de la compétition WMT23 (Kocmi *et al.*, 2023) montrent que l'adoption des LLM par la communauté de la traduction automatique est encore très faible.

De nombreux travaux dans ce domaine se concentrent sur le problème de la traduction des pronoms (Guillou, 2012; Le Nagard & Koehn, 2010), ces derniers étant très dépendants du contexte par nature. De plus, certaines phrases peuvent être intraduisibles lorsqu'un pronom se s'accorde pas en genre et/ou en nombre dans la langue source, mais s'accorde dans la langue cible. C'est typiquement le cas avec le pronom anglais « they », pouvant se traduire par « ils » ou « elles » en français. Ainsi, Hardmeier & Guillou (2018) montrent que les modèles reposant sur l'architecture Transformer ont tendance à traduire correctement les pronoms non-anaphoriques et ceux coréférents avec un élément de la même phrase. En revanche, ces modèles ne parviennent pas à traduire convenablement les pronoms faisant références à des éléments présents dans les phrases précédentes.

Lupo *et al.* (2022) estiment qu'il est difficile, en traduction automatique, de prendre en compte le contexte, car les informations pertinentes sont très éparses : généralement, seuls un nombre limité de tokens ont un impact sur la phrase courante. Par conséquent, les modèles ont tendance à ignorer le contexte, puisqu'il ne contient finalement que peu de signaux utiles à la traduction. Pourtant, ces signaux sont d'une importance capitale afin de résoudre certaines ambiguïtés linguistiques, comme les ellipses ou les deixis (Voita *et al.*, 2019). Ils proposent de réduire ce problème à l'aide d'un entraînement en trois étapes. La première étape consiste à entraîner un modèle de traduction standard, sans données contextuelles. La seconde étape, appelée `d&r` (*divide and rule*), consiste à réitérer l'entraînement, mais en scindant les phrases en K morceaux. Les $K - 1$ premiers morceaux sont alors utilisés en tant que contexte et le modèle est entraîné à traduire le dernier morceau. De cette manière, le modèle est incité à porter plus d'attention sur le contexte, puisque la dernière portion de phrase a de fortes chances d'être difficile à traduire sans avoir connaissance du début de la phrase. Une dernière étape d'entraînement est requise afin de recadrer le modèle, puisqu'on attend de celui-ci qu'il traduise des phrases complètes, et non pas des fragments de phrases.

Les récentes avancées en Traitement Automatique de la Langue (TAL) ont montré que les grands modèles de langage (LLM) avec architecture décodeur sont en mesure de répondre efficacement à ce problème, puisqu'ils sont capables de traiter de très longues séquences (plusieurs milliers de tokens). Ces grands modèles se sont également montrés surprenant performants sur des tâches de traduction, sans pour autant avoir été entraînés spécifiquement pour (Chowdhery *et al.*, 2022; Scao *et al.*, 2022;

Touvron *et al.*, 2023). Toutefois, la flexibilité offerte par ces modèles requiert d'énormes capacités de calculs, qui ne sont pas nécessairement justifiées si le but visé ne concerne que la traduction. En outre, Raffel *et al.* (2020) ont montré que l'architecture décodeur semble moins adaptée à la traduction automatique que l'architecture encodeur-décodeur. C'est pourquoi nous nous concentrerons sur des modèles encodeur-décodeur de tailles raisonnables dans la suite de ce document.

Les améliorations apportées par la bonne prise en compte du contexte par les systèmes de traduction automatique se révèlent assez difficile à observer à l'aide des métriques traditionnelles, typiquement BLEU (Papineni *et al.*, 2002), car les quelques tokens dépendant du contexte sont noyés par la masse de tokens n'en dépendant pas. C'est pourquoi Lopes *et al.* (2020) introduisent `ContraPro`, un jeu de données dédié à l'évaluation de modèles contextuels. Ce jeu de données est constitué de phrases provenant de OpenSubtitles2018 (Lison *et al.*, 2018) dans lesquelles le genre de certains pronoms ont été remplacés par le genre opposé. Par exemple, certains « il » sont changés en « elle ». Les phrases présentes dans le jeu de données sont sélectionnées afin que le contexte puisse permettre de sélectionner le bon pronom. Les modèles sont alors évalués selon les probabilités qu'ils attribuent aux phrases originales et à leurs versions altérées : la phrase originale doit être considérée comme plus probable que l'autre.

Post & Junczys-Dowmunt (2023) pensent que les modèles doivent être évalués en fonction des phrases qu'ils génèrent réellement, pas seulement selon les probabilités qu'ils donnent à des phrases prédéfinies. En effet, si l'on considère une traduction A et sa version altérée B, `ContraPro` considère le modèle comme étant juste si et seulement si la probabilité $P(A)$ est supérieure à $P(B)$. Or, $P(A)$ peut tout à fait être tellement faible qu'en pratique cette phrase ne puisse pas être générée par le modèle. Dans ce cas, doit-on considérer le modèle comme juste ? Post & Junczys-Dowmunt (2023) cherchent à limiter ce biais en proposant `GenPro`, une méthode d'évaluation basée sur `ContraPro` valorisant un modèle quand il génère effectivement le pronom attendu.

3 Méthode

Les approches existantes cherchent généralement à introduire l'information contextuelle ajoutant un module dédié à l'architecture transformer, couramment utilisée de nos jours pour répondre aux problématiques de TAL. Or, Li *et al.* (2020) ont montré que ces modifications n'aident pas réellement le modèle à mieux prendre en compte le contexte. Comme montré par Kim *et al.* (2019), les modèles tendent à considérer le contexte comme du bruit, et apprennent à ne pas en tenir compte. Ce comportement n'est pas déraisonnable, puisque l'immense majorité des informations nécessaires se trouvent effectivement dans la phrase à traduire, et non dans le contexte passé. Par conséquent, le modèle est naturellement incité, lors du processus d'entraînement, à se concentrer sur la phrase en cours de traduction uniquement, et à ignorer le reste. Cela permet au modèle de converger rapidement vers une solution acceptable. Toutefois, nous supposons qu'une fois ce niveau atteint, il est difficile, pour le modèle, de *faire marche arrière* afin de sortir de cet état où le contexte est ignoré.

Nous pensons que l'architecture Transformer originale est naturellement capable de sélectionner et d'utiliser les bonnes informations contextuelles. Le mécanisme d'attention est suffisamment puissant pour extraire les signaux pertinents pour chaque token. Selon nous, si le modèle traite les phrases contextuelles comme du bruit, c'est parce qu'il a été entraîné de façon sous-optimale, et que le modèle doit être davantage incité à regarder du côté du contexte, au lieu d'apprendre à l'ignorer.

Lupo *et al.* (2022) proposent une méthode dont afin de forcer le modèle à extraire de l'information du côté du contexte. Toutefois, leur approche en trois étapes nous semble contraignante et perfectible. En effet, le principe de leur méthode est de fournir au modèle des phrases dont la structure est *cassée* de manière à ce qu'elles ne puissent pas être traduites sans leur contexte. Le modèle est par la suite ré-entraîné sur des phrases non altérées, tout en figeant les paramètres du modèle chargés d'extraire l'information contextuelle. Le modèle résultant de ce processus est donc censé extraire de l'information contextuelle, mais il est nécessairement sous-optimal puisqu'il n'a été entraîné que sur des contextes synthétiques et incomplets.

Nous nous inspirons de ces travaux et proposons une approche en deux étapes consistant à entraîner le modèle sur un large jeu de données de phrases parallèles, puis à adapter ce modèle sur un jeu de données de taille réduite que l'on augmente avec des données synthétiques. Nous considérons le même jeu de données utilisé par Lupo *et al.* (2022) , c'est-à-dire WMT14 [1] pour entraîner le modèle de base et IWSLT17 [2] pour affiner le modèle sur des données contextuelles. Dans ces travaux, nous n'étudions que le cas de la traduction de l'anglais au français (EN-FR). Le contexte et la phrase à traduire sont encodés de la manière suivante : `<s> <ctx> CONTEXT TOKENS </ctx> SOURCE TOKENS </s>`.

Nous pensons que le comportement du modèle vis-à-vis du contexte doit être guidé par les données vues durant la phase d'entraînement, c'est pourquoi nous explorons différentes manières de pousser le modèle à extraire de l'information du contexte, lorsque cela s'avère nécessaire.

3.1 Sur-échantillonnage négatif

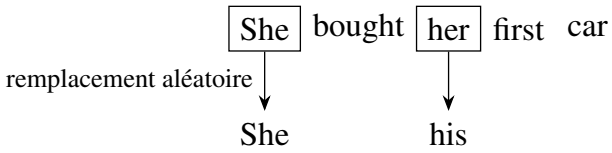


FIGURE 1 – Augmentation du jeu de données avec des échantillons négatifs. La phrase source est scannée pour détecter des mots genrés, qui sont remplacé selon une certaine probabilité.

L'objectif du sur-échantillonnage négatif (*negative sampling*) est de fournir au modèle davantage d'exemples négatifs afin de l'entraîner à ne pas les reproduire. Pour chaque phrase cible, nous générons une version altérée dans laquelle les mots *critiques* (pronoms et déterminants) sont remplacés par leurs homologues du genre opposé, comme illustré en Figure 1. Nous avons sélectionné les mots à remplacer en effectuant une simple recherche basée sur les caractères au lieu de reposer sur des méthodes plus raffinées, par exemple en s'appuyant sur l'analyse morphosyntaxique des mots. Ensuite, seul un sous-ensemble aléatoire des mots ainsi sélectionnés sont remplacés. La liste des déterminants et pronoms que nous considérons est la suivante : il/elle, ils/elles, le/la, un/une, mon/ma, ton/ta, son/sa, ce/cette, tous/toutes, quel/quelle et lequel/laquelle.

Le modèle est entraîné de façon à minimiser la fonction de perte associée au sur-échantillonnage (*negative sampling loss*) $\mathcal{L}_{ns}$, correspondant à la moyenne des probabilités des tokens remplacés, telles que prédites par le modèle (les *logits*). Puisque les tokens remplacés sont incorrects, leurs

1. http://www.statmt.org/wmt14/translation-task.html
2. https://sites.google.com/site/iwsltevaluation2017/data-provided

probabilités doivent être faibles, nous cherchons donc à minimiser $\mathcal{L}_{ns}$. La fonction objectif finale est donnée par $\mathcal{L} = \mathcal{L}_{translation} + \mathcal{L}_{ns}$ où $\mathcal{L}_{translation}$ est l'entropie-croisée, couramment utilisée pour entraîner les modèles de traduction.

3.2 Masquage ciblé

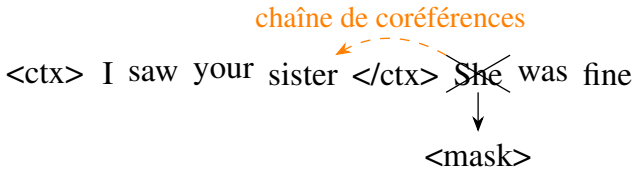


FIGURE 2 – Masquage des mots coréférents. Les éléments de la chaine de coréférence [she, sister] présents dans la source sont masqués selon une certaine probabilité.

L'idée du masquage ciblé est relativement similaire à celle du sur-échantillonnage négatif, mais au lieu d'altérer la cible, nous modifions la phrase de façon à masquer certains mots, censés être très dépendants du contexte. Puisque l'objectif est d'inciter le modèle à extraire de l'information du contexte, les tokens du contexte ne sont jamais masqués, seuls les tokens de la phrase à traduire le sont. Nous avons exploré deux stratégies, l'une consiste à masquer les mots coréférents comme montré en Figure 2, l'autre à masquer les mots apparaissant à la fois dans le contexte et la source. Nous avons utilisé `coreferee` [3] et `spaCy` [4] pour extraire les chaînes de coréférence. Par exemple, dans la phrase « J'ai vu ta sœur. Elle allait bien. », les mots « sœur » et « elle » sont coréférents. Nous supposons que masquer le pronom « elle » forcera le modèle à analyser le contexte pour en déduire le bon pronom. De même que pour la stratégie précédente, seulement une portion aléatoire d'éléments des chaînes de coréférence sont masqués, et seulement s'ils apparaissent dans la source afin de ne pas décourager le modèle à puiser de l'information dans le contexte.

3.3 Divide and rule

L'objectif de la stratégie divide-and-rule proposée par Lupo *et al.* (2022) est de forcer le modèle à observer le contexte en divisant la phrase initiale en morceaux de tailles similaires, et de réutiliser les premiers morceaux en guise de contexte. De cette manière, le modèle est entraîné sur des phrases complètement déstructurées, de sorte qu'il soit nécessaire de prendre en compte le contexte pour traduire la phrase. Dans les travaux originaux, les auteurs entraînent un encodeur dédié au contexte en plus de celui dédié à la phrase source à traduire, puis fusionnent les représentations de ces deux encodeurs avant de générer la traduction par le décodeur.

Dans le cadre de nos travaux, nous pensons que l'encodeur dédié au contexte est superflu, mais la stratégie divide-and-rule nous semble toutefois pertinente, notamment car elle permet de créer de l'information contextuelle à partir de phrases isolées. C'est pourquoi nous avons généré des données synthétiques en divisant chaque phrase source en trois morceaux x, y et z de tailles équivalentes. Ainsi, chaque phrase nous permet de générer les trois paires contexte/source suivantes :

3. https://github.com/richardpaulhudson/coreferee
4. https://spacy.io/

— <s> <ctx> </ctx> x y z </s>
— <s> <ctx> x </ctx> y z </s>
— <s> <ctx> x y </ctx> z </s>

Les phrases cibles sont générées de la même manière, ce qui s'avère être l'un des défauts majeurs de cette stratégie, puisque les mots présents dans la source et la cible n'apparaissent pas nécessairement dans le même ordre.

4 Expérimentations

Plusieurs modèles ont été entraînés afin de mesurer l'impact des différentes stratégies présentées dans ces travaux : sur-échantillonnage négatif, masquage ciblé et *divide and rule*. Les modèles ont été entraînés dans un premier temps sur la direction EN-FR du corpus WMT14 pendant une *epoch*, puis affinés sur la direction EN-FR du corpus IWSLT17 pour 15 *epochs*. Nos modèles reprennent l'architecture encodeur-décodeur standard, avec 6 couches d'encodeurs et 6 couches de décodeurs, 8 têtes d'attentions et la dimension des embeddings est fixée à 512. Les modèles sont entraînés avec l'optimiseur AdamW, un taux d'apprentissage de $5e^{-5}$ et la taille du batch est fixée à 32 phrases.

Les scores BLEU sont calculés sur l'ensemble de test du corpus IWSLT17 à l'aide de SacreBLEU (Post, 2018). Nous calculons également les scores ContraPro (Lopes *et al.*, 2020) et GenPro (Post & Junczys-Dowmunt, 2023), puisqu'ils sont censés mieux capturer la capacité d'un modèle à tenir compte des informations contextuelles. Ces deux dernières métriques mesurent essentiellement la capacité du modèle à traduire correctement les pronoms, qui sont des éléments très dépendants du contexte.

Chaque modèle a été entraîné avec et sans contexte afin d'observer si l'information portée par le contexte permet réellement d'améliorer la qualité des traductions. Dans cette expérience, le contexte correspond à la phrase précédant la phrase source. Les scores de tous les modèles évalués sont donnés en Tableaux 1 et 2. Les modèles dont le nom portent la mention « coref » et « samewords » ont été entraînés avec, respectivement, les stratégies consistant à masquer les mots coréférents et les mots partagés entre le contexte et la source. Pour chaque stratégie, deux expériences ont été faites avec des probabilités de masquage de 0,5 et 0,8. Le modèle « d&r » a été entraîné à l'aide de l'approche *divide and rule*.

Notre première observation est que les modèles contextuels (auxquels on fournit la phrase précédente) sont globalement plus performants que les autres modèles, comme l'indiquent les scores en Tableau 1, supérieurs à ceux en Tableau 2. Les différences entre chaque scores sont statistiquement significatives ($p < 0.05$), il semble donc qu'il soit bénéfique d'augmenter l'entrée du modèle, la phrase à traduire, avec son contexte. Cela est d'autant plus perceptible sur les scores ContraPro et GenPro, là où les fluctuations sur les scores BLEU sont plus limitées. Cette dernière observation corrobore les conclusions d'autres travaux, notamment ceux ayant motivé la création de ContraPro et GenPro, selon lesquelles BLEU ne serait pas adapté pour l'évaluation de modèles contextuels (Nakhlé, 2023; Jin *et al.*, 2023). De plus, ces expériences montrent qu'il est effectivement nécessaire d'inclure des informations contextuelles afin de traduire correctement les pronoms. Il en va certainement de même pour d'autres éléments fortement dépendants du contexte, mais ContraPro et GenPro ne nous permettent de nous prononcer que sur les pronoms.

Nous remarquons également que les différentes stratégies de masquage, de même que la méthode

	affiné avec SN			*affiné sans SN*		
Modèle	**BLEU**	**ContraPro**	**GenPro**	**BLEU**	**ContraPro**	**GenPro**
Pré-entraîné avec SN						
no masking	39,0	0,84	46,4	39,3	0,83	52,1
d&r	39,3	0,83	43,5	39,9	0,82	51,7
coref50	38,9	0,84	46,3	39,4	0,83	52,2
coref80	38,9	0,84	46,5	39,3	0,83	52,3
samewords50	39,0	0,83	47,2	39,7	0,82	52,3
samewords80	39,0	0,83	47,4	39,7	0,82	52,4
Pré-entraîné sans SN						
no masking	39,0	0,83	43,3	39,8	0,79	46,8
d&r	39,3	0,82	39,7	40,1	0,78	44,5
coref50	39,0	0,83	43,3	39,8	0,79	46,8
coref80	39,0	0,83	43,3	39,7	0,79	46,8
samewords50	39,4	0,82	41,6	39,9	0,78	43,8
samewords80	39,9	0,78	43,8	40,0	0,78	43,8
Lupo et al. *(2022)*						
K1				41,93	0,84	
K1 d&r				41,78	0,79	

TABLE 1 – Performances sur la traduction de phrases accompagnées de leurs contextes. SN signifie « sur-échantillonnage négatif ». Tous les scores sont à maximiser. Les différences entre les modèles affinés avec SN (gauche) et sans (droite) sont significatives ($p < 0.05$). Les différences entre les modèles pré-entraînés avec (haut) et sans SN (bas) ne sont significatives ($p < 0.05$) que pour les scores *Pro.

divide and rule, n'ont finalement que très peu d'impact sur les performances finales. Pire, le sur-échantillonnage négatif semble dégrader légèrement les scores BLEU des modèles contextuels. Cependant, cette stratégie semble tout de même améliorer les scores ContraPro et GenPro. C'est en effet la seule méthode permettant d'améliorer systématiquement le score ContraPro, indiquant que cette stratégie permet bel et bien au modèle d'extraire de meilleures informations contextuelles. Toutefois, cette stratégie semble aussi dégrader les scores GenPro lorsqu'elle est appliquée pour affiner le modèle. Cela signifie que le sur-échantillonnage négatif permet au modèle de mieux identifier les erreurs sur les pronoms, sans pour autant l'inciter à générer les bons. Cela reste tout de même à considérer avec un peu plus de hauteur, puisque GenPro repose sur une méthode heuristique pour aligner les pronoms entre les phrases sources et cibles. Il est possible que de nombreux faux négatifs soient comptabilisés.

Enfin, nous comparons les résultats obtenus par nos modèles avec ceux publiés par Lupo *et al.* (2022). Nos expérimentations sont comparables à celles effectuées sur le modèle intitulé *K1 Low Res*, c'est pourquoi nous ne reprenons que ces résultats en Tableau 1. Les scores BLEU des modèles *K1* sont plus élevés que les nôtres, nous supposons que cela provient de paramétrages plus optimaux, soit au niveau des hyperparamètres des modèles, soit au niveau de l'entraînement (taille du batch, durée de l'entraînement, ...). Nous nous intéressons davantage aux gains apportées par l'approche d&r sur les scores ContraPro, et nous observons que nos modèles parviennent à concurrencer le modèle *K1 d&r*, même en l'absence d'encodeur dédié à la prise en charge du contexte. Cela semble valider notre

	affiné avec SN			*affiné sans SN*		
Modèle	**BLEU**	**ContraPro**	**GenPro**	**BLEU**	**ContraPro**	**GenPro**
Pré-entraîné avec SN						
no masking	37,8	0,78	33,6	38,6	0,78	40,5
d&r	39,3	0,78	31,6	39,7	0,77	40,8
coref50	38,2	0,78	33,6	38,5	0,78	40,4
coref80	38,0	0,78	33,7	38,7	0,78	40,5
samewords50	38,4	0,78	34,6	39,1	0,77	40,6
samewords80	38,5	0,78	34,7	39,0	0,77	40,6
Pré-entraîné sans SN						
no masking	35,2	0,77	30,9	36,0	0,77	39,7
d&r	39,5	0,78	31,1	40,1	0,76	41,4
coref50	35,2	0,77	31,0	35,9	0,77	39,7
coref80	35,2	0,77	30,9	35,9	0,77	39,6
samewords50	35,9	0,78	30,5	36,8	0,76	39,2
samewords80	36,7	0,76	39,2	36,8	0,76	39,2

TABLE 2 – Performances sur la traduction de phrases individuelles. SN signifie « sur-échantillonnage négatif ». Tous les scores sont à maximiser. Les différences entre les modèles affinés avec SN (gauche) et sans (droite) ne sont significatives ($p < 0.05$) que pour GenPro. Les différences entre les modèles pré-entraînés avec (haut) et sans SN (bas) ne sont significatives ($p < 0.05$) que pour les scores BLEU et ContraPro.

hypothèse initiale (il n'est pas nécessaire de complexifier l'architecture du modèle pour améliorer la prise en charge du contexte).

En résumé, nos expériences montrent que la meilleure stratégie semble être d'augmenter le jeu de données de pré-entraînement avec des exemples négatifs (sur-échantillonnage négatif), puisque les modèles entraînés de cette manière tendent à trouver le meilleur équilibre entre la qualité de la traduction (BLEU) et prise en compte du contexte (ContraPro et GenPro). Les autres stratégies évaluées ne semblent pas apporter de réels gains.

5 Conclusion

Nous proposons, des méthodes permettant d'introduire des éléments d'information contextuelles au sein du processus de traduction. Cette étude se concentre sur la prise en compte des informations présentes dans la phrase précédente, mais nous envisageons d'étendre le contexte à davantage de phrases. Nous proposons et évaluons différentes stratégies d'entraînement et d'augmentation de données destinées à améliorer la prise en compte du contexte par modèles de traduction automatique. En effet, comme le montrent nos expériences, inclure le contexte, même s'il ne s'agit que de la phrase précédente, permet d'augmenter les performances du système de traduction. L'une de nos stratégies consiste à augmenter le jeu de données initiales afin d'ajouter des exemples négatifs. Les résultats montrent que cette stratégie permet effectivement d'améliorer la prise en charge du contexte par le modèle, ce qui se matérialise par des scores ContraPro et GenPro plus élevés, deux métriques conçues

pour évaluer les modèles contextuels. Toutefois, cette stratégie semble dégrader la qualité générale de la traduction, telle que reportée par le score BLEU. Nous montrons qu'il est possible d'éviter ce phénomène en appliquant le sur-échantillonnage négatif lors du pré-entraînement uniquement. Notre approche se limite à introduire un signal négatif sur le genre et le nombre de certains pronoms et déterminants. L'une des pistes à explorer serait d'envisager d'autres méthodes de corruptions, par exemple remplacer un mot par son antonyme, ou encore modifier le temps de certains verbes.

La seconde stratégie que nous proposons consister à sélectionner des termes dépendants du contexte et d'en masquer une proportion aléatoire. Nous avons exploré deux pistes : masquer les mots apparaissant à la fois dans le contexte et la source et masquer certains éléments des chaînes de coréférences. Nos expériences montrent que cette approche n'apporte aucune plus-value. Cependant, cela ne suffit pas pour rejeter complètement cette stratégie, puisqu'il est possible qu'elle porte ses fruits avec des choix plus avisés de mots à masquer.

Références

AI@META (2024). Llama 3 model card.

BAHDANAU D., CHO K. & BENGIO Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*.

CHOWDHERY A., NARANG S., DEVLIN J., BOSMA M., MISHRA G., ROBERTS A., BARHAM P., CHUNG H. W., SUTTON C., GEHRMANN S. *et al.* (2022). Palm : Scaling language modeling with pathways. *arXiv preprint arXiv :2204.02311*.

GUILLOU L. (2012). Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 1–10.

HARDMEIER C. & GUILLOU L. (2018). Pronoun translation in english-french machine translation : An analysis of error types. *arXiv preprint arXiv :1808.10196*.

JIANG A. Q., SABLAYROLLES A., ROUX A., MENSCH A., SAVARY B., BAMFORD C., CHAPLOT D. S., CASAS D. D. L., HANNA E. B., BRESSAND F. *et al.* (2024). Mixtral of experts. *arXiv preprint arXiv :2401.04088*.

JIN L., HE J., MAY J. & MA X. (2023). Challenges in context-aware neural machine translation. *arXiv preprint arXiv :2305.13751*.

KIM Y., TRAN D. T. & NEY H. (2019). When and why is document-level context useful in neural machine translation ? *arXiv preprint arXiv :1910.00294*.

KOCMI T., AVRAMIDIS E., BAWDEN R., BOJAR O., DVORKOVICH A., FEDERMANN C., FISHEL M., FREITAG M., GOWDA T., GRUNDKIEWICZ R. *et al.* (2023). Findings of the 2023 conference on machine translation (wmt23) : Llms are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, p. 1–42.

KOEHN P. (2005). Europarl : A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x : papers*, p. 79–86.

LE NAGARD R. & KOEHN P. (2010). Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, p. 252–261 : Association for Computational Linguistics.

LI B., LIU H., WANG Z., JIANG Y., XIAO T., ZHU J., LIU T. & LI C. (2020). Does multi-encoder help ? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, p. 3512–3518, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.322.

LISON P., TIEDEMANN J. & KOUYLEKOV M. (2018). Opensubtitles2018 : Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)* : European Language Resources Association (ELRA).

LOPES A. V., FARAJIAN M. A., BAWDEN R., ZHANG M. & MARTINS A. F. (2020). Document-level neural mt : A systematic comparison. In *22nd Annual Conference of the European Association for Machine Translation*, p. 225–234.

LUONG M.-T., PHAM H. & MANNING C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv :1508.04025*.

LUPO L., DINARELLI M. & BESACIER L. (2022). Divide and rule : Effective pre-training for context-aware multi-encoder translation models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 4557–4572.

MARUF S., MARTINS A. F. & HAFFARI G. (2019). Selective attention for context-aware neural machine translation. *arXiv preprint arXiv :1903.08788*.

NAKHLÉ M. (2023). L'évaluation de la traduction automatique du caractère au document : un état de l'art. In *Actes de CORIA-TALN 2023. Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL)*, p. 143–159.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, p. 311–318.

POST M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 186–191, Belgium, Brussels : Association for Computational Linguistics.

POST M. & JUNCZYS-DOWMUNT M. (2023). Escaping the sentence-level paradigm in machine translation. *arXiv preprint arXiv :2304.12959*.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, **21**(1), 5485–5551.

SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIĆ S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F., GALLÉ M. *et al.* (2022). Bloom : A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv :2211.05100*.

SCHWENK H., CHAUDHARY V., SUN S., GONG H. & GUZMÁN F. (2019a). Wikimatrix : Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv :1907.05791*.

SCHWENK H., WENZEK G., EDUNOV S., GRAVE E. & JOULIN A. (2019b). Ccmatrix : Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv :1911.04944*.

STANCZAK K. & AUGENSTEIN I. (2021). A survey on gender bias in natural language processing. *arXiv preprint arXiv :2112.14168*.

SUTSKEVER I., VINYALS O. & LE Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, **27**.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S. *et al.* (2023). Llama 2 : Open foundation and fine-tuned chat models. *arXiv preprint arXiv :2307.09288*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

VOITA E., SENNRICH R. & TITOV I. (2019). When a good translation is wrong in context : Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. *arXiv preprint arXiv :1905.05979*.

ZHENG Z., YUE X., HUANG S., CHEN J. & BIRCH A. (2020). Towards making the most of context in neural machine translation. *arXiv preprint arXiv :2002.07982*.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Améliorer les modèles de langue pour l'analyse des émotions : perspectives venant des sciences cognitives

Constant Bonard[1] Gustave Cortal[2, 3]
(1) Université de Berne, Département de Philosophie, Hochschulstrasse 4, 3012 Berne, Suisse
(2) Université Paris-Saclay, ENS Paris-Saclay, CNRS, LMF, 91190, Gif-sur-Yvette, France
(3) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
constant.bonard@gmail.com, gcortal@ens-paris-saclay.fr

RÉSUMÉ

Nous proposons d'exploiter les recherches en sciences cognitives sur les émotions et la communication pour améliorer les modèles de langue pour l'analyse des émotions. Tout d'abord, nous présentons les principales théories des émotions en psychologie et en sciences cognitives. Puis, nous présentons les principales méthodes d'annotation des émotions en traitement automatique des langues et leurs liens avec les théories psychologiques. Nous présentons aussi les deux principaux types d'analyses de la communication des émotions en pragmatique cognitive. Enfin, en s'appuyant sur les recherches en sciences cognitives présentées, nous proposons des pistes pour améliorer les modèles de langue pour l'analyse des émotions. Nous suggérons que ces recherches ouvrent la voie à la construction de nouveaux schémas d'annotation et d'un possible *benchmark* pour la compréhension émotionnelle, prenant en compte différentes facettes de l'émotion et de la communication chez l'humain.

ABSTRACT

Improving Language Models for Emotion Analysis : Insights from Cognitive Science.

We propose leveraging cognitive science research on emotions and communication to improve language models for emotion analysis. First, we present the main emotion theories in psychology and cognitive science. Then, we introduce the main methods of emotion annotation in natural language processing and their connections to psychological theories. We also present the two main types of analyses of emotional communication in cognitive pragmatics. Finally, based on the cognitive science research presented, we propose directions for improving language models for emotion analysis. We suggest that these research efforts pave the way for constructing new annotation schemes and a possible benchmark for emotional understanding, considering different facets of human emotion and communication.

MOTS-CLÉS : Analyse des émotions, modèle de langue, sciences cognitives, sciences affectives, pragmatique.

KEYWORDS: Emotion analysis, language model, cognitive science, affective sciences, pragmatics.

1 Introduction

L'analyse des émotions dans le traitement automatique des langues vise à développer des modèles computationnels capables de discerner les émotions humaines dans le texte. Récemment, les modèles de langue ont largement été utilisés pour résoudre diverses tâches en traitement automatique des

[*] Les auteurs ont contribué à parts égales et apparaissent par ordre alphabétique.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 307–322.

langues, dont l'analyse des émotions (Devlin *et al.*, 2019; Brown *et al.*, 2020). Ce domaine de recherche fait face à plusieurs limites. Tout d'abord, les différentes façons de conceptualiser les émotions amènent à différents schémas d'annotation et jeux de données (Klinger, 2023). En conséquence, la capacité de généralisation des modèles est limitée et il est souvent impossible de comparer les études. Pour remédier à ces limites, il a été proposé d'unifier certains schémas d'annotation en se basant sur la proximité sémantique des catégories d'émotions (Bostan & Klinger, 2018), de trouver automatiquement les catégories d'émotions d'intérêts à partir des données (De Bruyne *et al.*, 2020) ou d'obtenir des plongements émotionnels indépendants des schémas d'annotation (Buechel *et al.*, 2021). En s'inspirant de débats récents en sciences cognitives (Scherer, 2022), nous pensons qu'il serait possible de construire un schéma d'annotation unifiant différentes perspectives sur l'émotion.

De plus, les *benchmarks* existants évaluent certains aspects de la compréhension émotionnelle, mais sans prendre en compte toute sa complexité (Campagnano *et al.*, 2022; Zhang *et al.*, 2023; Paech, 2024). Par exemple, Paech (2024) propose d'évaluer la compréhension émotionnelle des modèles de langue à travers la prédiction de l'intensité des émotions dans des scènes de conflits. Ce type d'évaluation est trop limité : les *benchmarks* devraient refléter autant que possible la richesse de la compréhension émotionnelle chez les humains, une richesse documentée ces dernières décennies dans différentes branches des sciences affectives (Green, 2007; Wharton, 2016; Scarantino, 2017; Barrett *et al.*, 2019; Bonard & Deonna, 2023).

Un autre domaine de recherche connexe se concentre sur la théorie de l'esprit des modèles de langue, c'est-à-dire leur capacité à attribuer correctement des états mentaux aux autres. Cette littérature est prometteuse selon nous en cela qu'elle relie les développements récents des modèles de langue aux théories et aux méthodes empiriques des sciences cognitives (Bonard, 2024, section 5). Notamment, plusieurs tâches et *benchmarks* ont été développés pour mesurer la capacité des modèles de langue à réussir différentes versions de la tâche de la fausse croyance (*False Belief Task*) (Trott *et al.*, 2022; Aru *et al.*, 2023; Gandhi *et al.*, 2023; Holterman & van Deemter, 2023; Kosinski, 2023; Mitchell & Krakauer, 2023; Shapira *et al.*, 2023; Stojnić *et al.*, 2023; Ullman, 2023). Cependant, la théorie de l'esprit et, plus généralement, les capacités de raisonnement social vont au-delà de la capacité à réussir la tâche de la fausse croyance (Apperly & Butterfill, 2009; Langley *et al.*, 2022; Ma *et al.*, 2023). La capacité à interpréter correctement les émotions exprimées ne peut s'y réduire. Le degré de possession de cette compétence émotionnelle par les modèles de langue mérite d'être étudiée en soi.

D'une façon générale, la recherche portant sur les modèles de langue pour l'analyse des émotions bénéficierait d'un apport de la recherche en sciences cognitives. Notamment, nous pensons que cela peut mener à de meilleures manières d'annoter les émotions exprimées dans le texte, mais aussi à une meilleure évaluation de la compréhension émotionnelle des modèles de langue en développant de nouveaux *benchmarks*. Nous présentons un panorama général sur les théories psychologiques des émotions (section 2) et sur les manières d'annoter les émotions dans le traitement automatique des langues (section 3). Puis, en s'inspirant de certaines théories psychologiques et linguistiques (section 4), nous proposons des directions de recherche pour remédier à certaines limites actuelles de l'analyse des émotions (section 5).

Contributions. Pour améliorer l'analyse automatique des émotions, nous proposons d'intégrer différentes théories des sciences cognitives avec le TAL. Nous expliquons pourquoi et comment l'analyse des émotions devrait utiliser des théories en psychologie des émotions – en particulier le cadre intégré – ainsi que des théories en pragmatique cognitive – en particulier l'analyse du détective. Cela conduit à l'élaboration d'un nouveau schéma d'annotation et à une meilleure évaluation des modèles de langue.

2 Les théories des émotions dans les sciences cognitives

Cette section présente les trois principales théories des émotions en psychologie afin de fournir un arrière-plan quant à notre projet de mieux connecter l'analyse des émotions en traitement automatique des langues avec les sciences cognitives.

La théorie des émotions de base. La théorie des émotions de base est certainement la plus influente aujourd'hui. Inspirée par les recherches de Darwin sur les émotions (Darwin, 1872), elle postule un certain nombre d'émotions discrètes et fondamentales qui sont universelles et innées chez les humains en raison de leurs origines évolutives. Les émotions sont comprises comme des « programmes » psycho-physiologiques qui ont été sélectionnés pour aider à surmonter les défis évolutifs récurrents (Cosmides & Tooby, 2000). Une version importante de cette théorie est celle de Paul Ekman (Ekman, 1999), qui a cherché à montrer, comme le prévoyait Darwin, que certaines émotions sont associées aux mêmes expressions faciales à travers toutes les cultures - Ekman a utilisé la liste d'émotions proposées par Darwin (Darwin, 1872) : colère, peur, surprise, dégoût, bonheur et tristesse. Il a notamment mené des études auprès d'individus n'ayant pas été exposés à la culture occidentale, indiquant qu'ils pouvaient correctement identifier les expressions faciales pour ces six émotions sur des photographies (Ekman & Friesen, 1971). Des tentatives ont également été faites pour soutenir la théorie des émotions de base en identifiant des signatures physiologiques et neurologiques des émotions de base (Moors, 2022, p. 129–131).

Il convient de noter qu'Ekman n'a pas précisé le nombre exact d'émotions de base. Outre les six émotions énumérées, les candidats comprennent l'amusement, le mépris, la gêne, la culpabilité, la fierté et la honte (Ekman, 1999). D'autres versions de la théorie des émotions de base proposent différentes listes (Tomkins, 1962; Izard, 1992; Panksepp, 1998; Plutchik, 2001).

Le constructivisme psychologique. Le constructivisme psychologique est l'alternative la plus influente à la théorie des émotions de base aujourd'hui. Il rejette l'idée qu'il existe des émotions discrètes et fondamentales partagées universellement par les humains et postule au contraire que les types d'émotions tels que la colère, la peur et la joie sont construits à travers l'interaction de facteurs biologiques, psychologiques et socioculturels. Parmi les adeptes de la première heure figurent Schachter & Singer (1962). Cette théorie est aujourd'hui principalement associée à James Russell et Lisa Feldman Barrett (Russell, 1980, 2003; Barrett, 2006; Russell, 2009; Barrett, 2017). Les constructivistes psychologiques se concentrent sur les sentiments subjectifs associés aux émotions qui sont interprétés comme un continuum sans barrière catégorique. Les sentiments sont généralement représentés dans un espace bidimensionnel avec un axe de valence (sentiments agréables-désagréables) et un axe d'excitation (sentiments d'activation-désactivation). L'impression qu'il existe des émotions distinctes est considérée comme une construction sociale : différentes formes d'acculturation suscitent différentes façons de conceptualiser ou d'étiqueter nos sentiments corporels en types d'émotions distincts.

La théorie de l'évaluation (*appraisal*). La troisième théorie psychologique majeure de l'émotion est la théorie de l'évaluation (*appraisal*), dont la version empirique a été initiée par Magda Arnold (Arnold, 1960). Elle a été développée pour expliquer l'absence de correspondance bijective, une-à-une, entre les types d'émotions et les types de stimuli émotionnels, c'est-à-dire le fait que le même type de stimuli peut déclencher des émotions différentes et que des types de stimuli différents peuvent déclencher le même type d'émotion. Pour expliquer ce phénomène, des évaluations (*appraisals*) sont postulées comme médiatrices entre les stimuli et les réactions émotionnelles.

Les évaluations en question sont des catégorisations cognitives (inconscientes, rapides et souvent erronées) de la pertinence des stimuli par rapport aux préoccupations de la personne et de la manière dont elle doit y réagir. La théorie de l'évaluation postule que, par exemple, Simon a peur de la souris dans la cuisine parce qu'il l'évalue comme une menace imminente pour sa sécurité, tandis que Sylvie, au contraire, est en colère qu'il y ait une souris dans la cuisine parce qu'elle l'évalue comme un intrus à chasser. Ainsi, chaque type d'émotion peut être analysé par le type d'évaluation qui lui est associée. Ainsi, Lazarus (1991) propose *le danger imminent* pour la peur, *l'offense dégradante* pour la colère, *la perte irrévocable* pour la tristesse et *le progrès vers un but* pour la joie.

Dans les années 1980, des adeptes de cette théorie ont proposé d'analyser les évaluations comme des régions dans un espace multidimensionnel (Moors *et al.*, 2013). Ces dimensions d'évaluation comprennent généralement : (a) la pertinence du stimulus par rapport aux objectifs de l'individu, (b) la capacité de l'individu à faire face à la situation, (c) l'urgence de la réponse nécessaire, (d) la cause de l'événement déclencheur (moi, quelqu'un d'autre, intentionnelle ou non) et (e) la compatibilité avec les normes personnelles de l'individu. Par exemple, la peur est déclenchée par l'évaluation d'un stimulus comme étant (a) fortement contraire aux objectifs, (b) difficile à gérer et (c) nécessitant une réponse urgente.

Cadre intégré pour les théories des émotions. Bien que les trois théories examinées soient habituellement considérées comme rivales, il a été défendu qu'il fallait au contraire les intégrer dans un cadre commun (Scherer & Moors, 2019; Bonard, 2021b; Scherer, 2022). En effet, on peut affirmer que les trois théories diffèrent avant tout vis-à-vis de l'objet de leur enquête, leur axe de recherche et les aspects de l'émotion sur lesquels elles mettent l'accent. La théorie des émotions de base se concentre sur les traits universels hérités de l'évolution et en particulier sur leurs expressions physiologiques et motrices (réactions corporelles). Le constructivisme psychologique se concentre sur les dimensions du ressenti et la façon dont les individus les catégorisent (sentiment subjectif). La théorie de l'évaluation se concentre sur le déclenchement émotionnel (processus d'évaluation) et les tendances à l'action qui en découlent. Nous pensons qu'un cadre intégrant les différents éléments étudiés par ces théories est possible et souhaitable. Ce que nous appelons « le cadre intégré pour les théories des émotions » propose de le faire en postulant que les épisodes émotionnels paradigmatiques sont faits de changements synchronisés et causalement interconnectés dans quatre composantes : (1) processus d'évaluation (*appraisal*), (2) tendances à l'action, (3) changements corporels (expressions motrices et réponses physiologiques), (4) sentiments subjectifs. Pour une discussion d'un tel cadre intégré, voir Scherer (2022).

3 L'analyse des émotions dans le texte

L'émotion est une catégorie. L'analyse des émotions dans le texte s'appuie sur les théories de l'émotion de base pour définir les différentes catégories d'émotion à associer aux unités textuelles (un empan de texte, une phrase ou un document). Par exemple, la phrase « J'adore la philosophie. » pourrait être associée automatiquement à l'émotion discrète *joie*. Plusieurs schémas d'annotation se concentrent sur des sous-ensembles de catégories alors que d'autres considèrent un plus large ensemble, pouvant atteindre plus de 28 catégories différentes (Demszky *et al.*, 2020; Bostan & Klinger, 2018).

L'émotion est une valeur continue ayant un sens affectif. Au lieu de représenter l'émotion par une catégorie, certains schémas d'annotation considèrent que l'émotion est un point dans un espace

multidimensionnel et associent à des unités textuelles des valeurs continues (Buechel & Hahn, 2017). Ces dimensions portent un sens affectif. Deux dimensions sont dominantes dans la littérature et proviennent des théories du constructivisme psychologique qui considèrent qu'une émotion peut être caractérisée par son degré d'*agréabilité* et son degré d'*activation physiologique*. Ainsi, la phrase « Sa voix m'apaise. » pourrait être associée automatiquement à deux valeurs continues : un degré d'*agréabilité* de 4 sur 5 et un degré d'*activation physiologique* de 1 sur 5.

L'émotion est une valeur continue ayant un sens cognitif. Ces dimensions peuvent aussi porter un sens cognitif. Récemment, une nouvelle ligne de recherche propose d'incorporer les théories psychologiques de l'évaluation cognitive dans les modèles d'analyse des émotions (Hofmann *et al.*, 2020; Troiano *et al.*, 2022; Zhan *et al.*, 2023). Depuis cette perspective, les émotions sont causées par des évènements évalués selon plusieurs dimensions cognitives. Par exemple, la phrase « J'ai reçu un cadeau surprise. » pourrait être associée automatiquement à plusieurs valeurs continues : l'évènement est *soudain* (4 sur 5), *contraire aux normes sociales* (0 sur 5) et la personne a le *contrôle* sur l'évènement (0 sur 5).

L'émotion est constituée de rôles sémantiques. Une émotion ne peut se réduire à une catégorie ou des valeurs continues ayant un sens affectif ou cognitif. Pour avoir une meilleure compréhension d'un évènement émotionnel, plusieurs approches associent à des empans de texte des rôles sémantiques comme la *cause*, la *cible*, l'*expérienceur·euse* et l'*indice* de l'émotion (Lee *et al.*, 2010; Kim & Klinger, 2018; Bostan *et al.*, 2020; Oberländer *et al.*, 2020; Campagnano *et al.*, 2022; Wegge *et al.*, 2023). Ainsi, au lieu de considérer l'émotion comme causée par un évènement, l'analyse des rôles sémantiques de l'émotion considère que l'émotion *est* un évènement (Klinger, 2023) qu'il faut reconstituer en répondant à la question : « Qui (*expérienceur·euse*) ressent quoi (*indice*) envers qui (*cible*) et pourquoi (*cause*) ? ». Dans cet exemple, chaque empan de texte peut être associé à un rôle sémantique : « Louise (*expérienceuse*) était en colère (*indice*) contre Paul (*cible*), car il ne l'a pas prévenue (*cause*). »

Première limite : il n'existe pas de schéma d'annotation unifié. Les divergences dans la définition de l'émotion en psychologie mènent vers des divergences dans la manière d'annoter l'émotion dans le texte. Les différentes théories psychologiques des émotions représentent différentes perspectives sur le phénomène émotionnel. Elles sont loin de se contredire et peuvent même tendre à s'unifier (section 2). Nous pensons que c'est aussi le cas pour les schémas d'annotation dans l'analyse des émotions. Dans la section 5, nous donnons des pistes pour la construction d'un schéma d'annotation unifié, inspiré par les débats récents en sciences cognitives (Scherer, 2022).

Seconde limite : la verbalisation de l'émotion est peu considérée. L'analyse des émotions considère rarement le processus de verbalisation de l'émotion. En conséquence, il est difficile d'obtenir des guides d'annotation qui définissent clairement les marqueurs linguistiques à annoter dans le texte. Nous voulons mettre en lumière la théorie linguistique de Raphaël Micheli, qui catégorise un large panel de marqueurs linguistiques en trois modes d'expression de l'émotion (Micheli, 2014) : l'émotion peut être *dite*, *montrée* ou *suggérée* (ou « étayée »). L'émotion peut être exprimée explicitement avec un terme du lexique émotionnel (« Je suis *triste* »), être montrée avec des caractéristiques de l'énoncé comme les interjections et les ponctuations (« *Ah !* C'est super *!* »), ou être suggérée avec la description d'une situation qui généralement, dans un contexte socioculturel donné, mène à une émotion (« *Elle m'a offert un cadeau* »). La majorité des schémas d'annotation se sont concentrés implicitement sur l'émotion *dite*, en occultant les deux autres modes d'expression. Récemment, les schémas d'annotation basés sur les théories de l'évaluation cognitive s'intéressent implicitement à l'émotion *suggérée*. La théorie de Micheli analyse donc les différents types de signes verbaux qui

sont utilisés, chez les humains, pour inférer les émotions exprimées. Par contraste, les théories de la pragmatique cognitive s'intéressent aux mécanismes psychologiques qui sont utilisés pour inférer ce qui est communiqué, dont notamment les émotions exprimées par ces différents types de signes. Dans la prochaine section, nous suggérerons l'hypothèse que les catégories de signes distinguées par Micheli correspondent à différentes sources d'inférences postulées par la pragmatique cognitive.

4 Pragmatique cognitive et communication émotionnelle

Deux analyses de la communication. La pragmatique cognitive est la branche des sciences cognitives qui s'intéresse à la façon dont les personnes utilisent et interprètent les signes dans la communication. Dans cette branche et d'autres disciplines connexes, il est courant de distinguer deux grandes manières d'analyser la communication : l'analyse du dictionnaire (également appelée « modèle du code », « sémiotique » ou « sémantique ») et l'analyse du détective (également appelée « analyse gricéene », « modèle inférentiel » ou « pragmatique ») (Sperber & Wilson, 1995; Schlenker, 2016; Heintz & Scott-Phillips, 2023).

Analyse du dictionnaire. L'analyse du dictionnaire décrit la communication comme suit : les expéditeur·ices *encodent* (intentionnellement ou non) des informations dans un signal que les destinataires *décodent*. De manière vitale, avant l'échange communicatif, les expéditeur·ices et destinataires doivent partager le même *code*. Par « code », on entend ici une association préétablie entre des types de stimuli (symbolisés par « <...> ») et des ensembles d'informations (symbolisés par « [...] »). Par exemple, le code Morse consiste en une association entre <combinaisons de signaux courts et longs> et [lettres] qui doit être partagée pour communiquer avec lui. Les codes peuvent être conventionnels, comme le code Morse, mais aussi comme la sémantique formelle d'une langue : un code fait de règles syntaxiques et lexicales qui associent des <chaînes de mots> à des [significations de phrases] (Heim & Kratzer, 1998). Les codes peuvent également être non conventionnels ou « naturels » (Wharton, 2003; Bonard, 2023a). Par exemple, les abeilles utilisent un code associant leurs <danses> à la [localisation du nectar]. Comme mentionné dans la section 2, Darwin ou Ekman postulent que les humains utilisent un code transmis génétiquement qui associe des types d'<expressions faciales> à des types d'[émotions exprimées].

La principale limite de l'analyse du dictionnaire est que, parfois, les codes *sous-déterminent* le sens : les associations préétablies entre <types de stimuli> et [ensembles d'informations] sont parfois insuffisantes pour rendre compte de l'information communiquée. De manière paradigmatique, les *implicatures conversationnelles* (Grice, 1975) communiquent implicitement des informations au-delà de ce qui est linguistiquement encodé, au-delà de ce qui est déterminé par les règles syntaxiques et lexicales de la langue utilisée. Par exemple (Wilson & Sperber, 2006), si Pierre demande : « Est-ce que Jean t'a remboursé l'argent qu'il te devait ? » et Marie répond : « Il a oublié d'aller à la banque. », Pierre comprendra facilement que Marie veut dire « non » bien que le code pertinent - les règles associant <la grammaire et le lexique français> à [la signification des phrases] - soit insuffisant à lui seul pour en rendre compte, puisque le code ne dit que Jean a oublié d'aller à la banque.

Cette limite de l'analyse du dictionnaire concerne également l'expression verbale des émotions. Pour l'illustrer, revenons à la typologie de Micheli : émotions *dites*, *montrées* et *suggérées* (Micheli, 2013). En ce qui concerne les émotions *dites*, l'analyse du dictionnaire fonctionne assez bien grâce à l'association entre <mots d'émotion> (par exemple, « heureux », « merveilleux », « tristement ») et les [types d'émotion] auxquels ils font référence. Cependant, même les émotions *dites* n'encodent parfois

pas tout ce qui est communiqué. Par exemple, « Je suis triste. » est explicite sur le type d'émotion exprimée, mais n'encode pas ce sur quoi porte l'émotion. Néanmoins, dans le contexte pertinent, nous comprenons en général à propos de quoi porte la tristesse en question. L'analyse du dictionnaire s'en sort encore moins bien avec les émotions *montrées*, car celles-ci sont souvent ambiguës. Par exemple, des interjections telles que « Wow ! », « Oulalala ! », « Diantre ! », « Ah ! » et « Oh ! », bien qu'elles montrent de façon évidente qu'une émotion est exprimée, peuvent en fait exprimer une variété d'émotions positives et négatives. De plus, ces interjections n'encodent pas non plus ce sur quoi porte l'émotion en question. Cependant, les destinataires réussissent généralement à inférer ces informations. L'analyse du dictionnaire est encore plus limitée quand il s'agit d'émotions *suggérées*. En fonction de ce que croit ou souhaite la personne exprimant son émotion, une phrase ne faisant que suggérer l'émotion peut en communiquer une multitude indéfinie. Imaginez, par exemple, que quelqu'un dise « Le navire a des voiles noires. ». Dans un certain contexte, cette phrase apparemment dénuée d'affectivité peut en fait transmettre de manière poignante une émotion intense - parce que, disons, elle signifie que le fils de celui qui prononce la phrase est mort, comme dans l'histoire d'Égée et Thésée. Il convient de noter qu'au-delà de l'expression verbale, la plupart, voire tous les types d'expressions émotionnelles, sous-déterminent également ce qui est communiqué par les émotions exprimées. Les expressions faciales ou les indices acoustiques (par exemple, cris, rires, soupirs) communiquent également différentes émotions en fonction des contextes (Aviezer *et al.*, 2008; Teigen, 2008; Vlemincx *et al.*, 2009; Barrett *et al.*, 2011, 2019; Bonard, 2023b). L'analyse du code est donc aussi insuffisante pour ce genre d'expressions émotionnelles.

Comment donc les humains désambiguïsent-ils les expressions émotionnelles dans les cas où les codes des expressions émotionnelles sous-déterminent ce qui est communiqué ? Si l'on se fie à la pragmatique cognitive contemporaine, la réponse devrait se trouver dans l'analyse de la communication dite « du détective ».

L'analyse du détective. Ce que nous appelons « l'analyse du détective » est constitué d'une famille de théories développées par Paul Grice (Grice, 1957, 1989) et ses héritier·ères (pour un compte rendu, voir Bonard (2021a), chapitre 1 et appendice). Notez que bien que notre présentation vise à rester équilibrée entre différentes théories, il n'existe pas de version universellement acceptée de cette analyse.

Comme mentionné, l'analyse du détective a été développée pour rendre compte des implicatures conversationnelles, des cas où ce qui est communiqué va au-delà de ce qui est transmis par le sens littéral des mots utilisés, comme dans l'exemple de Pierre et Marie ci-dessus. Pour ce faire, l'analyse du détective conceptualise l'interprétation linguistique comme un type de raisonnement *abductif* - c'est-à-dire une inférence qui cherche la conclusion la plus simple et la plus probable en fonction des données probantes disponibles. L'analyse décrit trois sources principales de données probantes :

1. *Les codes*, par exemple les règles syntaxiques et lexicales de l'anglais ou encore les codes des expressions émotionnelles verbales et non verbales. Comme nous l'avons vu avec la typologie de Micheli (Micheli, 2013), les expressions utilisant des émotions dites (par exemple, « Je suis triste. ») et montrées (par exemple, « Wow ! ») sont partiellement comprises grâce à de tels codes, bien que ces derniers soient trop ambigus pour rendre compte de tout ce qui est communiqué ;
2. *Les attentes pragmatiques*, c'est-à-dire les attentes concernant la façon dont les gens sont censés se comporter dans des contextes donnés et en particulier en fonction du type de signal qu'ils ont reçu. Par exemple, dans les conversations, on s'attend à ce que soient dites des choses *pertinentes* quant à la question discutée (voir les maximes de conversation de Grice (Grice,

1975)). Pour cette raison, bien que ce qui est littéralement encodé dans la réponse de Marie soit que Jean a oublié d'aller à la banque, Pierre s'attendra néanmoins à ce que cela soit pertinent à la question qu'il a posée. De même, nous nous attendons à ce que les expressions émotionnelles de quelqu'un portent sur quelque chose qui lui importe particulièrement (Wharton *et al.*, 2021; Bonard, 2022). Par exemple, si quelqu'un dit « Diantre ! » après avoir reçu un compliment étonnamment gentil, nous nous attendons à ce que le compliment importe particulièrement à la personne et interpréterons l'interjection en fonction de cela ;

3. *Les connaissances partagées* (en anglais *common ground*), c'est-à-dire l'information que les participants à l'échange présument partager (Stalnaker, 2002). Par exemple, Marie et Pierre présument qu'une banque est un endroit où l'on peut retirer de l'argent. De même, on présume généralement que recevoir un compliment est une chose que l'on recherche, surtout s'il est agréablement surprenant – bien que cela ne fasse pas toujours partie des connaissances partagées, par exemple si l'on sait que le compliment vient de l'ennemi juré de la personne complimentée. C'est aussi les connaissances partagées qui nous permettent de comprendre qu'Égée peut exprimer un profond désespoir avec la phrase « Le navire a des voiles noires. ».

En se basant sur ces trois sources de données probantes, l'analyse du détective postule ensuite que l'interprète utilise ses capacités de « lecture de l'esprit » (en anglais *mindreading*, aussi appelée « théorie de l'esprit », « mentalisation » ou « cognition sociale ») pour inférer l'information la plus probable qui est implicitement communiquée. Par exemple, Pierre déduit que Marie voulait dire « non, il ne m'a pas rendu mon argent » et nous inférons que la personne qui dit « Diantre ! » après avoir reçu le compliment est probablement contente (sauf si le compliment vient de son ennemi juré). Enfin, l'analyse du détective précise que l'information ainsi inférée est ajoutée aux connaissances partagées des personnes participant à l'échange, de sorte qu'elle puisse devenir une nouvelle source de données probantes dans la suite de l'échange ou les échanges suivants.

Il est intéressant de noter que l'analyse du détective prédit que la capacité à inférer correctement ce qui est communiqué par les expressions émotionnelles dépend fortement de nos capacités de « lecture de l'esprit ». En corroboration de cette prédiction, les personnes autistes ou les enfants peuvent avoir du mal à inférer correctement le sens implicite, par exemple dans les implicatures conversationnelles (Foppolo & Mazzaggio, 2024) ou dans les expressions utilisant des émotions suggérées (Blanc & Quenette, 2017; Etienne *et al.*, 2022).

5 Les directions de recherche pour l'analyse des émotions

Vers un schéma d'annotation unifié. Pour améliorer la compréhension émotionnelle des modèles, il est souhaitable de les entraîner sur des données annotées avec un schéma rendant compte fidèlement d'une situation émotionnelle. Un tel schéma devrait intégrer différentes perspectives sur le phénomène émotionnel pour permettre de meilleures comparaisons entre les études ainsi qu'augmenter les performances et la généralisation des modèles.

Les tentatives d'unification. Plusieurs études récentes essayent d'unir différentes manières d'annoter l'émotion dans le texte. Campagnano *et al.* (2022) proposent un nouveau schéma d'annotation qui unifie plusieurs schémas sur les rôles sémantiques des émotions. Pour choisir un ensemble de catégories partagées, les différentes émotions discrètes des schémas ont été converties vers les émotions de base de la théorie de Plutchik (Plutchik, 2001). Klinger (2023) explore les divergences et les points communs entre l'analyse des rôles sémantiques de l'émotion et les approches basées sur

l'évaluation cognitive. L'étude identifie plusieurs directions de recherche, comme l'utilisation des variables de l'évaluation cognitive pour améliorer la tâche de détection des causes de l'émotion, ou l'analyse des évaluations cognitives spécifiques aux expérienceur·euses (Wegge *et al.*, 2023). Ces études montrent que l'unification des schémas permet le transfert de connaissance entre différentes tâches, ce qui augmente les performances et la généralisation des modèles.

À la recherche d'un cadre commun. Ce que nous avons appelé plus haut « le cadre intégré pour les théories de l'émotion » (section 2) vise à réconcilier les principales théories de l'émotion en sciences cognitives (Scherer, 2022). Il constitue selon nous un bon candidat pour fournir un cadre commun aux schémas d'annotation. Pour rappel, ce modèle considère qu'une émotion est constituée de changements synchronisés dans différents composants : le processus d'évaluation, les tendances à l'action, les changements corporels (expressions motrices et réponses physiologiques) et les sentiments subjectifs. La recherche en analyse des émotions doit s'inspirer des récents débats en psychologie des émotions pour faire dialoguer les schémas d'annotation existants sur une base théorique solide et, idéalement, construire un schéma d'annotation unifié.

L'émotion est constituée de plusieurs composantes en interaction. Un schéma d'annotation unifié pourrait clarifier certaines zones d'ombre existantes dans l'analyse des émotions, comme l'absence de définitions claires des rôles sémantiques liés à l'émotion, comme l'expérienceur·euse, la cause et la cible. Il pourrait aussi permettre de mieux situer les schémas existants. Par exemple, l'annotation des émotions discrètes et des dimensions affectives met l'accent sur le sentiment subjectif, alors que l'annotation des dimensions cognitives met l'accent sur l'évaluation cognitive. Peu de schémas rendent compte des réponses physiologiques, des expressions motrices et des tendances à l'action. Plus généralement, peu de schémas considèrent la totalité des composantes. Kim & Klinger (2019) analysent la communication des émotions dans des fictions à travers des descriptions de sensations subjectives, de postures, d'expressions faciales et de relations spatiales entre les personnages. Casel *et al.* (2021) associent à des empans de texte des catégories correspondant aux composantes de Scherer. Cortal *et al.* (2023) structurent des récits narratifs émotionnels selon des composantes similaires à celles de Scherer. Chaque empan de texte correspond à des comportements observables, des pensées, des ressentis physiques ou des évaluations cognitives. À notre connaissance, il n'existe pas de schémas d'annotation qui essaient de capturer l'interaction entre les composantes. Généralement, l'analyse des émotions se concentre peu sur le caractère dynamique de l'émotion et la synchronisation des diverses composantes.

Améliorer la clarté des guides d'annotation. Nous soulignons que peu d'études justifient psychologiquement le choix des différents objets à détecter dans le texte. L'analyse des émotions a besoin de développer une approche systématique pour comparer les guides d'annotation entre eux et ainsi comprendre précisément comment l'émotion est capturée par les différents schémas d'annotation. Avec des guides d'annotation clairs, il sera plus facile pour les équipes de recherche de se concentrer sur les points de convergence entre les schémas. Ainsi, ces schémas devront s'inspirer des théories en psychologie des émotions (section 2) mais aussi des théories linguistiques (sections 3 et 4) pour identifier les marqueurs linguistiques qui verbalisent l'émotion.

Vers une meilleure évaluation de la compréhension émotionnelle. Récemment, les *benchmarks* sur les émotions évaluent des modèles de langue sur certains aspects de la compréhension émotionnelle (Wang *et al.*, 2023; Paech, 2024), sans prendre en compte toute sa richesse (Scherer, 2007; Mayer *et al.*, 2008; O'Connor *et al.*, 2019). Par exemple, Paech (2024) évalue la compréhension émotionnelle à travers la prédiction de l'intensité de plusieurs émotions dans des scènes de conflits. Il existe aussi des *benchmarks* qui évaluent des modèles sur des tâches connexes, comme l'analyse du sentiment

(Zhang *et al.*, 2023) et la théorie de l'esprit (Zhou *et al.*, 2023; Ma *et al.*, 2023; Kim *et al.*, 2023; Gandhi *et al.*, 2023). Ainsi, il n'existe aucun *benchmark* qui propose d'évaluer spécifiquement plusieurs aspects du phénomène émotionnel. Il est donc difficile de savoir si les modèles actuels sont performants pour la compréhension émotionnelle.

Cette limite s'ajoute au fait qu'il est difficile de déterminer clairement les propriétés de la compréhension émotionnelle à évaluer. Nous pensons qu'il faudrait s'inspirer de la communication émotionnelle chez les humains pour évaluer les modèles de langue, et notamment des travaux en psycholinguistique. Ainsi, avant dix ans, les émotions de base (par exemple, la joie ou la tristesse) sont mieux retenues que les émotions complexes (par exemple, la fierté ou la culpabilité) (Davidson *et al.*, 2001; Creissen & Blanc, 2017). De six à dix ans, les émotions *dites* sont mieux comprises que les émotions *suggérées* (Blanc, 2010; Creissen & Blanc, 2017). Un autre exemple d'études pertinentes concerne la plus ou moins grande facilité qu'ont les personnes sur le spectre de l'autisme à comprendre différents types d'expressions émotionnelles (Foppolo & Mazzaggio, 2024). Ces études montrent que, pour les humains, différents types d'émotions et différents modes d'expressions émotionnels sont plus ou moins difficiles à interpréter. Il serait souhaitable que les *benchmarks* évaluent les modèles de langue de manière à refléter la plus ou moins grande difficulté des tâches pour les humains. Un tel projet bénéficierait certainement des recherches en pragmatique cognitive (section 4) sachant, par exemple, que des personnes souffrant de troubles de la communication ont du mal à comprendre les implicatures conversationnelles (Foppolo & Mazzaggio, 2024), ce qui indique que les différentes sources de données probantes distinguées par l'analyse du détective impliquent différents degrés de difficultés.

Nous pensons que le concept d'émotion doit être adressé à travers sa relation avec la compréhension du texte, c'est-à-dire la capacité qu'à un·e lecteur·rice à construire une représentation mentale d'une situation dans un texte (Zwaan & Radvansky, 1998). Ainsi, il faudrait aller au-delà des conceptualisations courantes de l'émotion en traitement automatique des langues (section 3) pour prendre en compte la diversité des marqueurs linguistiques employés pour verbaliser l'émotion (section 3) ainsi que les différents types d'émotion (basique ou complexe) issus des travaux en psycholinguistiques. Etienne *et al.* (2022) ont proposé un schéma d'annotation inspiré par les études précédentes qui considère les modes d'expression de l'émotion et les types d'émotion. De futurs *benchmarks* évaluant les capacités des modèles de langue à analyser les émotions devraient prendre en compte de tels schémas d'annotation qui, comme nous l'avons recommandé, cherchent à solidement se baser sur les recherches pertinentes en sciences cognitives.

6 Conclusion

Pour remédier à certaines limites dans l'analyse des émotions, nous avons proposé d'exploiter les recherches en sciences cognitives sur les émotions et la communication. Nous avons expliqué pourquoi et comment l'analyse des émotions devrait utiliser des théories en psychologie des émotions – en particulier le cadre intégré – ainsi que des théories en pragmatique cognitive – en particulier l'analyse du détective. Ces recherches ouvrent la voie à la construction de nouveaux schémas d'annotation et d'un possible *benchmark* pour la compréhension émotionnelle, considérant différentes facettes de l'émotion et de la communication chez l'humain.

Références

APPERLY I. A. & BUTTERFILL S. A. (2009). Do humans have two systems to track beliefs and belief-like states ? *Psychological review*, **116**(4), 953. Publisher : American Psychological Association.

ARNOLD M. B. (1960). *Emotion and Personality*. New York : Columbia University Press.

ARU J., LABASH A., CORCOLL O. & VICENTE R. (2023). Mind the gap : challenges of deep learning approaches to Theory of Mind. *Artificial Intelligence Review*, **56**(9), 9141–9156. DOI : 10.1007/s10462-023-10401-x.

AVIEZER H., HASSIN R. R., RYAN J., GRADY C., SUSSKIND J., ANDERSON A., MOSCOVITCH M. & BENTIN S. (2008). Angry, disgusted, or afraid ? Studies on the malleability of emotion perception. *Psychological science*, **19**(7), 724–732. Publisher : SAGE Publications Sage CA : Los Angeles, CA.

BARRETT L. F. (2006). Solving the Emotion Paradox : Categorization and the Experience of Emotion. *Personality and Social Psychology Review*, **10**(1), 20–46. DOI : 10.1207/s15327957pspr1001_2.

BARRETT L. F. (2017). *How Emotions Are Made : The Secret Life of the Brain*. Boston & New York : Houghton Mifflin Harcourt.

BARRETT L. F., ADOLPHS R., MARSELLA S., MARTINEZ A. M. & POLLAK S. D. (2019). Emotional expressions reconsidered : Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*. Publisher : SAGE PublicationsSage CA : Los Angeles, CA, DOI : 10.1177/1529100619832930.

BARRETT L. F., MESQUITA B. & GENDRON M. (2011). Context in emotion perception. *Current Directions in Psychological Science*, **20**(5), 286–290. Publisher : Sage Publications Sage CA : Los Angeles, CA.

BLANC N. (2010). La compréhension des contes entre 5 et 7 ans : Quelle représentation des informations émotionnelles ? [The comprehension of the tales between 5 and 7 year-olds : Which representation of emotional information ?]. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, **64**(4), 256–265. DOI : 10.1037/a0021283.

BLANC N. & QUENETTE G. (2017). La production d'inférences émotionnelles entre 8 et 10 ans : quelle méthodologie pour quels résultats ? *Enfance*, **4**(4), 503–511. Publisher : NecPlus.

BONARD C. (2021a). *Meaning and emotion : The extended Gricean model and what emotional signs mean*. Doctoral dissertation, University of Geneva and University of Antwerp.

BONARD C. (2021b). Émotions et sensibilité aux valeurs : quatre conceptions philosophiques contemporaines. *Revue de métaphysique et de morale*, **110**(2), 209–229. Place : Paris cedex 14 Publisher : Presses Universitaires de France, DOI : 10.3917/rmm.212.0209.

BONARD C. (2022). Beyond ostension : Introducing the expressive principle of relevance. *Journal of Pragmatics*, **187**, 13–23. DOI : 10.1016/j.pragma.2021.10.024.

BONARD C. (2023a). Natural meaning, probabilistic meaning, and the interpretation of emotional signs. *Synthese*, **201**(5), 167. Publisher : Springer, DOI : https ://doi.org/10.1007/s11229-023-04144-z.

BONARD C. (2023b). Underdeterminacy without ostension : A blind spot in the prevailing models of communication. *Mind & Language*. DOI : https ://doi.org/10.1111/mila.12481.

BONARD C. (2024). Can AI and humans genuinely communicate ? In A. STRASSER, Éd., *Anna's AI Anthology. How to live with smart machines ?* Berlin : Xenemoi.

BONARD C. & DEONNA J. (2023). Emotion and language in philosophy. In G. L. SCHIEWER, J. ALTARRIBA & B. C. NG, Éds., *Language and emotion : An international handbook*, volume 1, p. 54–72. Berlin : de Gruyter.

BOSTAN L. A. M., KIM E. & KLINGER R. (2020). GoodNewsEveryone : A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 1554–1566, Marseille, France : European Language Resources Association.

BOSTAN L.-A.-M. & KLINGER R. (2018). An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 2104–2119, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I. & AMODEI D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, p. 1877–1901 : Curran Associates, Inc.

BUECHEL S. & HAHN U. (2017). EmoBank : Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 578–585, Valencia, Spain : Association for Computational Linguistics.

BUECHEL S., MODERSOHN L. & HAHN U. (2021). Towards label-agnostic emotion embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 9231–9249, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.728.

CAMPAGNANO C., CONIA S. & NAVIGLI R. (2022). SRL4E – Semantic Role Labeling for Emotions : A unified evaluation framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 4586–4601, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.314.

CASEL F., HEINDL A. & KLINGER R. (2021). Emotion recognition under consideration of the emotion component process model. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, p. 49–61, Düsseldorf, Germany : KONVENS 2021 Organizers.

CORTAL G., FINKEL A., PAROUBEK P. & YE L. (2023). Emotion recognition based on psychological components in guided narratives for emotion regulation. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, p. 72–81, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.latechclfl-1.8.

COSMIDES L. & TOOBY J. (2000). Evolutionary psychology and the emotions. In M. LEWIS & J. M. HAVILAND-JONES, Éds., *Handbook of emotions*, p. 91–115. New York : Guilford Press, 2nd édition. Publisher : Citeseer.

CREISSEN S. & BLANC N. (2017). Quelle représentation des différentes facettes de la dimension émotionnelle d'une histoire entre l'âge de 6 et 10 ans ? Apports d'une étude multimédia. *Psychologie Française*, **62**(3), 263–277. DOI : 10.1016/j.psfr.2015.07.006.

DARWIN C. (1872). *The expression of the emotions in man and animals*. London : John Murray.

DAVIDSON D., LUO Z. & BURDEN M. J. (2001). Children's recall of emotional behaviours, emotional labels, and nonemotional behaviours : Does emotion enhance memory ? *Cognition and Emotion*, **15**(1), 1–26. DOI : 10.1080/0269993004200105.

DE BRUYNE L., DE CLERCQ O. & HOSTE V. (2020). An emotional mess ! deciding on a framework for building a Dutch emotion-annotated corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 1643–1651, Marseille, France : European Language Resources Association.

DEMSZKY D., MOVSHOVITZ-ATTIAS D., KO J., COWEN A., NEMADE G. & RAVI S. (2020). GoEmotions : A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 4040–4054, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.372.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

EKMAN P. (1999). Basic emotions. In T. DALGLEISH & M. J. POWER, Éds., *Handbook of cognition and emotion*, p. 45–60. Chichester : John Wiley & Sons Ltd.

EKMAN P. & FRIESEN W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, **17 2**, 124–9.

ETIENNE A., BATTISTELLI D. & LECORVÉ G. (2022). A (psycho-)linguistically motivated scheme for annotating and exploring emotions in a genre-diverse corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 603–612, Marseille, France : European Language Resources Association.

FOPPOLO F. & MAZZAGGIO G. (2024). Conversational Implicature and Communication Disorders. In M. J. BALL, N. MÜLLER & E. SPENCER, Éds., *The Handbook of Clinical Linguistics, Second Edition*, p. 15–27. Wiley, 1 édition. DOI : 10.1002/9781119875949.ch2.

GANDHI K., FRÄNKEN J.-P., GERSTENBERG T. & GOODMAN N. D. (2023). Understanding Social Reasoning in Language Models with Language Models. DOI : 10.48550/arXiv.2306.15448.

GREEN M. (2007). *Self-expression*. Oxford : Oxford University Press.

GRICE H. P. (1957). Meaning. *The Philosophical Review*, **66**(3), 377–388.

GRICE H. P. (1975). Logic and conversation. In *Speech acts*, p. 41–58. Leiden : Brill.

GRICE H. P. (1989). *Studies in the way of words*. Cambridge (MA) : Harvard University Press.

HEIM I. & KRATZER A. (1998). *Semantics in generative grammar*. Hoboken : Wiley. Google-Books-ID : jAvR2DB3pPIC.

HEINTZ C. & SCOTT-PHILLIPS T. (2023). Expression unleashed : The evolutionary & cognitive foundations of human communication. *Behavioral and Brain Sciences*, **46**, E1. type : article, DOI : 10.31234/osf.io/mcv5b.

HOFMANN J., TROIANO E., SASSENBERG K. & KLINGER R. (2020). Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 125–138, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.11.

HOLTERMAN B. & VAN DEEMTER K. (2023). Does ChatGPT have Theory of Mind ? arXiv :2305.14020 [cs].

IZARD C. E. (1992). Basic Emotions, Relations Among Emotions, and Emotion-Cognition Relations. *Psychological Review*, **99**(3), 561–565.

KIM E. & KLINGER R. (2018). Who Feels What and Why ? Annotation of a Literature Corpus with Semantic Roles of Emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 1345–1359, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

KIM E. & KLINGER R. (2019). An analysis of emotion communication channels in fan-fiction : Towards emotional storytelling. In *Proceedings of the Second Workshop on Storytelling*, p. 56–64, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-3406.

KIM H., SCLAR M., ZHOU X., BRAS R. L., KIM G., CHOI Y. & SAP M. (2023). FANToM : A Benchmark for Stress-testing Machine Theory of Mind in Interactions. DOI : 10.48550/arXiv.2310.15421.

KLINGER R. (2023). Where are We in Event-centric Emotion Analysis ? Bridging Emotion Role Labeling and Appraisal-based Approaches. In Y. ELAZAR, A. ETTINGER, N. KASSNER, S. RUDER & N. A. SMITH, Éds., *Proceedings of the Big Picture Workshop*, p. 1–17, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.bigpicture-1.1.

KOSINSKI M. (2023). Theory of Mind Might Have Spontaneously Emerged in Large Language Models. arXiv :2302.02083 [cs], DOI : 10.48550/arXiv.2302.02083.

LANGLEY C., CIRSTEA B. I., CUZZOLIN F. & SAHAKIAN B. J. (2022). Theory of Mind and Preference Learning at the Interface of Cognitive Science, Neuroscience, and AI : A Review. *Frontiers in Artificial Intelligence*, **5**.

LAZARUS R. S. (1991). Progress on a cognitive-motivational-relational theory of emotion. *American psychologist*, **46**(8), 819.

LEE S. Y. M., CHEN Y. & HUANG C.-R. (2010). A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, p. 45–53, Los Angeles, CA : Association for Computational Linguistics.

MA Z., SANSOM J., PENG R. & CHAI J. (2023). Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. DOI : 10.48550/arXiv.2310.19619.

MAYER J. D., ROBERTS R. D. & BARSADE S. G. (2008). Human Abilities : Emotional Intelligence. *Annual Review of Psychology*, **59**(1), 507–536. DOI : 10.1146/annurev.psych.59.103006.093646.

MICHELI R. (2013). Esquisse d'une typologie des différents modes de sémiotisation verbale de l'émotion. *Semen*, (35). DOI : 10.4000/semen.9795.

MICHELI R. (2014). *Les émotions dans les discours*. De Boeck Supérieur. DOI : 10.3917/dbu.mchel.2014.01.

MITCHELL M. & KRAKAUER D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, **120**(13), e2215907120. Publisher : Proceedings of the National Academy of Sciences, DOI : 10.1073/pnas.2215907120.

MOORS A. (2022). *Demystifying emotions : A Typology of theories in psychology and philosophy*. Cambridge, cambridge university press édition.

MOORS A., ELLSWORTH P. C., SCHERER K. R. & FRIJDA N. H. (2013). Appraisal theories of emotion : state of the art and future development. *Emotion Review*, **5**(2), 119–124. Publisher : Sage Publications Sage UK : London, England.

OBERLÄNDER L., REICH K. & KLINGER R. (2020). Experiencers, Stimuli, or Targets : Which Semantic Roles Enable Machine Learning to Infer the Emotions ? *arXiv :2011.01599 [cs]*.

O'CONNOR P. J., HILL A., KAYA M. & MARTIN B. (2019). The measurement of emotional intelligence : A critical review of the literature and recommendations for researchers and practitioners. *Frontiers in psychology*, **10**, 1116. Publisher : Frontiers.

PAECH S. J. (2024). EQ-Bench : An Emotional Intelligence Benchmark for Large Language Models. DOI : 10.48550/arXiv.2312.06281.

PANKSEPP J. (1998). *Affective neuroscience : the foundations of human and animal emotions*. New York : Oxford University Press.

PLUTCHIK R. (2001). The Nature of Emotions : Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, **89**(4), 344–350.

RUSSELL J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, **39**(6), 1161. Publisher : American Psychological Association.

RUSSELL J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, **110**(1), 145. Publisher : American Psychological Association.

RUSSELL J. A. (2009). Emotion, core affect, and psychological construction. *Cognition and Emotion*, **23**(7), 1259–1283. DOI : 10.1080/02699930902809375.

SCARANTINO A. (2017). How to do things with emotional expressions : The theory of affective pragmatics. *Psychological Inquiry*, **28**(2-3), 165–185. Publisher : Taylor & Francis.

SCHACHTER S. & SINGER J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological review*, **69**(5), 379. Publisher : American Psychological Association.

SCHERER K. R. (2007). Componential emotion theory can inform models of emotional competence. Publisher : Oxford University Press.

SCHERER K. R. (2022). Theory convergence in emotion science is timely and realistic. *Cognition and Emotion*, **36**(2), 154–170. DOI : 10.1080/02699931.2021.1973378.

SCHERER K. R. & MOORS A. (2019). The emotion process : event appraisal and component differentiation. *Annual Review of Psychology*, **70**, 719–745. Publisher : Annual Reviews.

SCHLENKER P. (2016). The semantics-pragmatics interface. In M. ALONI & P. DEKKER, Éds., *The Cambridge Handbook of Formal Semantics*, p. 664–727. Cambridge : Cambridge University Press.

SHAPIRA N., LEVY M., ALAVI S. H., ZHOU X., CHOI Y., GOLDBERG Y., SAP M. & SHWARTZ V. (2023). Clever Hans or Neural Theory of Mind ? Stress Testing Social Reasoning in Large Language Models. arXiv :2305.14763 [cs], DOI : 10.48550/arXiv.2305.14763.

SPERBER D. & WILSON D. (1995). *Relevance : Communication and cognition*. Oxford and Cambridge (MA) : Blackwell, 2nd edition édition.

STALNAKER R. (2002). Common ground. *Linguistics and philosophy*, **25**(5/6), 701–721.

STOJNIĆ G., GANDHI K., YASUDA S., LAKE B. M. & DILLON M. R. (2023). Commonsense psychology in human infants and machines. *Cognition*, **235**, 105406. DOI : https ://doi.org/10.1016/j.cognition.2023.105406.

TEIGEN K. H. (2008). Is a sigh "just a sigh" ? Sighs as emotional signals and responses to a difficult task. *Scandinavian journal of Psychology*, **49**(1), 49–57. Publisher : Wiley Online Library.

TOMKINS S. (1962). *Affect imagery consciousness*, volume Volume I : The positive affects. New York : Springer.

TROIANO E., OBERLÄNDER L. & KLINGER R. (2022). Dimensional Modeling of Emotions in Text with Appraisal Theories : Corpus Creation, Annotation Reliability, and Prediction. *Computational Linguistics*, p. 1–71. DOI : 10.1162/coli_a_00461.

TROTT S., JONES C., CHANG T., MICHAELOV J. & BERGEN B. (2022). Do Large Language Models know what humans know ? *arXiv preprint arXiv :2209.01515*.

ULLMAN T. (2023). Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. arXiv :2302.08399 [cs].

VLEMINCX E., VAN DIEST I., DE PEUTER S., BRESSELEERS J., BOGAERTS K., FANNES S., LI W. & VAN DEN BERGH O. (2009). Why do you sigh ? Sigh rate during induced stress and relief. *Psychophysiology*, **46**(5), 1005–1013. Publisher : Wiley Online Library.

WANG X., LI X., YIN Z., WU Y. & LIU J. (2023). Emotional intelligence of Large Language Models. *Journal of Pacific Rim Psychology*, **17**, 18344909231213958. DOI : 10.1177/18344909231213958.

WEGGE M., TROIANO E., OBERLÄNDER L. & KLINGER R. (2023). Experiencer-Specific Emotion and Appraisal Prediction. DOI : 10.48550/arXiv.2210.12078.

WHARTON T. (2003). Natural pragmatics and natural codes. *Mind & language*, **18**(5), 447–477. Publisher : Wiley Online Library.

WHARTON T. (2016). That bloody so-and-so has retired : Expressives revisited. *Lingua*, **175**, 20–35. Publisher : Elsevier.

WHARTON T., BONARD C., DUKES D., SANDER D. & OSWALD S. (2021). Relevance and emotion. *Journal of Pragmatics*, **181**, 259–269.

WILSON D. & SPERBER D. (2006). Relevance theory. In L. HORN, Éd., *The Handbook of pragmatics*. Oxford : Blackwell.

ZHAN H., ONG D. & LI J. J. (2023). Evaluating Subjective Cognitive Appraisals of Emotions from Large Language Models. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 14418–14446, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.962.

ZHANG W., DENG Y., LIU B., PAN S. J. & BING L. (2023). Sentiment Analysis in the Era of Large Language Models : A Reality Check. DOI : 10.48550/arXiv.2305.15005.

ZHOU P., MADAAN A., POTHARAJU S. P., GUPTA A., MCKEE K. R., HOLTZMAN A., PUJARA J., REN X., MISHRA S., NEMATZADEH A., UPADHYAY S. & FARUQUI M. (2023). How FaR Are Large Language Models From Agents with Theory-of-Mind ? DOI : 10.48550/arXiv.2310.03051.

ZWAAN R. A. & RADVANSKY G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, **123**(2), 162–185. DOI : 10.1037/0033-2909.123.2.162.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Analyse de la perception de l'offre INTERCITÉS de jour : Classification multi-étiquettes des émotions dans les tweets

Chang Liu[1] Hélène Flamein[2] Luce Lefeuvre[2] Fanny Hanen[1]
(1) SNCF Voyageurs INTERCITÉS Direction Marketing, 2 rue Traversière, 75012 PARIS
(2) SNCF DTIPG, 1-3 avenue François Mitterrand, 93210 Saint-Denis, France
ext.chang.liu@sncf.fr, helene.flamein2@gmail.com, luce.lefeuvre@sncf.fr, fanny.hanen@sncf.fr

RÉSUMÉ

La Direction Marketing de SNCF Voyageurs INTERCITÉS souhaite améliorer l'expérience des voyageurs en procédant à l'analyse automatique de la perception de son offre à travers les ressentis partagés sur les réseaux sociaux. L'un des axes de notre recherche se focalise sur la détection des émotions en multi-étiquettes qui traduisent cette perception. Pour accomplir cette tâche, nous ajustons tout d'abord un modèle de langue pré-entraîné à l'aide d'un corpus préalablement annoté en émotions, puis nous le spécialisons sur notre corpus, axé sur le contexte ferroviaire d'INTERCITÉS. Notre approche obtient un F1-Micro score de 0,55, un F1-Macro score de 0,44 et une exactitude de 0,826.

ABSTRACT

Analysis of the perception of the INTERCITÉS day train service : Multi-Label classification of emotions in tweets

The Marketing Department of SNCF Voyageurs INTERCITÉS aims to improve the passenger experience by conducting an automatic analysis of the perception of its service through feelings shared on social media. Our research focuses on the detection of multi-label emotions that reflect this perception. To accomplish this task, we first adjust a pre-trained language model using a corpus previously annotated with emotions. Then we specialize our model on our corpus, specific to the INTERCITÉS railway context. Our approach achieves a F1-Micro score of 0.55, a F1-Macro score of 0,44, and an accuracy of 0.826.

MOTS-CLÉS : Perception, Détection des émotions, Classification multi-étiquettes, CamemBERT, Mesures d'évaluation.

KEYWORDS: Perception, Emotion Detection, Multi-label Classification, CamemBERT, Evaluation Measures.

1 Introduction

L'amélioration constante de l'expérience client demeure une préoccupation centrale pour les entreprises opérant dans le secteur des services, et plus particulièrement dans le domaine du transport ferroviaire. Dans cette ère numérique, où les clients comme les non-clients expriment leurs opinions et ressentis de manière instantanée et publique sur les plateformes en ligne, l'analyse des données issues des réseaux sociaux offre de nouvelles perspectives pour mieux comprendre les attentes, les

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 323–337.

préoccupations et les satisfactions des voyageurs. En nous basant sur un corpus de tweets concernant l'offre INTERCITÉS de jour, notre recherche s'attache à analyser les émotions des utilisateurs telles qu'elles transparaissent dans les tweets. L'analyse des émotions des tweets est une tâche complexe, non seulement en raison de leur nature textuelle, mais aussi parce qu'un seul tweet peut véhiculer des émotions complexes, définies comme une combinaison d'émotions simples ou de processus cognitifs plus nuancés, tels que l'amour ou la culpabilité. L'annotation des émotions en multi-étiquettes semble nécessaire et permet de saisir la nuance des messages humains. Dans un tweet, un auteur peut en effet exprimer des avis divergents sur différentes thématiques. Par exemple, dans le tweet suivant issu de notre corpus INTERCITÉS, «@Intercites Bah t'es bien t'es coincé tu ne peux plus bosser», l'auteur exprime à la fois de l'«angoisse» et de la «colère». Nous avons ainsi exploré les stratégies permettant de surmonter la complexité de la détection de plusieurs émotions au sein d'un même tweet. De plus, nous nous sommes interrogées sur les moyens de gérer un corpus d'étude non annoté. En réponse à ces défis, un modèle de classification multi-étiquettes a été entraîné sur un corpus préalablement annoté en émotions selon une typologie adaptée à nos données et aux objectifs du projet.

2 État de l'art : détection des émotions dans les tweets

Dans le domaine de l'analyse des émotions, plusieurs auteurs se sont intéressés à leur classification, notamment en psychologie. Ainsi, Ekman (1992) identifie six émotions fondamentales : la colère, le dégoût, la peur, la joie, la tristesse et la surprise. Pour classifier les émotions, Plutchik (1980), s'inspirant du modèle d'Ekman, propose une visualisation en forme de roue qui comprend quatre ensembles bipolaires : joie et tristesse ; colère et peur ; confiance et dégoût ; surprise et anticipation. En TAL, les méthodes élémentaires de détection des émotions reposent sur l'utilisation de lexiques pour identifier des émotions liées à des états psychologiques via des mots-clés, une approche soulignée dans l'étude de Rabeya *et al.* (2017). Deux lexiques sont principalement utilisées dans ces approches : WordNet-Affect (Strapparava *et al.*, 2004) et NRC Word-Emotion Association Lexicon (Mohammad & Turney, 2013) également appelé EmoLex.

Les données issues des réseaux socio-numériques, et plus particulièrement de Twitter/X, ont donné lieu à de nombreux travaux en *sentiment analysis*, *emotion recognition* ou *opinion mining*, notamment parce qu'ils sont un lieu de polémique. D'un point de vue linguistique, les tweets peuvent inclure des textes linéaires simples, des émoticônes, des liens URL vers des sites externes, ainsi que des éléments spécifiques aux réseaux sociaux tels que les hashtags (marqués par un #) pour organiser l'information et les pseudos (précédés de @) qui identifient les utilisateurs (Paveau, 2017). D'après Baziotis *et al.* (2018), l'utilisation de ces formes de communication spécifiques éloigne les tweets des structures linguistiques traditionnelles et rend leur traitement complexe.

Ainsi, diverses approches sont étudiées pour détecter les émotions dans les tweets. Lora *et al.* (2020) évaluent plusieurs techniques d'apprentissage automatique, comme les modèles bayésiens naïfs, les SVM (*Support Vector Machine*), et la Régression Logistique, ainsi que diverses architectures de réseaux de neurones. Les méthodes d'apprentissage automatique affichent un score F1 supérieur ; toutefois, l'approche utilisant les CNN (*Convolutional Neural Networks*) avec des plongements lexicaux pré-entraînés surpasse les autres approches en termes d'exactitude. Plusieurs travaux suggèrent l'utilisation de réseaux de neurones récurrents, en particulier les LSTM (*Long Short-Term Memory*), pour effectuer la détection des émotions dans les tweets (Kabir & Madria, 2021 ; Javed & Muralidhara, 2022). Aslam *et al.* (2022) combinent deux réseaux neuronaux récurrents différents, les LSTM et les

GRU (*Gated Recurrent Units*), et leur modèle atteint une exactitude de 0.91. Les auteurs observent que diminuer la taille du corpus d'entraînement entraîne une baisse de performance du modèle. Par ailleurs, l'utilisation des modèles basés sur les transformers (Vaswani *et al.*, 2017) est largement répandue (Yu *et al.*, 2018; Camacho-Collados *et al.*, 2022), et plus particulièrement le modèle de langage BERT (Devlin *et al.*, 2018). Notamment, Chowdhury & Pal (2023) appliquent des LSTM après avoir ajusté le modèle BERT pour la classification des émotions ; cette approche combinatoire permet d'atteindre un score F1 de 0,71.

Quant à la classification multi-étiquettes des émotions, Kim & Klinger (2018) présentent une analyse des émotions dans les textes littéraires, ayant pour objectif d'identifier les émotions et de les relier aux personnages, à leur origine (causes) et à leurs cibles. Ils construisent pour cela des modèles LSTM bidirectionnels avec une couche de CRF (*Conditional Random Fields*). L'étude montre que l'intégration des CRF améliore l'étiquetage des séquences lorsque les données montrent des dépendances entre les éléments. Enfin, dans sa thèse, Etienne (2023) introduit un modèle capable de réaliser des prédictions simultanées du caractère émotionnel d'un texte, des modes d'expression présents dans les textes, et des types d'émotions véhiculées, tout en effectuant une classification multi-étiquettes des catégories émotionnelles. L'auteure démontre que l'approche prenant en compte plusieurs aspects émotionnels peut améliorer la robustesse du modèle.

Beaucoup plus récemment, dans le domaine des LLMs (*Larges Language Models*) génératifs, Wang *et al.* (2023) développent un test psychométrique nommé SECEU (*Situational Evaluation of Complex Emotional Understanding*) et l'appliquent sur différents LLMs pour évaluer la compréhension émotionnelle des LLMs et des humains. Le LLM génératif le plus efficace pour comprendre les émotions s'est avéré être GPT-4. Liu *et al.* (2024) décrivent quant à eux EmoLLMs, formés par l'affinage de différents LLMs en utilisant un ensemble de données nommé AAID (*Affective Analysis Instruction Dataset*) construit par le SemEval-2018 (Mohammad *et al.*, 2018). Dans les tâches SemEval-2018, ces modèles surpassent les LLMs génératifs open source.

Certaines études se heurtent au manque de données annotées pour détecter les émotions dans les textes. Pour répondre à cette problématique, différentes solutions sont proposées. Par exemple, Baziotis *et al.* (2018) utilisent un corpus de 550 millions de tweets en anglais pour entraîner les plongements lexicaux et un deuxième corpus de 61 854 tweets constitué par SemEval-2017 (Rosenthal *et al.*, 2019) dans le but de réaliser un apprentissage par transfert. Le modèle est ensuite affiné à l'aide des données dans le corpus de SemEval-2018. Cette étude montre que l'utilisation de plusieurs corpus permet de créer un ensemble de données diversifié et représentatif des différentes expressions des émotions. Cependant, cela augmente le coût en temps pour normaliser et harmoniser les données. Guibon *et al.* (2021), pour leur part, explorent l'utilisation de deux corpus distincts pour la classification des émotions dans des dialogues : le premier est utilisé pour la phase d'entraînement, et le second pour évaluer la performance du modèle. L'utilisation de différents corpus permet de tester la généralisation du modèle à diverses sources de données et également d'offrir une évaluation robuste de sa performance.

Compte tenu des travaux existants, plusieurs paramètres doivent être considérés afin de capter les émotions des utilisateurs à propos des services INTERCITÉS : le genre textuel ainsi que les émotions significatives présentes dans notre corpus, et l'architecture du système de prédiction. Ainsi, nous envisageons d'explorer l'utilisation de modèles basés sur les transformers, en particulier l'architecture BERT (Luo & Wang, 2019; Huang *et al.*, 2019; Camacho-Collados *et al.*, 2022). La combinaison de divers modèles nous parait aussi pertinente comme approche, car cela permet d'avoir une architecture plus complexe et de traiter plus finement les données afin d'avoir de meilleures performances, notamment pour la classification des émotions (Aslam *et al.*, 2022; Chowdhury & Pal, 2023).

3 Jeux de données et typologie des émotions

3.1 Présentation du corpus d'étude

Le corpus principal de notre étude englobe l'ensemble des tweets en français diffusés au cours de l'année 2022 et intégrant les termes « intercités » ou « intercité ». Notre étude se concentre exclusivement sur les tweets des utilisateurs relatifs aux lignes INTERCITÉS de jour. L'objectif est de cerner précisément les publications provenant de voyageurs actuels ou potentiels. Dans cette optique, nous avons essayé d'exclure les tweets évoquant l'offre INTERCITÉS de nuit et ceux venant de comptes affiliés à la SNCF ou à des médias, ainsi que ceux se référant à d'autres lignes de transport. Nous avons également éliminé les hyperliens tout en préservant les emojis, ces derniers étant convertis en format textuel cf. (Ex. 1).

(Ex. 1) *1/3 :stop_sign : Service dégradé :stop_sign : Des nouvelles de la ligne POLT ! L'INTERCITÉS de Toulouse est arrivé à Austerlitz avec 2h10 de retard. INTERCITÉS 3665 retardé d'une heure. Pourquoi ? Réponse d'un contrôleur » @Intercites @SNCF*

Après nettoyage, le corpus d'étude comporte 11 025 tweets non-annotés pour un total de 276 716 tokens, hors mentions et hashtags. Il recense 24 718 mentions et 2 272 hashtags. Les tailles minimale, moyenne et maximale des tweets sont respectivement de 1, 24 et 72 tokens.

3.2 Constitution du corpus de référence

Pour analyser les émotions présentes dans notre corpus, une typologie adaptée aux données et à nos objectifs a été choisie. Nous travaillons notamment sur les émotions simples, parce que nous considérons qu'elles répondent aux besoins d'identifier la perception qu'ont les clients et non-clients de l'offre INTERCITÉS, et parce qu'elles facilitent la conception et l'annotation du corpus. Nous nous sommes ainsi inspirées du modèle d'émotions élaboré par Plutchik (1980) et l'avons ajusté. Les émotions sur lesquelles nous travaillons sont les suivantes : joie, tristesse, angoisse, colère, dégoût, neutre. Nous avons décidé de regrouper l'anticipation et la peur du modèle de Plutchik sous l'étiquette «angoisse». En effet, la peur est peu représentée dans les données ferroviaires. L'analyse linguistique de quelques tweets nous a amenées à considérer que l'angoisse est une émotion fréquemment ressentie par les voyageurs en cas de manque d'information et/ou de situation perturbée. L'intensité de cette émotion se situe entre celle de la peur et celle de l'anticipation, avec une nuance plus pondérée. Deux annotateurs spécialistes en linguistique ont annoté un échantillon de 249 tweets en multi-étiquettes selon la typologie mentionnée plus haut, et le taux d'accord global pour l'ensemble des étiquetages, mesuré par le coefficient Kappa de Cohen (Cohen, 1960), est de 0,69. Cette valeur suggère un accord substantiel selon Landis & Koch (1977) mais imparfait entre les annotateurs, indiquant que certaines ambiguïtés subsistent dans le processus d'annotation. Néanmoins, pour les cas les plus difficiles, les annotateurs sont parvenus à un consensus sur l'annotation de l'échantillon de tweets. Cette annotation permet de confirmer l'hypothèse selon laquelle un tweet peut être marqué de plusieurs émotions. Les statistiques de l'annotation sont présentées dans les Tables 1 et 2.

Cet échantillon est considéré comme notre corpus de référence, et sert à valider le comportement du modèle dédié à la détection des émotions. Il est à noter que, dans ce corpus de référence, la répartition des 301 émotions identifiées reste déséquilibrée : les émotions de tristesse (4%) et de dégoût (6%) sont nettement moins représentées que les autres émotions. En outre, bien que les tweets annotés en

uni-étiquette restent largement majoritaires (77%), la part de tweets annotés en multi-étiquettes (23%) reste significative.

Émotions	Nombre	Pourcentage
Colère	82	27%
Joie	74	25%
Neutre	58	19%
Angoisse	57	19%
Tristesse	19	6%
Dégoût	11	4%
Total	301	100%

TABLE 1 – Répartition des émotions dans le corpus de référence annoté en uni-étiquettes et multi-étiquettes

Type d'annotation	Nombre de tweets	Pourcentage
Uni-étiquette	192	77%
Multi-étiquette	57	23%
Total	249	100%

TABLE 2 – Répartition des étiquettes uniques ou multiples dans le corpus de référence

3.3 Constitution du corpus d'entraînement

Tout entraînement de modèle pour des tâches de traitement de données langagières nécessite un nombre important de données d'exemples. Comme mentionné précédemment, le corpus INTERCITÉS n'est pas annoté en émotions et nous ne disposons pas de données équivalentes annotées. Pour franchir ce premier obstacle, notre stratégie consiste à entraîner le modèle sur un corpus annoté pré-existant, et à appliquer ensuite les connaissances acquises par le modèle pour annoter notre corpus d'étude en émotions. En effet, selon l'état de l'art, la fusion de données provenant de différents corpus est une solution viable en cas de manque de données (Baziotis *et al.*, 2018; Guibon *et al.*, 2021). Après l'examen de plusieurs corpus, nous décidons d'utiliser le corpus DEFT 2015 [1] (Hamon *et al.*, 2015), qui regroupe des tweets axés sur le thème du changement climatique. Dans ce corpus, chaque tweet est annoté pour refléter une émotion, une opinion ou un sentiment, en utilisant un système de 19 étiquettes définies par Fraisse & Paroubek (2014) [2]. Il existe une disparité notable entre la typologie des étiquettes d'annotation utilisées dans ce corpus et celle que nous avons établie pour notre projet. En conséquence, nous avons converti les étiquettes pour aligner les annotations du corpus DEFT avec notre typologie (cf. section 3.2). Ce processus a donné naissance à notre corpus d'entraînement, que nous appelons CoarseDEFT.

La distribution des émotions dans le corpus d'entraînement est déséquilibrée (Table 3) et diffère de celle constatée dans le corpus de référence. L'émotion de joie est dominante dans le corpus d'entraînement, tandis que les émotions de tristesse et de dégoût sont sous-représentées. Les sous-représentations de certaines étiquettes font que le modèle a moins d'exemples pour apprendre à les identifier correctement.

3.4 Synthèse des données utilisées

La première phase de notre travail nous a amenées à constituer quatre corpus (cf. Table 4), utilisés à différentes phases dans les expériences que nous présentons dans la section suivante. Nous avons

1. https ://deft.lisn.upsaclay.fr/2015/
2. https ://deft.lisn.upsaclay.fr/2015/guideAnnotation.fr.php ?lang=fr

Étiquettes DEFT 2015	Étiquettes CoarseDEFT	Proportions
Amour, Apaisement, Plaisir, Satisfaction, Valorisation	Joie	52,1%
Colère, Ennui, Insatisfaction, Dévalorisation, Mépris	Colère	25,7%
Accord, Désaccord	Neutre	11,7%
Peur	Angoisse	8,5%
Tristesse, Déplaisir	Tristesse	1,6%
Dérangement	Dégoût	0,4%

TABLE 3 – Adaptation de la typologie des émotions du corpus DEFT 2015 vers la nouvelle typologie utilisée dans le corpus CoarseDEFT

divisé le corpus CoarseDEFT en différentes parties distinctes. L'entraînement initial de notre modèle est réalisé à partir d'une partie de corpus CoarseDEFT (70%). L'évaluation de sa performance, tout au long des expériences, s'effectue sur deux corpus : un échantillon du corpus CoarseDEFT (15%) et notre corpus de référence INTERCITÉS (cf. section 3.2). La seconde phase de notre travail a consisté à enrichir notre corpus d'entraînement pour améliorer les performances de notre modèle (cf. section 4.2). Pour cela, nous avons construit de manière itérative un autre corpus de 863 486 tweets, extraits de Twitter/X avec le mot-clé «SNCF».

Corpus	Nombre de tweets	Sources de tweets
Train (avant l'enrichissement)	3 002	CoarseDEFT
Val (avant l'enrichissement)	659	CoarseDEFT
Test (avant l'enrichissement)	659	CoarseDEFT
Corpus de référence	249	Corpus INTERCITÉS
Corpus d'enrichissement 1	10 776	Corpus INTERCITÉS
Corpus d'enrichissement 2	863 486	Corpus SNCF

TABLE 4 – Corpus utilisés pour l'apprentissage

4 Méthodologie

La démarche mise en place vise à spécialiser notre modèle de prédiction en fonction de notre corpus d'étude, permettant ainsi une classification multi-étiquettes efficace et précise sur les données ferroviaires. Notre objectif est donc d'entraîner un classifieur multi-étiquettes sur le corpus d'entraînement (CoarseDEFT) annoté selon différentes émotions, puis de le spécialiser sur nos données. Pour cela, nous avons mis en place une méthode itérative, qui enrichit progressivement le corpus d'entraînement avec davantage de données (cf. schéma 1). Le modèle s'entraîne sur le corpus d'entraînement et après chaque phase d'apprentissage, nous évaluons sa performance sur notre corpus de référence (cf. section 3.2). Plus précisément, nous effectuons une analyse manuelle des prédictions pour observer le comportement du modèle. Si l'analyse indique une pertinence insuffisante, nous intégrons de nouvelles données issues de nos corpus d'enrichissement (cf.Table 4) au corpus d'entraînement en fonction de cette analyse. Les itérations se poursuivent ensuite jusqu'à atteindre une performance satisfaisante.

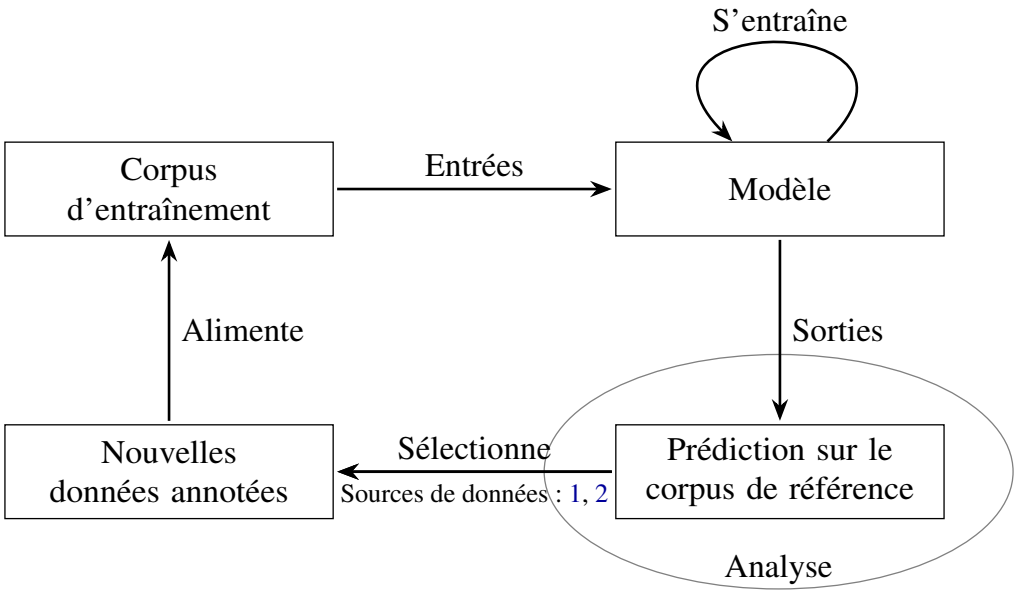


FIGURE 1 – Chaîne de traitement de la spécialisation du modèle

4.1 Configuration du modèle dédié à la classification des émotions en multi-étiquettes

Notre modèle repose sur l'affinage des deux dernières couches du modèle de langue CamemBERT (Martin *et al.*, 2020), architecture basée sur les transformers largement utilisés pour les tâches de classification de textes en français (Lincker *et al.*, 2023; Etienne, 2023). Cette représentation est enrichie par l'ajout de six couches d'encodeurs transformers, chacune avec 8 têtes d'attention. Notre structure des transformers s'inspire de l'approche explorée par Blivet *et al.* (2023) dans le cadre de leur participation au défi DEFT 2023 [3] qui traite de la classification multi-étiquettes.

Dans un modèle de classification multi-étiquettes, chaque label est traité comme une entité distincte. Si l'on considère un ensemble de N labels, le modèle génère N scores de probabilité $p_1, p_2, \ldots, p_N$ indépendants, chacun correspondant à une étiquette spécifique. Ces scores sont obtenus par la fonction d'activation *sigmoid* appliquée à la sortie du modèle pour chaque étiquette. Un seuil de décision θ est appliqué à chaque probabilité prédite pour déterminer la présence ou l'absence d'une étiquette. Si la probabilité prédite pour une étiquette donnée i, p_i est supérieure ou égale au seuil θ , alors l'étiquette i est assignée à l'instance ; sinon, elle est rejetée. La valeur de θ est généralement fixée à 0,5 dans de nombreux cas par défaut, mais cette valeur peut être ajustée pour répondre à des critères de performance spécifiques. Le choix de θ implique un équilibre entre la sensibilité (taux de vrais positifs) et la spécificité (taux de vrais négatifs) du modèle. Dans le cadre de notre étude, nous définissons le seuil à 0,35 au début de nos expériences, puisqu'initialement le modèle n'est pas assez sensible pour générer des probabilités plus élevées permettant des prédictions fiables. Cependant, à mesure que la performance du modèle s'améliore, nous augmentons progressivement ce seuil jusqu'à atteindre 0,5, comme le montre la Table 6.

3. https ://deft2023.univ-avignon.fr/

4.2 Spécialisation du modèle par enrichissement du corpus d'entraînement

Les prédictions sur le corpus de référence sont analysées pour évaluer la sensibilité du modèle. Les premières analyses indiquent un faible rappel, signifiant que le modèle ne généralise pas correctement sur les données INTERCITÉS. Bien qu'il identifie mieux les émotions fréquentes (joie, colère, neutre), le modèle génère de nombreux faux positifs, suggérant une insuffisance de données pour les émotions sous-représentées comme la tristesse et le dégoût.

Pour augmenter la performance du modèle sur le corpus d'étude, nous avons adopté les stratégies suivantes pour enrichir le corpus d'entraînement :

1. Adapter le contenu du corpus d'entraînement pour le rendre plus similaire à celui du corpus d'étude, en utilisant des données extraites du corpus INTERCITÉS. Cela permet de mieux aligner notre modèle aux spécificités des tweets d'INTERCITÉS.
2. Améliorer l'équilibre des émotions représentées dans le corpus d'entraînement, en ciblant spécifiquement les émotions sous-représentées telles que la tristesse et le dégoût. Pour cela, nous avons extrait de nouveaux tweets en utilisant le mot-clé « SNCF » et sélectionné ceux qui contiennent des termes associés à la tristesse et au dégoût, identifiés grâce au dictionnaire EmoLex (Mohammad & Turney, 2013).

L'enrichissement s'est fait selon plusieurs itérations dont le détail est présenté avec les informations précises concernant la répartition des émotions se trouve dans l'annexe A.

5 Évaluation et résultats

Afin de valider la performance des modèles, nous prenons en compte deux mesures d'évaluation : le F1-Micro, le F1-Macro et l'exactitude, fréquemment utilisées pour les tâches de classification.

Méthodes	**Exactitude**	**F1-Micro**	**F1-Macro**
SVM	**0,77**	0,18	0,07
Random Forest	**0,77**	0,20	0,10
Régression logistique	0.74	0,25	0,17
Naive Bayes	0.65	**0,36**	**0,29**
Arbre de décision	0,71	0.28	0,20

TABLE 5 – Évaluation de méthodes de référence pour la définition d'une *baseline*

Avant de mettre en oeuvre notre méthodologie, nous construisons tout d'abord une *baseline*. Pour cela, nous évaluons les performances des méthodes classiques de classification automatique sur notre corpus (cf. Table 5). Ces approches manifestent des difficultés à équilibrer l'exactitude avec les scores F1-Micro et F1-Macro. Naive Bayes se distingue par ses performances supérieures en F1-Micro et en F1-Macro. L'écart global entre F1-Micro et F1-Macro indique que les classes majoritaires peuvent influencer le résultat du F1-Micro. Les scores obtenus des approches soulignent la complexité de la tâche de classification multi-étiquettes et confirment qu'il existe une marge significative pour l'amélioration des modèles.

Ensuite, nous avons réalisé quatre cycles d'enrichissement de notre corpus d'entraînement, suivis chacun par une phase de ré-entraînement du modèle. Après chaque cycle, nous avons évalué la performance du modèle à l'aide des métriques sélectionnées. L'évaluation s'effectue d'abord sur le corpus de référence composé de 249 tweets (cf. Figure 1). Les résultats obtenus après chaque session d'entraînement sont présentés dans la Table 6.

Corpus	**Taille du corpus d'entraînement**	**Exactitude**	**F1-Micro**	**F1-Macro**	**Seuil**
CoarseDEFT	3661 tweets	0,77	0,39	0,22	0,35
CoarseDEFT-V2	4361 tweets	0,814	0,51	0,36	0,45
CoarseDEFT-V3	5561 tweets	0,816	0,50	0,43	0,5
CoarseDEFT-V4	6201 tweets	**0.826**	**0.55**	**0,44**	0,5

TABLE 6 – Évolution des performances du modèle au cours des différentes itérations de l'entraînement

Les résultats de l'évaluation montrent d'abord que notre approche permet d'atteindre la *baseline* définie dès la première itération en exactitude et en F1 Micro et de la dépasser ensuite au fur et à mesure que le corpus est alimenté par les données provenant des deux méthodes que nous avons mentionnées dans la partie précédente. Nous observons des améliorations significatives de F1-Micro lors de l'ajout des données aux corpus CoarseDEFT-V2 et CoarseDEFT-V4, lesquelles concernent des tweets relatifs à INTERCITÉS. Par ailleurs, l'intégration de données correspondant aux émotions sous-représentées dans le corpus CoarseDEFT-V3 contribue à une amélioration du score F1-Macro. Les scores F1 pour chaque étiquette d'émotion, obtenus lors des différentes itérations, sont présentés dans l'annexe B. L'augmentation du score d'exactitude prouve d'une part l'amélioration de la spécialisation du modèle et confirme que les nouvelles données ajoutées n'introduisent pas de bruit dans les prédictions. L'augmentation des scores F1 Micro et Macro semble valider nos hypothèses : l'adaptation du corpus aux données ferroviaires et l'équilibrage de la distribution des émotions améliorent la précision du modèle.

Compte tenu de la complexité de l'annotation des émotions, confirmée par le taux d'accord inter-annotateurs obtenu lors de l'annotation des émotions dans le corpus de référence (cf. section 3.2), il est pertinent de se demander si les les faibles scores F1 Micro et Macro observés dans la Table 6 peuvent être attribué à des divergences dans l'interprétation de la typologie des émotions parmi les annotateurs.

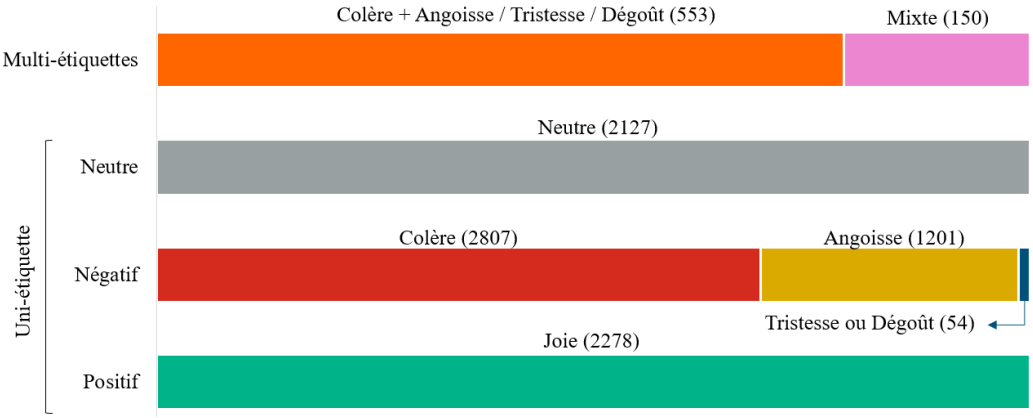


FIGURE 2 – Répartition des émotions dans le corpus INTERCITÉS

Bien que le modèle nécessite des améliorations, il affiche déjà une performance satisfaisante, en particulier en termes de spécificité. Appliqué à l'ensemble des tweets de notre corpus d'étude, il nous permet d'extraire des statistiques qui donnent une première indication de la répartition des émotions. D'après la Figure 2, les émotions les plus exprimées sont la colère, la joie, le neutre et l'angoisse. Les émotions négatives dominent dans les tweets. La sur-représentation des émotions négatives colère et angoisse peut certainement être relativisée et attribuée aux particularités de la plate-forme Twitter/X, sur laquelle les utilisateurs sont plus enclins à partager des commentaires et des messages négatifs.

6 Conclusion et discussion

Pour répondre aux enjeux commerciaux et industriels de notre projet, nous implémentons une chaîne de traitement qui entraîne un modèle de classification multi-étiquettes capable de détecter les émotions exprimées dans les tweets. Une partie importante de cette chaîne de traitement implique la constitution d'un corpus d'entraînement, annoté selon une typologie adaptée à notre étude. En l'occurrence, le corpus DEFT 2015 est repris, adapté à une nouvelle typologie d'émotions et enrichi en tweets émotionnellement marqués et traitant du domaine ferroviaire. Cette démarche permet au modèle de transférer ses acquis à notre corpus d'étude, lequel n'est pas annoté. Notre modèle, soit une combinaison du modèle CamemBERT et de transformers, dépasse d'emblée la baseline établie, et atteint un F1-Micro score de 0,55, un F1-Macro score de 0,44 et une exactitude de 0,826. Ce modèle se révèle être précis et efficace pour la classification des émotions dans les situations où une seule étiquette doit être prédite.

Dans le cadre de cette étude, nous avons fait face à plusieurs défis. Premièrement, l'annotation des émotions dans les tweets est une tâche subjective, qui complique la mise en place de critères d'annotation uniformes. Deuxièmement, notre corpus d'entraînement montre une répartition inégale entre les tweets uni-étiquettes, qui dominent, et les tweets multi-étiquettes. De surcroît, la distribution des différentes émotions reste déséquilibrée au sein du corpus, même après l'ajout de nouvelles données. Ces déséquilibres exercent une influence négative sur les performances du modèle. Cependant, il est essentiel de déterminer si ces inégalités sont propres et inhérentes à notre corpus INTERCITÉS. Au lieu d'essayer d'équilibrer artificiellement le corpus d'entraînement, il serait plus avisé d'adapter le modèle pour qu'il s'accommode de cette distribution inégale des étiquettes. Enfin, nous devons reconsidérer la nature des données utilisées lors de l'entraînement du modèle. Les données contenues dans le corpus d'entraînement (CoarseDEFT et ses versions successives), qui servent à l'apprentissage du modèle, se distinguent de celles du corpus de référence utilisé pour les prédictions et les évaluations. Le modèle présente des limitations dans le transfert des connaissances du corpus d'entraînement vers le corpus INTERCITÉS.

Malgré une sensibilité parfois limitée pour identifier l'intégralité des vrais positifs, les résultats permettent de dégager des tendances représentatives de la perception qu'ont les clients et les non-clients de l'offre INTERCITÉS et répondent aux besoins spécifiques de la la Direction Marketing. Pour aller au-delà de la classification multi-étiquettes des émotions, l'Analyse de Sentiments Basée sur les Aspects (ABSA) (Pontiki *et al.*, 2016) est une piste de recherche envisagée afin de relier les émotions aux thématiques évoquées dans les tweets. Cela permettra d'analyser plus précisément la manière dont sont perçus les différents aspects spécifiques de l'offre INTERCITÉS.

Références

ASLAM N., RUSTAM F., LEE E., WASHINGTON P. B. & ASHRAF I. (2022). Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble lstm-gru model. *Ieee Access*, **10**, 39313–39324. DOI : 10.1109/access.2022.3165621.

BAZIOTIS C., ATHANASIOU N., CHRONOPOULOU A., KOLOVOU A., PARASKEVOPOULOS G., ELLINAS N., NARAYANAN S. & POTAMIANOS A. (2018). Ntua-slp at semeval-2018 task 1 : Predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint arXiv :1804.06658*. DOI : 10.18653/v1/s18-1037.

BLIVET A., DEGRUTÈRE S., GENDRON B., RENAULT A., SIOUFFI C., GAUDRAY-BOUJU V., CERISARA C., FLAMEIN H., GUIBON G., LABEAU M. *et al.* (2023). Participation de l'équipe ttgv à deft 2023 : Réponse automatique à des qcm issus d'examens en pharmacie. *Actes de CORIA-TALN 2023. Actes du Défi Fouille de Textes@ TALN2023*, p. 23–38.

CAMACHO-COLLADOS J., REZAEE K., RIAHI T., USHIO A., LOUREIRO D., ANTYPAS D., BOISSON J., ESPINOSA-ANKE L., LIU F., MARTÍNEZ-CÁMARA E. *et al.* (2022). Tweetnlp : Cutting-edge natural language processing for social media. *arXiv preprint arXiv :2206.14774*. DOI : 10.18653/v1/2022.emnlp-demos.5.

CHOWDHURY M. S. M. & PAL B. (2023). Bert-based emotion classification approach with analysis of covid-19 pandemic tweets. In *Applied Informatics for Industry 4.0*, p. 109–121. Chapman and Hall/CRC. DOI : 10.1201/9781003256069-10.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), 37–46. DOI : 10.1177/001316446002000104.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv (Cornell University)*.

EKMAN P. (1992). An argument for basic emotions. *Cognition & emotion*, **6**(3-4), 169–200. DOI : 10.1080/02699939208411068.

ETIENNE A. (2023). *Analyse automatique des émotions dans les textes : contributions théoriques et applicatives dans le cadre de l'étude de la complexité des textes pour enfants*. Thèse de doctorat, Université de Nanterre-Paris X.

FRAISSE A. & PAROUBEK P. (2014). Toward a unifying model for opinion, sentiment and emotion information extraction. In *The 9th International Conference on Language Resources and Evaluation*, p. 3881–3886 : European Language Resources Association (ELRA).

GUIBON G., LABEAU M., FLAMEIN H., LEFEUVRE L. & CLAVEL C. (2021). Meta-learning for classifying previously unseen data source into previously unseen emotional categories. In *1st Workshop on Meta Learning and Its Applications to Natural Language Processing, ACL 2021*. DOI : 10.18653/v1/2021.metanlp-1.9.

HAMON T., FRAISSE A., PAROUBEK P., ZWEIGENBAUM P. & GROUIN C. (2015). Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de l'édition 2015 du défi fouille de texte (deft). In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2015)*.

HUANG C., TRABELSI A. & ZAÏANE O. R. (2019). Ana at semeval-2019 task 3 : Contextual emotion detection in conversations through hierarchical lstms and bert. *arXiv preprint arXiv :1904.00132*.

JAVED N. & MURALIDHARA B. (2022). Emotions during covid-19 : Lstm models for emotion detection in tweets. In *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications : ICMISC 2021*, p. 133–148 : Springer. DOI : 10.1007/978-981-16-6407-6_13.

KABIR M. Y. & MADRIA S. (2021). Emocov : Machine learning for emotion detection, analysis and visualization using covid-19 tweets. *Online Social Networks and Media*, **23**, 100135–100147. DOI : 10.1016/j.osnem.2021.100135.

KIM E. & KLINGER R. (2018). Who feels what and why ? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 1345–1359 : Association for Computational Linguistics.

LANDIS J. R. & KOCH G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174.

LINCKER E., GUINAUDEAU C., PONS O., DUPIRE J., BARBET I., HUDELOT C., MOUSSEAU V. & HURON C. (2023). Classification automatique de données déséquilibrées et bruitées : application aux exercices de manuels scolaires. In *18e Conférence en Recherche d'Information et Applications\\16e Rencontres Jeunes Chercheurs en RI\\30e Conférence sur le Traitement Automatique des Langues Naturelles\\25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, p. 121–130 : ATALA.

LIU Z., YANG K., ZHANG T., XIE Q., YU Z. & ANANIADOU S. (2024). Emollms : A series of emotional large language models and annotation tools for comprehensive affective analysis. *arXiv preprint arXiv :2401.08508*. DOI : 10.48550/arxiv.2401.08508.

LORA S. K., SAKIB N., ANTORA S. A. & JAHAN N. (2020). A comparative study to detect emotions from tweets analyzing machine learning and deep learning techniques. volume 12, p. 6–12.

LUO L. & WANG Y. (2019). Emotionx-hsu : Adopting pre-trained bert for emotion classification. *arXiv preprint arXiv :1907.09669*.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. *arXiv preprint arXiv :1911.03894*. DOI : 10.18653/v1/2020.acl-main.645.

MOHAMMAD S., BRAVO-MARQUEZ F., SALAMEH M. & KIRITCHENKO S. (2018). Semeval-2018 task 1 : Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, p. 1–17. DOI : 10.18653/v1/s18-1001.

MOHAMMAD S. M. & TURNEY P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, **29**(3), 436–465.

PAVEAU M.-A. (2017). *L'analyse du discours numérique. Dictionnaire des formes et des pratiques*. Hermann.

PLUTCHIK R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion*, p. 3–33. Elsevier. DOI : 10.1016/b978-0-12-558701-3.50007-7.

PONTIKI M., GALANIS D., PAPAGEORGIOU H., ANDROUTSOPOULOS I., MANANDHAR S., AL-SMADI M., AL-AYYOUB M., ZHAO Y., QIN B., DE CLERCQ O. *et al.* (2016). Semeval-2016 task 5 : Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, p. 19–30 : Association for Computational Linguistics. DOI : 10.18653/v1/s16-1002.

RABEYA T., FERDOUS S., ALI H. S. & CHAKRABORTY N. R. (2017). A survey on emotion detection : A lexicon based backtracking approach for detecting emotion from bengali text. In *2017 20th international conference of computer and information technology (ICCIT)*, p. 1–7 : IEEE. DOI : 10.1109/iccitechn.2017.8281855.

ROSENTHAL S., FARRA N. & NAKOV P. (2019). Semeval-2017 task 4 : Sentiment analysis in twitter. *arXiv preprint arXiv :1912.00741*. DOI : 10.18653/v1/s17-2088.

STRAPPARAVA C., VALITUTTI A. *et al.* (2004). Wordnet affect : an affective extension of wordnet. In *Lrec*, volume 4, p. 1083–1086 : Lisbon, Portugal.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention Is All You Need. arXiv :1706.03762 [cs].

WANG X., LI X., YIN Z., WU Y. & LIU J. (2023). Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*. DOI : 10.1177/18344909231213958.

YU J., MARUJO L., JIANG J., KARUTURI P. & BRENDEL W. (2018). Improving multi-label emotion classification via sentiment classification with dual attention transfer network. DOI : 10.18653/v1/d18-1137.

A Enrichissement du corpus d'entraînement

Corpus	Données	uni-étiquette	multi-étiquttes	Neutre	Joie	Tristesse	Colère	Angoisse	Dégoût
CoarseDEFT	**Train** : 3002 tweets	3002	0	352	1564	47	774	255	10
	Val : 659 tweets	659	0	73	347	9	170	57	3
CoarseDEFT-V2	**Train** : 3002 + 700 tweets d'Intercités	3484	218	484	1685	64	1042	404	21
	Val : 659 tweets	659	0	73	347	9	170	57	3
CoarseDEFT-V3	**Train** : 3002 + 700 tweets d'Intercité + 500 tweets «tristesse» + 500 tweets «dégoût»	3977	725	516	1781	190	1127	344	115
	Val : 659 + 100 tweets «tristesse» + 100 tweets «dégoût»	749	110	93	349	19	214	66	90
CoarseDEFT-V4	**Train** : 3002 + 1050 tweets d'Intercité + 800 tweets «tristesse» + 500 tweets «dégoût»	4553	799	599	1786	343	1275	435	115
	Val : 659 + 100 tweets «trsitesse» + 100 tweets «dégoût»	749	110	93	349	19	214	66	90

TABLE 7 – Évolution des données du modèle au cours des différentes itérations de l'entraînement

B Analyse des résultats par émotion

Corpus	F1 score par émotion						F1-Micro	F1-Macro
	Neutre	Joie	Tristesse	Colère	Angoisse	Dégoût		
CoarseDEFT	0.31	0.47	0	0.55	0	0	0.39	0.22
CoarseDEFT-V2	0.43	0.65	0	0.68	0.40	0	0.51	0.36
CoarseDEFT-V3	0.5	0.54	0	0.62	0.45	0.50	0.50	0.43
CoarseDEFT-V4	0.52	0.6	0	0.6	0.61	0.27	0.55	0.44

TABLE 8 – Évaluation de la classification par étiquette

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Approche multitâche pour l'amélioration de la fiabilité des systèmes de résumé automatique de conversation

Eunice Akani[1, 2] Benoit Favre[1] Frederic Bechet[1] Romain Gemignani[2]
(1) Aix-Marseille Univ, CNRS, LIS, Marseille, France
(2) Enedis, Marseille, France
prenom.nom@lis-lab.fr, romain.gemignani@enedis.fr

RÉSUMÉ

Le résumé de dialogue consiste à générer un résumé bref et cohérent d'une conversation ou d'un dialogue entre deux ou plusieurs locuteurs. Même si les modèles de langue les plus récents ont permis des progrès remarquables dans ce domaine, générer un résumé fidèle au dialogue de départ reste un défi car cela nécessite de prendre en compte l'interaction entre les locuteurs pour conserver les informations les plus pertinentes du dialogue. Nous nous plaçons dans le cadre des dialogues humain-humain avec but. Ce cadre nous permet d'intégrer des informations relatives à la tâche dans le cadre du résumé de dialogue afin d'aider le système à générer des résumés plus fidèles sémantiquement. Nous évaluons dans cette étude des approches multitâches permettant de lier la tâche de résumé à des tâches de compréhension du langage comme la détection de motifs d'appels. Les informations liées à la tâche nous permettent également de proposer des nouvelles méthodes de sélection de résumés basées sur l'analyse sémantique du dialogue ainsi que des métriques d'évaluation basées également sur cette même analyse. Nous avons testé ces méthodes sur DECODA, un corpus français de dialogue collecté dans le centre d'appel de la RATP entre des usagers et des téléconseillers. Nous montrons que l'ajout d'informations liées à la tâche augmente la fiabilité des résumés générés.

ABSTRACT

Multitask approaches for improving reliability in goal-oriented dialogue summarization

Dialogue summarization consists of generating a brief and coherent summary of a conversation or dialogue between two or more speakers. Although the most recent language models have led to remarkable progress in this field, generating a summary faithful to the original dialogue remains a challenge because it requires taking into account the interaction between the speakers to retain the most relevant information from the dialogue. In this study we will consider human-human goal-oriented dialogues. This framework allows us to integrate task-related information into the dialogue summary framework to assist the system in generating more semantically faithful summaries. In this study, we evaluate multitask approaches that link the summary task to language comprehension tasks such as calltype classification. Task-related information also enables us to propose new methods of summary selection based on semantic analysis of the dialogue as well as semantic-based evaluation metrics. We tested these methods on DECODA, a French dialogue corpus collected in the RATP call center between users and teleadvisors. We demonstrate that the addition of task-related information increases the reliability of the generated summaries.

MOTS-CLÉS : Résumé Dialogue orienté tâche, compréhension de la parole, approches multitâches.

KEYWORDS: Task-oriented dialog summarization, spoken language understanding, multitask approaches.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 338–351.

1 Introduction

La tâche de résumé automatique est une sous-tâche de la génération automatique de texte qui consiste à condenser un contenu pour en avoir les informations essentielles. Il existe deux types de résumés automatiques : le résumé par extraction qui consiste à extraire les informations importantes, les agencer pour en faire un résumé et le résumé par abstraction qui consiste à résumer un texte en employant de nouveaux termes ou des paraphrases. Bien que les systèmes pour le résumé par abstraction aient une meilleure représentation syntaxique et une bonne compréhension sémantique, un problème majeur demeure. Il s'agit de la fidélité du résumé généré par rapport au document source. En effet, les systèmes de l'état-de-l'art basés sur les modèles de langue pré-entraînés générèrent des résumés qui peuvent contenir des informations erronées n'étant pas présentes dans le document original (souvent appelées *hallucinations*) (Maynez *et al.*, 2020).

Le résumé de conversation, une extension du résumé automatique de texte sur les interactions orales ou écrites, n'échappe pas à ce problème majeur. En effet, le style du discours spontané des transcriptions ne correspond pas au style attendu des résumés. Ce qui implique donc l'utilisation d'approches abstractives qui sont significativement affectées par ces hallucinations. De plus, résumer une conversation implique de *comprendre* les interactions entre les participants afin de ne pas faire de contresens sur son contenu.

Une étude menée par Cao *et al.* (2018) révèle que 30% des résumés générés par divers systèmes de résumé de texte contenaient des informations erronées par rapport au document original, qualifiées d'« hallucination » par Maynez *et al.* (2020). Plusieurs approches ont été proposées pour évaluer la fidélité des résumés générés, telles que l'implication textuelle (Falke *et al.*, 2019; Maynez *et al.*, 2020; Luo *et al.*, 2023), la vérification des questions dérivées d'un résumé (Durmus *et al.*, 2020; Fabbri *et al.*, 2022), et l'analyse des entités nommées absentes dans la source (Nan *et al.*, 2021; Ji *et al.*, 2023). Pour les résumés de dialogues, les études sur l'hallucination demeurent moins nombreuses que pour les résumés de documents textuels. Wang *et al.* (2022) a constaté que 35% des résumés du jeu de données SAMSum (Gliwa *et al.*, 2019) ne sont pas cohérents vis-à-vis des dialogues sources. Tang *et al.* (2022) ont classé huit types d'erreurs factuelles dans les résumés de dialogue, cinq étant spécifiques à ceux-ci, tandis que Wang *et al.* (2022) en ont identifié six. Ils ont proposé un schéma d'évaluation au niveau du modèle pour évaluer la fidélité. Ils ont utilisé un modèle de résumé basé sur des probabilités de génération conditionnelles pour différencier les résumés positifs des résumés négatifs.

Le résumé de dialogue, une tâche récente, englobe divers types de conversations, notamment les résumés de réunions introduits en premier par des corpus comme AMI/ICSI (Carletta *et al.*, 2005; Janin *et al.*, 2003). Des approches neuronales ont été appliquées à ces corpus, mais leur efficacité est limitée par la structure des dialogues et la diversité des données d'entrée, incluant des conversations de service à la clientèle, des discussions informelles et techniques. Ainsi, des méthodes utilisant des informations auxiliaires, telles que les actes de dialogue (Goo & Chen, 2018) ou la terminologie du domaine (Koay *et al.*, 2020), ont été proposées. Des jeux de données plus récents, comme SAMSum (Gliwa *et al.*, 2019), offrent de nouveaux défis, notamment la modélisation des participants dans les dialogues.

Une tâche récente dans le domaine du résumé de conversation est le résumé de dialogue orienté tâche. Le dialogue orienté tâche fait référence à un type de conversation dont l'objectif principal est d'accomplir une tâche ou un objectif spécifique. Ces dialogues sont typiques des interactions avec un service clientèle, une assistance technique et d'autres scénarios similaires. Plusieurs corpus ont été

proposés pour la tâche entre autre TODSUM (Zhao *et al.*, 2021), DECODA (Bechet *et al.*, 2012).

Dans un dialogue orienté tâche, les participants interagissent spontanément pour résoudre un problème. Chacun joue un rôle spécifique, utilisant un langage naturel avec des hésitations. Les objectifs des participants, les procédures, entités nommées, etc. sont donc cruciaux pour caractériser l'objectif du dialogue. Ainsi, les résumés doivent refléter ces aspects pour rester fidèles à la conversation.

Dans ce papier, nous proposons d'évaluer et d'améliorer la fidélité du résumé de dialogue en utilisant des informations spécifiques à la tâche telle que le motif d'appel. Nous nous concentrons sur les centres d'appels et nous nous appuyons sur le corpus DECODA (Bechet *et al.*, 2012), qui est l'un des rares corpus de dialogue parlé humain-humain à grande échelle, enregistré dans des centres d'appels (dans des conditions réelles), avec des annotations en motif d'appel et d'entités nommées spécifiques au domaine. A notre connaissance, il n'y a pas de corpus de parole en anglais plus volumineux contenant ce type d'interactions (dialogue humain-humain avec but) annotés sémantiquement et possédant des résumés annotés.

Nos contributions sont les suivantes :

— Nous proposons d'utiliser des informations spécifiques à la tâche pour améliorer la fiabilité des modèles de résumé automatique en comparant plusieurs méthodes multitâches.
— Nous introduisons une mesure basée sur la prédiction de la distribution des motifs d'appel permettant de sélectionner des résumés à partir de critères sémantiques.
— Nous proposons une évaluation montrant l'amélioration de la qualité des résumés sur le corpus DECODA.

2 Guidage de la génération de résumés par les informations sémantiques

Dans ce chapitre, nous décrivons plusieurs méthodes permettant d'intégrer des informations spécifiques à la tâche, notamment les motifs d'appel et les entités nommées du domaine, pour la génération de résumés.

Informations sémantiques liées à la tâche Le résumé de la conversation orientée tâche implique plusieurs échanges portant sur des informations spécifiques ainsi que des instructions relatives à la tâche que les participants cherchent à accomplir. Les informations dans la conversation dépendent ainsi de l'objectif à atteindre. Le type de représentation sémantique permettant d'encoder ces informations a été particulièrement étudié dans le cadre du dialogue humain-machine avec les modèles de Compréhension Automatique de la Parole (Spoken Language Understanding) développés pour des tâches de réservation de transport (ex : corpus ATIS) ou encore de restaurants ou hotels (ex : corpus MEDIA). Dans ce type d'étude on défini généralement 3 niveaux (Lee *et al.*, 2018) sémantiques :

— **domaine** : le domaine représente le cadre sémantique dans lequel se déroule le dialogue. Par exemple la réservation de billets d'avion pour le corpus ATIS. Dans notre étude il s'agit du domaine des transports publics dans Paris pour le corpus DECODA.
— **intention** : l'intention représente le type de requête pour un système de communication humain-machine, par exemple une confirmation, une demande de renseignement, etc. Elle est le plus souvent associée à un tour de parole, mais dans le corpus DECODA les étiquettes d'intention sont mises au niveau de l'ensemble du dialogue. Elles correspondent aux motifs

d'appels tels que *demande d'itinéraire* ou *réclamation objet perdu*.

— **paires concepts/valeurs** : ces paires correspondent aux arguments des relations sémantiques exprimées dans les intentions, comme par exemple la destination pour une demande d'itinéraire. Dans le corpus DECODA les entités nommées du domaine des transports parisiens sont annotées dans le corpus.

Dans cette étude, nous explorons comment introduire ces informations dans le processus de génération de résumés, comment elles peuvent être utilisées pour créer une mesure d'évaluation de la fidélité du résumé et comment leur utilisation influence la fidélité des résumés.

Approches pipeline et multi-tâches pour le résumé de conversation avec but Les informations spécifiques à une tâche ne sont pas directement disponibles dans les enregistrements ou les transcriptions de conversations, elles doivent donc être déduites. Nous pouvons, soit exploiter un système distinct pour prédire les catégories correspondantes et les utiliser comme entrée du système de résumé automatique, soit laisser le système les apprendre dans le cadre d'une supervision multitâche. Pour le moment nous nous sommes concentré sur la prédiction du motif d'appel à travers ces deux approches.

Soit C la conversation en entrée, R le résumé généré et M le motif d'appel. Nous considérons les deux méthodes suivantes :

1. **Multitache$_X$: Motif d'appel puis synopsis** : De manière générale, il s'agit de faire générer au modèle de langue une séquence composée de l'information sémantique X suivi le résumé. $\{X\}, R =$ résumé $\circ$ sémantique$C$. Pour notre cas particulier, nous avons considéré le motif d'appel. On a donc : $\{M\}, R =$ résumé $\circ$ sémantique(C).
2. **Pipeline$_X$** : Il s'agit de conditionner la génération du résumé par par une information sémantique X. On a donc $R =$ résumé$(X,$ information-sémantique$(M))$. Ici nous utilisons le motif d'appel comme information sémantique. Ainsi on obtient : $R =$ summary$(C,$ information-sémantique$(M))$. Ce motif en entrée peut être issu des motifs de référence (expérience *oracle*) ou encore d'un classifieur en motif d'appel.

Dans nos expériences un modèle de langue a été affiné sur une tâche de résumé automatique correspondant à chaque scénario. En complément de l'utilisation des informations sémantiques directement dans le processus de génération de résumé, nous proposons également de les utiliser en sortie du système de génération afin de sélectionner le résumé le plus fiable sémantiquement selon nos modèles.

3 Sélection de résumés sur des critères sémantiques

Il est possible de faire varier certains paramètres dans les processus de génération de texte par modèles de langue afin d'obtenir plusieurs sorties pour une même entrée. Dans cette étude, nous avons généré plusieurs résumés en utilisant 4 méthodes d'échantillonnage différentes pour sélectionner le prochain token comme dans l'article Akani *et al.*, 2023 : la recherche de type *beam-search* qui conserve les n-meilleurs chemins de probabilité les plus élevés à chaque étape ; l'échantillonnage de température qui consiste à redimensionner les logits avant d'appliquer la fonction softmax ; l'échantillonnage Top-K (Fan *et al.*, 2018) qui ne conserve que les K mots suivants les plus probables et redistribue la probabilité parmi ces K mots ; et l'échantillonnage Top-P (Holtzman *et al.*, 2020) qui consiste, étant donné une probabilité p, à prendre le plus petit ensemble possible de mots suivants dont la probabilité cumulative dépasse une masse de probabilité donnée et redistribue la probabilité parmi eux.

A partir de cet ensemble de résumés possibles, nous proposons deux méthodes de sélection basées sur les informations sémantiques liées à la tâche, l'une utilisant le motif d'appel et l'autre les entités nommées liées à la tâche.

Sélection de résumés basés sur la distribution des motifs d'appel. Le motif d'appel d'une conversation peut être vu comme une signature sémantique de l'objectif visé par la conversation. Ainsi, pour un résumé généré à partir d'une conversation, préserver le motif d'appel signifie préserver les caractéristiques sémantiques les plus importantes de cette conversation. Nous avons émis l'hypothèse qu'un résumé généré produisant un motif d'appel différent du motif d'appel de référence de la conversation contiendrait davantage d'informations incorrectes.

Pour utiliser ce critère pour choisir un résumé parmi un ensemble d'hypothèses, il est possible d'entraîner un classifieur de motifs d'appel pour prédire, à partir d'un résumé généré, son motif d'appel et de le comparer au motif d'appel qui pourrait être prédit par un autre classifieur sur la conversation entière.

Ici se pose un problème, certaines conversations peuvent avoir plusieurs motifs d'appel. En effet, un locuteur peut appeler pour une raison particulière et faire une autre demande dans la même conversation. Cela crée une frontière ambiguë entre les différents motifs d'appel. Pour traiter ce problème, nous proposons d'utiliser la divergence de Kullback-Leibler (KL) (Kullback & Leibler, 1951) sur la distribution de probabilité des motifs d'appel fournie par le classifieur, afin d'évaluer les résumés générés. La divergence de KL est une mesure de distance statistique qui quantifie la dissemblance entre deux distributions de probabilités. Elle évalue la différence entre une distribution de probabilité et une distribution de référence.

Ici, nous avons utilisé la divergence de KL pour évaluer un résumé généré sur la base de la distribution de probabilité sur les motifs d'appel. Cela est possible par l'utilisation de deux classifieurs en motifs d'appel : le premier appris sur les conversations entières et qui fournit une distribution de probabilités de motifs d'appels pour toute la conversation, et le second qui est appris directement sur les résumés de références et qui sera utilisé pour obtenir la distribution de motifs des résumés à sélectionner.

Pour $1...n$ les motifs d'appel, $G = \{g_1, ..., g_n\}$ la distribution de probabilité du résumé généré et $R = \{r_1, ..., r_n\}$ celle de la conversation entière, la KL divergence entre G et R se définit comme suit :

$$D_{\mathrm{KL}}(G \parallel R) = \sum_{x \in \{1..n\}} G(x) \log \left(\frac{G(x)}{R(x)} \right) \tag{1}$$

Il est maintenant possible de sélectionner le résumé qui minimise la distance D_{KL} parmi l'ensemble des résumés générés.

Sélection de résumés basée sur le risque d'hallucination. En complément du motif d'appel comme critère de sélection, nous utilisons le NEHR Akani *et al.*, 2023 qui correspond au pourcentage d'entités dans le résumé qui ne se trouvent pas dans le document source (la conversation dans notre cas). Il permet d'évaluer le risque d'hallucination sur les entités nommées. Lorsqu'un système de génération de résumés produit une entité nommée qui n'appartient pas au document original, cela augmente le risque d'*hallucination* de la part de ce modèle. En effet il peut s'agir d'une entité valide,

variante d'une des entités du document, mais il peut également s'agir d'une erreur de sur-génération de la part du modèle Akani, 2023.

Pour augmenter la fidélité du résumé, nous proposons de combiner au critère sur les motifs d'appels D_{KL}, la minimisation explicite du risque *NEHR*. Nous appliquons la règle de sélection suivante du résumé : $\hat{s}$. Soit H l'ensemble des résumés issus de l'échantillonnage pour la conversation C, V l'ensemble des résumés avec le NEHR minimum m, $D_{KL}(x, C)$ la divergence KL entre la résumé généré x et et la conversation C, et $\hat{s}$ la sortie finale :

$$m = \min_{x \in H} NEHR(x) \quad \text{and} \quad V = \{x \in H | NEHR(x) = m\}$$

$$\hat{s} = \min_{x \in V} D_{\mathrm{KL}}(x \parallel C) \tag{2}$$

4 Expériences et Résultats

4.1 Contexte expérimental : Jeu de données et modèles utilisés

Le corpus DECODA Bechet *et al.*, 2012 contient pour chaque transcription de conversation un résumé entre plusieurs agents du service client de la RATP et un usager. Ce sont des résumés très courts appelés synopsis. Les synopsis présentent les principaux évènements de la conversation tels que les objectifs des participants, le processus de résolution. Ce corpus se compose de 3 parties qui représentent les années d'enregistrement des conversations et ont des particularités (Voir Annexe A.1 pour la différence entre les deux parties). Pour notre part, nous n'avons utilisé que deux (DECODA-1, DECODA-3) des trois parties qui ont des synopsis associés créés et écrits par des annotateurs humains avec deux types de consignes. Tandis que les synopsis de DECODA-3 sont plus longs, écrits avec plus de détails, ceux de DECODA-1 sont synthétiques et écris en français abrégé. Le corpus a été aussi annoté en termes d'entités nommées spécifiques au domaine, avec un étiquetage morpho-syntaxique, avec des lemmes et des dépendances syntaxiques. Pour nos expériences, nous avons fusionné les deux parties. Les conversations issues du corpus DECODA couvrent une variété de motif d'appel notamment : *Information Trafic*, *Itinéraire*, *Objets perdus/trouvés*, *Abonnement*, *Horaires*, *Tickets* (Trione, 2014). La distribution de chaque motif d'appel est donné en Annexe A.2. De plus DECODA possède des entités nommées appartenant à une ontologie du domaine. On peut notamment citer *Produit, Transport, Horaire*, etc. Ces différentes annotations ainsi que le fait qu'il soit un corpus enregistré dans des conditions réelles ont motivé notre choix.

Dans le cadre de nos expériences, nous avons divisé le jeu de données en trois partitions (train pour l'entraînement des modèles, val pour le développement et test pour l'évaluation). Afin de s'assurer que chaque motif apparaît dans chaque partitions, une stratification des données en suivant le motif d'appel a été appliquée lors de la division. Nous avons calculé quelques statistiques sur les données et les avons consignées dans le tableau 1.

Stats	Train	Val.	Test
# examples	717	99	140
# conv. mots	486.3	487.9	465.4
# synopsis mots	28.8	30.3	29.5
# conv. tokens	608.2	604.7	579.8
# synopsis tokens	35.9	37.5	36.2

TABLE 1 – Distribution du jeu de données DECODA

Résumé automatique Pour l'entraînement des systèmes pour le résumé automatique, nous avons utilisé BARThez (Kamal Eddine *et al.*, 2021), un modèle séquence-à-séquence pré-entraîné sur la partie française de CommonCrawl, Wikipédia, NewsCrawl et d'autres corpus disponibles en français. Il a été introduit pour la

Conseiller : euh NNAAMMEE bonjour *Appelant :* oui bonjour je voulais savoir de la Madeleine quel bus je dois prendre pour me rendre au Saint-Philippe+du+Roule *Conseiller :* alors Madeleine Saint-Philippe+du+Roule *Appelant :* oui *Conseiller :* je recherche un instant *Appelant :* merci *Conseiller :* le cinquante-deux en direction du parc de Saint-Cloud *Appelant :* alors le cinquante-deux je le prends où ah+ben oui *Appelant :* d'accord *Conseiller :* Madeleine+Vignon *Appelant :* d'accord *Appelant :* donc cinquante-deux direction Porte+de+Saint-Cloud *Conseiller :* parc de Saint-Cloud *Appelant :* parc de Saint-Cloud et il descend à Saint-Philippe+du+Roule *Conseiller :* tout+à+fait *Appelant :* merci *Appelant :* au monsieur merci au+revoir *Conseiller :* bonne journée au+revoir	***REFERENCE SUMMARY*** Un appelant demande quel bus prendre pour se rendre de La Madeleine à Saint-Philippe-du-Roule. Le conseiller lui indique de prendre le bus 52 en direction du parc de Saint-Cloud, qui s'arrête à Saint-Philippe-du-Roule. ***CT-syn REFERENCE SUMMARY*** [MOTIF] ITNR [SUMMARY] Un appelant demande quel bus prendre pour se rendre de La Madeleine à Saint-Philippe-du-Roule. Le conseiller lui indique de prendre le bus 52 en direction du parc de Saint-Cloud, qui s'arrête à Saint-Philippe-du-Roule.

TABLE 2 – Un exemple issu du jeu de données DECODA

tâche de résumé automatique de texte. BARThez est un modèle composé de couches de Transformers (Vaswani *et al.*, 2017) et est basé sur l'architecture de BART (Lewis *et al.*, 2020). Il existe deux versions de l'architecture BARThez, une version de base et une large mBARThez. Pour notre part, nous avons utilisé la version de base qui possède 6 couches d'encodeur et 6 de decodeur comme BART base. Pour l'entraînement des modèles, nous avons utilisé le modèle pré-entraîné fourni par les auteurs et disponible sur la librairie Transformers [1] de Hugging Face.

Dans le tableau 1, nous avons consigné le nombre moyen de tokens de chaque conversation ainsi que chaque synopsis en utilisant le tokeniseur de BARThez. En ce qui concerne les hyper-paramètres, nous avons utilisé un *learning rate* de 5×10^{-5} pour l'optimiseur AdamW et avons fixé la taille maximale pour les conversations à 1024 et ceux des synopsis à 128. Chaque modèle a été entraîné sur 12 époques en ne sauvegardant que celle qui minimise la perte sur le jeu de validation. Nous avons modifié les synopsis du jeu de données en y ajoutant les motifs pour être dans les conditions pour entraîner le modèle Multitache$_M$. Un exemple d'illustration des modifications apportées au résumé de référence est dans le Tableau 2.

Pour le modèle Pipeline$_M$, les motifs associés à chaque conversation sont nécessaires en entrée, ce qui implique de les avoir préalablement. Nous avons utilisé un classifieur de motifs d'appel entraîné via une validation croisée (k-fold) pour prédire les motifs sur 25% des données non utilisées lors de l'entraînement initial (75%). Ce processus a été répété quatre fois pour obtenir les prédictions sur l'ensemble d'entraînement, puis utilisé pour l'entraînement du système de résumé automatique. Les motifs et les conversations ont été séparés par un marqueur afin d'aider le modèle à tenir compte des motifs lors de la prédiction des résumés.

Classification des motifs d'appel Nous avons entraîné un modèle CamemBERT-base (Martin *et al.*, 2020) pour la tâche de classification de séquences pour prédire les motifs d'appel. Pour ce faire, deux classifieurs ont été utilisés. L'un appliqué aux conversations *Conv-M-classif* et l'autre aux synopsis *Syn-M-classif*. Le classifieur *Conv-M-classif* prend en entrée la conversation, et prédit le motif d'appel. Ainsi, le nombre maximum de tokens pris en entrée est 512. Pour *Syn-M-classif*, on part du synopsis pour prédire le motif d'appel. La longueur maximale de l'entrée pour ce modèle est de 128. Chaque modèle a été entraîné sur 13 époques avec un batch size de 8. Le modèle ayant la loss minimale sur les données de validation a été conservé.

Reconnaissance des entités nommées du domaine DECODA possède 14 entités nommées spécifiques à un domaine, parmi lesquelles le numéro de téléphone, le prix, le type de produit, le type de transport, etc. Nous avons entraîné CamemBERT-base (Martin *et al.*, 2020) pour la tâche NER et avons obtenu un micro F1 et un macro F1 de 0,93 et 0,84 respectivement en utilisant la bibliothèque Seqeval [2]. Les entités détectées sont utilisées dans les métriques d'évaluation, telles que NEHR.

1. https://huggingface.co/moussaKam/barthez
2. https://github.com/chakki-works/seqeval

4.2 Résultats de l'évaluation

Classification des motifs d'appel Après avoir entraîné les deux modèles de classification en motifs d'appel, les résultats sont consignés dans le tableau 3. Les données étant déséquilibrées, nous avons calculé l'accuracy et le score F1 de chaque modèle. Le tableau 3 montre que la classification effectuée sur la conversation donne de meilleurs résultats que celle effectuée sur les synopsis. Ce résultat est probablement dû à la quantité d'informations supplémentaires disponibles dans la conversation qui révèlent la raison de l'appel. Le modèle Multitache$_M$ qui prédit les entités en même temps que le résumé donne des résultats significativement plus faibles que les classifieurs.

Système	Acc.	F1	W-F1
Conv-M-classif	86	65	86
Syn-M-classif	83	54	81
Multitache$_M$	79	50	76

TABLE 3 – Classification des motifs d'appel en utilisant un classifieur sur les conversations (Conv-M-classif), les synopsis (Syn-M-classif) et le modèle Multitache$_M$ générant des motifs. F1 et W-F1 sont respectivement le F1 score en macro moyenne et celui en moyenne pondérée.

Résumé automatique basé sur les motifs d'appel Dans la section 2, nous avons décrit le processus pour entraîner les différents modèles : Multitache$_M$, Pipeline$_M$. Multitache$_M$ est conçu pour une utilisation multitâche (génération du motif d'appel suivi du résumé). Il faut donc modifier les données en sortie du système pour les faire correspondre à la sortie attendue. Ainsi, nous avons ajouté le motif d'appel de référence au résumé pendant l'entraînement (voir Table 2). Pour Pipeline$_M$, nous avons construit trois différents modèles qui se distinguent par le motif d'appel donné en entrée au côté de la conversation lors de l'entraînement et de la prédiction. Le premier est un système d'oracle (Pipeline$_M$ − oracle). En effet, en entrée du modèle, nous donnons le motif d'appel de référence, que ce soit pour l'entraînement ou pour le test. Le second est Pipeline$_M$ − oracle/pred ; nous supposons avoir accès au motif d'appel de référence pendant la phase d'entraînement, mais pas pendant la phase de test.

Par conséquent, en utilisant le classifieur (Conv-M-classif), nous avons prédit les motifs d'appel pour chaque exemple du jeu de test. Ces prédictions ont ensuite été combinées à la conversation et transmises au modèle pour la génération de résumé. Pour le dernier modèle (Pipeline$_M$), nous supposons que nous n'avons pas accès aux motifs d'appel de référence de chaque conversation pendant l'entraînement. Ainsi, nous avons donc dû les prédire. Pour la phase de test, nous avons procédé de la même manière que pour le modèle (Pipeline$_M$ − oracle/pred).

Nous avons comparé ces systèmes avec un modèle baseline qui génère un synopsis à partir de la conversation uniquement en entrée.

System	R1	R2	RL
Baseline	34.04	14.91	28.83
Multitache$_M$	33.59	14.23	28.42
Pipeline$_M$ − oracle	34.65	15.10	29.33
Pipeline$_M$ − oracle/pred	34.46	14.93	29.09
Pipeline$_M$	34.09	14.79	28.99

TABLE 4 – R1, R2, RL : ROUGE-Score des différents systèmes.

Métrique	Système	1st
Fidélité	Baseline	38.9
	Multitache$_M$	16.7
	Pipeline$_M$	44.4
Informativité	Baseline	33.3
	Multitache$_M$	23.8
	Pipeline$_M$	42.9

TABLE 5 – Évaluation manuelle : 1st - Nombre de fois (en pourcentage) où chaque système a été classé premier en termes de fidélité et d'informativité.

— **Évaluation automatique** Pour évaluer les différents systèmes, nous avons calculé ROUGE (Lin, 2004)

afin de tester la capacité du modèle à générer des n-grammes proches du résumé de référence. Le tableau 4 montre les résultats obtenus. Les différents systèmes donnent des résultats similaires en termes de ROUGE. Comme Zhou *et al.* (2022) montrent que les résumés générés qu'ils ont obtenus de DECODA contiennent des hallucinations, des omissions et des erreurs grammaticales, nous évaluons manuellement le niveau de fidélité des résumés.

— **Évaluation manuelle** Nous évaluons manuellement 30 résumés de systèmes en fonction de leur fidélité et de leur caractère informatif. L'idée est de choisir entre les trois systèmes les plus fidèles et les plus informatifs. Un système est dit fidèle à la conversation s'il ne contient aucune information la contredisant. Un système est informatif s'il couvre les informations les plus essentielles de la conversation. Ceci peut être mesuré en comparant le résumé d'un système avec le résumé de référence correspondant, qui devrait contenir les informations les plus essentielles de la conversation. Les résultats sont présentés dans le tableau 5. Nous observons que Pipeline$_M$ est plus souvent considéré comme le meilleur résultat, tant en termes de fidélité que d'informativité. Il est à noter que Multitache$_M$ est moins bon que la ligne de base, probablement en raison de la difficulté du système à générer des motifs d'appel.

Sélection de résumé Dans le but de maximiser la similarité en termes de motif d'appel entre le résumé prédit et la référence, nous avons généré une cohorte de résumés et effectué une sélection afin de minimiser la KL CONV. Nous n'avons utilisé que le modèle Baseline et le *Pipeline*$_M$ pour cette section. En moyenne 40 résumés ont été générés pour chaque système, et le meilleur en fonction de deux critères de sélection a été choisi. Le premier critère consiste à choisir le résumé présentant la divergence KL minimale entre la distribution de probabilité de la conversation et la distribution de probabilité du résumé généré ($\min D_{\text{KL}}(G \parallel \text{Conv})$). Le second critère est basé sur les équations 2 décrites dans la section 3. D'abord une sélection de l'ensemble des résumés ayant le NEHR minimum, puis, parmi eux, le résumé minimisant la $\min D_{\text{KL}}(G \parallel \text{Conv})$ ($\min \text{NEHR} + D_{\text{KL}}(G \parallel Conv)$).

Pour l'évaluation des résumés sémantiquement, nous avons introduit quatre métriques basées sur les informations sémantiques de la conversation tels que le motif d'appel et les entités nommées.

— **CT-Acc** : Il s'agit de la mesure de l'accuracy du classifieur en motif d'appel appliqué aux résumés générés par rapport aux motifs de référence. Nous considérons que plus l'accuracy est haute, plus le résumé contient d'éléments cohérents avec le motif d'appel que le classifieur utilise pour prédire la bonne étiquette.

— **NE-R/NE-P/NE-F1** : Nan *et al.*, 2021 a introduit des métriques qui comparent les entités nommées des résumés générés à ceux des résumés de référence. Basé sur leur metrique, nous mesurons la précision, le rappel et le F1 concernant les entités nommées détectées dans le résumé automatique par rapport aux entités dans le résumé de référence. La différence fondamentale avec leur proposition réside dans ce qui est considéré comme étant une entité nommée. Pour notre part, nous intégrons les quantités, les nombres, dans la définition des entités nommées et basons nos métriques sur les entités nommées du domaine à notre étude. Nous considérons que plus ces valeurs sont élevés, moins grand est le risque d'hallucination de la part du modèle.

Les métriques étant définies, nous nous attendons à ce que la minimisation de la divergence KL ($\min D_{\text{KL}}(G \parallel \text{Conv})$) augmente le **CT-Acc** et que le minimisation du NEHR augmente le rappel et la précision sur les entités nommées (NE-R et NE-P). Ainsi, la combinaison des deux métriques devrait nous permettre d'augmenter aussi bien les mesures sur les entités que l'accuracy sur les motifs d'appel.

Le tableau 6 présente les résultats obtenus après la sélection du résumé, tout d'abord selon le score ROUGE par rapport aux résumés de référence, puis par les mesures sémantiques définies précédemment.

Pour chaque modèle, nous indiquons également le score obtenu par le modèle sans sélection prenant juste la meilleure hypothèse (BEAM). Nous avons aussi rajouté le résumé qui permet d'avoir le NEHR minimim en suivant le papier Akani *et al.*, 2023.

Le tableau 6 montrent de faibles variations en terme de score ROUGE entre les différents modèles. Par contre les scores sémantiques évoluent fortement : la minimisation explicite de la divergence sur les motifs d'appels augmentent très significativement la mesure **CT-Acc**, particulièrement pour le modèle *Pipeline*$_M$ et en combinant

la minimisation du risque sur les entités et la divergence sur les motifs, nous montrons une augmentation significative de la précision sur les entités en conservant un gain important en accuracy sur les motifs d'appels et une dégradation légère en terme de rappel.

	R1 ↑	**R2** ↑	**RL** ↑	**CT-Acc** ↑	**NE-R** ↑	**NE-P** ↑	**NE-F1** ↑
Baseline model - BARThez finetuné pour générer les synopsis							
BEAM	34.04	14.91	28.83	76.4	0.46	0.57	0.51
min NEHR	34.12	14.13	27.79	73.6	**0.50**	0.58	**0.54**
min $D_{\text{KL}}(G \parallel \text{Conv})$	33.99	14.23	28.13	**82.9**	0.43	0.54	0.48
min NEHR + $D_{\text{KL}}(G \parallel Conv)$	33.70	14.57	28.04	82.1	0.45	**0.60**	0.51
Pipeline$_M$ - baseline + motif prédit en entrée, le synopsis en sortie							
BEAM	34.09	14.79	28.99	76.4	**0.47**	0.49	0.48
min NEHR	34.18	14.69	28.52	75.7	**0.47**	0.55	**0.51**
min $D_{\text{KL}}(G \parallel \text{Conv})$	33.85	14.60	28.87	**85.7**	0.42	0.53	0.47
min NEHR + $D_{\text{KL}}(G \parallel Conv)$	33.32	14.48	28.58	84.3	0.45	**0.57**	0.50

TABLE 6 – R1,R2,RL : ROUGE ; CT-Acc : Prédiction des motifs depuis les synopsis (ref=oracle synopsis) ; NE-R et NE-P : Le rappel et la précision des entités des résumés générés par rapport à la référence

5 Discussion et Conclusion

Nous avons examiné l'impact de deux modèles basés sur le motif d'appel pour améliorer la représentation sémantique des résumés générés. Bien que les métriques d'évaluation automatique comme ROUGE montrent des résultats similaires à la baseline, une évaluation humaine révèle que les résumés Pipeline$_M$ ont été choisi comme plus fidèle et informative en comparaison aux systèmes Baseline et Multitache$_M$.

L'utilisation de KL divergence comme critère de sélection des résumés montre des promesses, avec une augmentation d'environ 10% de l'accuracy (CT-Acc) lorsque le modèle Syn-M-classif est appliqué aux résumés générés. Cependant, certains motifs d'appel n'ont pas été suffisamment représentés dans le jeu de données, suggérant la nécessité d'une étude plus approfondie pour améliorer la prédiction du motif et, par conséquent, la qualité du résumé associé.

La combinaison de la KL divergence sur la conversation et du NEHR comme critère de sélection améliore la précision du modèle sur les entités nommées par rapport à la référence. Cependant, une évaluation manuelle plus poussée est nécessaire pour confirmer l'efficacité de ce critère combiné en termes de fidélité et d'informativité. Étant donné que toutes les méta-données du corpus n'ont pas été exploitées, nous envisageons d'intégrer d'autres aspects, notamment la structure de la conversation, pour continuer à améliorer la fidélité des résumés générés.

Les expériences ont été menées sur le corpus DECODA en utilisant le modèle BARThez, pre-entraîné sur du français. Le choix du corpus a été motivé par sa richesse en terme d'annotations mais également parce qu'il est un corpus enregistré dans un cas réel ; il a donc des conversations naturelles. Notre méthodologie a été utilisée dans ce cas précis mais nous croyons que les conclusions seront les mêmes quelques soit le corpus ou le modèle qui sera utilisé.

De plus, on peut envisager d'utiliser le corpus X-RiSaWoz (Moradshahi *et al.*, 2023) pour nos expériences. L'inconvénient de ce corpus est qu'il n'est pas issu de conversations naturelles et qu'il ne possède pas de résumé associé à chaque conversation. Néanmoins, il s'agit d'un grand corpus qui peut être utiliser suivant notre méthodologie grâce à l'usage des LLMs pour la génération de résumés. C'est une piste que nous souhaiterons explorer.

Remerciements

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2021-AD011012525 attribuée par GENCI.

Références

AKANI E. (2023). Étude de la fidélité des entités dans les résumés par abstraction. In M. CANDITO, T. GERALD & J. G. MORENO, Éds., *Actes de CORIA-TALN 2023. Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL)*, p. 21–36, Paris, France : ATALA.

AKANI E., FAVRE B., BECHET F. & GEMIGNANI R. (2023). Reducing named entity hallucination risk to ensure faithful summary generation. In C. M. KEET, H.-Y. LEE & S. ZARRIESS, Éds., *Proceedings of the 16th International Natural Language Generation Conference*, p. 437–442, Prague, Czechia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.inlg-main.33.

BECHET F., MAZA B., BIGOUROUX N., BAZILLON T., EL-BÈZE M., DE MORI R. & ARBILLOT E. (2012). DECODA : a call-centre human-human spoken conversation corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 1343–1347, Istanbul, Turkey : European Language Resources Association (ELRA).

CAO Z., WEI F., LI W. & LI S. (2018). Faithful to the original : Fact aware neural abstractive summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**(1). DOI : 10.1609/aaai.v32i1.11912.

CARLETTA J., ASHBY S., BOURBAN S., FLYNN M., GUILLEMOT M., HAIN T., KADLEC J., KARAISKOS V., KRAAIJ W., KRONENTHAL M., LATHOUD G., LINCOLN M., MASSON A. L., MCCOWAN I., POST W., REIDSMA D. & WELLNER P. D. (2005). The ami meeting corpus : A pre-announcement. In *Machine Learning for Multimodal Interaction*.

DURMUS E., HE H. & DIAB M. (2020). FEQA : A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the ACL*, p. 5055–5070, Online : ACL. DOI : 10.18653/v1/2020.acl-main.454.

FABBRI A., WU C.-S., LIU W. & XIONG C. (2022). QAFactEval : Improved QA-based factual consistency evaluation for summarization. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 2587–2601, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.187.

FALKE T., RIBEIRO L. F. R., UTAMA P. A., DAGAN I. & GUREVYCH I. (2019). Ranking generated summaries by correctness : An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the ACL*, p. 2214–2220, Florence, Italy : ACL. DOI : 10.18653/v1/P19-1213.

FAN A., LEWIS M. & DAUPHIN Y. (2018). Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the ACL (Volume 1 : Long Papers)*, p. 889–898, Melbourne, Australia : ACL. DOI : 10.18653/v1/P18-1082.

GLIWA B., MOCHOL I., BIESEK M. & WAWER A. (2019). SAMSum corpus : A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, p. 70–79, Hong Kong, China : ACL. DOI : 10.18653/v1/D19-5409.

GOO C.-W. & CHEN Y.-N. (2018). Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, p. 735–742. DOI : 10.1109/SLT.2018.8639531.

HOLTZMAN A., BUYS J., DU L., FORBES M. & CHOI Y. (2020). The curious case of neural text degeneration. In *International Conference on Learning Representations*.

JANIN A., BARON D., EDWARDS J., ELLIS D., GELBART D., MORGAN N., PESKIN B., PFAU T., SHRIBERG E., STOLCKE A. & WOOTERS C. (2003). The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, p. I–I. DOI : 10.1109/ICASSP.2003.1198793.

JI Z., LEE N., FRIESKE R., YU T., SU D., XU Y., ISHII E., BANG Y. J., MADOTTO A. & FUNG P. (2023). Survey of hallucination in natural language generation. **55**(12). DOI : 10.1145/3571730.

KAMAL EDDINE M., TIXIER A. & VAZIRGIANNIS M. (2021). BARThez : a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 9369–9390, Online and Punta Cana, Dominican Republic : ACL. DOI : 10.18653/v1/2021.emnlp-main.740.

KOAY J. J., ROUSTAI A., DAI X., BURNS D., KERRIGAN A. & LIU F. (2020). How domain terminology affects meeting summarization performance. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 5689–5695, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.499.

KULLBACK S. & LEIBLER R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, **22**(1), 79–86. Publisher : Institute of Mathematical Statistics.

LEE J., KIM D., SARIKAYA R. & KIM Y.-B. (2018). Coupled representation learning for domains, intents and slots in spoken language understanding. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, p. 714–719 : IEEE.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). BART : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the ACL*, p. 7871–7880, Online : ACL. DOI : 10.18653/v1/2020.acl-main.703.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : ACL.

LUO Z., XIE Q. & ANANIADOU S. (2023). Chatgpt as a factual inconsistency evaluator for text summarization.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the ACL*, p. 7203–7219, Online : ACL. DOI : 10.18653/v1/2020.acl-main.645.

MAYNEZ J., NARAYAN S., BOHNET B. & MCDONALD R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the ACL*, p. 1906–1919, Online : ACL. DOI : 10.18653/v1/2020.acl-main.173.

MORADSHAHI M., SHEN T., BALI K., CHOUDHURY M., DE CHALENDAR G., GOEL A., KIM S., KODALI P., KUMARAGURU P., SEMMAR N., SEMNANI S., SEO J., SESHADRI V., SHRIVASTAVA M., SUN M., YADAVALLI A., YOU C., XIONG D. & LAM M. (2023). X-RiSAWOZ : High-quality end-to-end multilingual dialogue datasets and few-shot agents. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Findings of the Association for Computational Linguistics : ACL 2023*, p. 2773–2794, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.174.

NAN F., NALLAPATI R., WANG Z., NOGUEIRA DOS SANTOS C., ZHU H., ZHANG D., MCKEOWN K. & XIANG B. (2021). Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the ACL : Main Volume*, p. 2727–2733, Online : ACL. DOI : 10.18653/v1/2021.eacl-main.235.

TANG X., NAIR A., WANG B., WANG B., DESAI J., WADE A., LI H., CELIKYILMAZ A., MEHDAD Y. & RADEV D. (2022). CONFIT : Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the ACL : Human Language Technologies* : ACL. DOI : 10.18653/v1/2022.naacl-main.415.

TRIONE J. (2014). Extraction methods for automatic summarization of spoken conversations from call centers (méthodes par extraction pour le résumé automatique de conversations parlées provenant de centres d'appels) [in French]. In *Proceedings of TALN 2014 (Volume 4 : RECITAL - Student Research Workshop)*, p. 104–111, Marseille, France : Association pour le Traitement Automatique des Langues.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.

WANG B., ZHANG C., ZHANG Y., CHEN Y. & LI H. (2022). Analyzing and evaluating faithfulness in dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 4897–4908, Abu Dhabi, United Arab Emirates : ACL. DOI : 10.18653/v1/2022.emnlp-main.325.

ZHAO L., ZHENG F., HE K., ZENG W., LEI Y., JIANG H., WU W., XU W., GUO J. & MENG F. (2021). Todsum : Task-oriented dialogue summarization with state tracking.

ZHOU Y., PORTET F. & RINGEVAL F. (2022). Effectiveness of French language models on abstractive dialogue summarization task. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 3571–3581, Marseille, France : European Language Resources Association.

A Corpus DECODA

A.1 Example du corpus DECODA

DIALOGUE : *autre :* NNAAMMEE va vous répondre *autre :* NNAAMMEE bonjour *appelant :* bonjour monsieur *appelant :* monsieur j' ai un renseignement à vous demander *autre :* oui *appelant :* j' ai grand garçon handicapé mental qui a une un coupon Améthyste *autre :* hm *appelant :* bon évidemment ça ne dure pas un an il se démagnétise et avant j' allais euh trente rue Championnet *autre :* oui *appelant :* le faire changer *autre :* hm *appelant :* mais leur numéro de téléphone a changé auriez vous l' amabilité de lu de me donner le numéro le nouveau numéro *autre :* alors VGC à Championnet c' est le zéro un cinquante-huit *appelant :* oui *autre :* soixante-dix-sept trois fois *appelant :* soixante-dix-sept trois fois *autre :* hm *appelant :* très bien je vous en remercie *appelant :* monsieur *autre :* bonne *autre :* journée madame au revoir *appelant :* au revoir **RÉSUMÉ :** numéro VGC pour faire changer carte Améthyste démagnétisée
autre : va vous répondre *conseiller :* bonjour *appelant :* oui allô bonjour *appelant :* je vous appelle car euh j'ai perdu mon porte-monnaie dans le... dans le train, et ça serait pour savoir quelles sont, euh comment on fait en fait ? *conseiller :* oui quel euh, quel train ? *appelant :* euh c'était le train pour aller à Montreau *conseiller :* oui il faut voir avec la SNCF madame *appelant :* d'accord *appelant :* je vous remercie *conseiller :* mais je vous en prie *appelant :* au revoir, bonne journée *conseiller :* bonne journée, au revoir **RÉSUMÉ :** Une appelante a perdu son porte-monnaie en Gare pour aller à Montreau. Le conseiller lui rappelle donc qu'il s'agit d'une requête à adresser à la SNCF et non à la RATP.

TABLE 7 – Différence entre les parties 1 (Premier exemple : 20091112_RATP_SCD_0152) et 3 (deuxième exemple : 20110704_RATP_SCD_0018) de DECODA.

A.2 Distribution des motifs d'appel de DECODA

Cette annexe présente la distribution des motifs d'appel dans notre jeu de données. On distingue 15 motifs d'appel selon la demande de l'appelant. On peut en citer : ITNR (demande d'itinéraire), OBJT (Objet trouvé/perdu) NVGO (Pass navigo), etc.

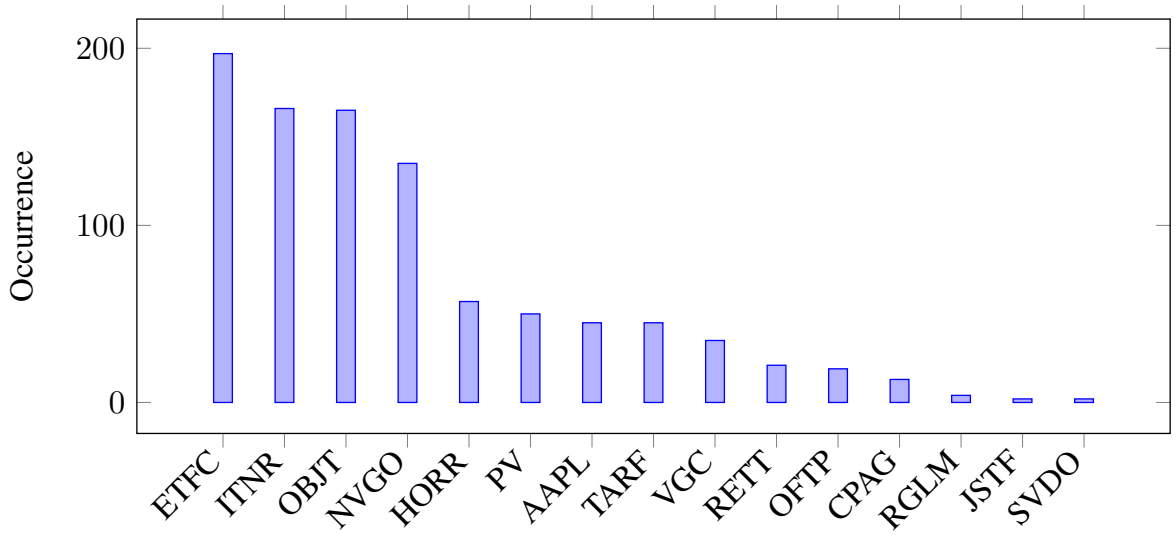


FIGURE 1 – Distribution du motif d'appel de DECODA pour le jeu de données présenté au tableau 1

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Auto-correction et oracle dynamique : certains effets n'apparaissent qu'à taille réduite

Fang Zhao Timothée Bernard
Laboratoire de linguistique formelle, Université Paris Cité
fang.zhao@etu.u-paris.fr, timothee.bernard@u-paris.fr

RÉSUMÉ

Nous étudions l'effet de la capacité d'auto-correction, de l'utilisation d'un oracle dynamique et de la taille du modèle, sur la performance d'un analyseur joint (morpho)syntaxe/sémantique. Nous montrons qu'avec un modèle de taille réduite, la possibilité d'auto-correction est nuisible en sémantique mais bénéfique en syntaxe, tandis que l'utilisation d'un oracle dynamique augmente la performance en sémantique. Nous constatons également que ces effets sont souvent atténués pour des modèles de taille plus importante.

ABSTRACT

Self-correction and dynamic oracle : some effects only appear at reduced size

We study the effect of the possibility of self-correction, the use of a dynamic oracle, and model size, on the performance of a joint (morpho)syntax/semantics parser. We show that with a reduced model size, the possibility of self-correction is detrimental to semantics performance but beneficial for syntax performance, and that the use of a dynamic oracle increases semantic performance. We also find that these effects are often mitigated for larger models.

MOTS-CLÉS : auto-correction, oracle dynamique, taille des modèles, syntaxe, sémantique.

KEYWORDS: self-correction, dynamique oracle, model size, syntax, semantics.

1 Introduction

Nous avons tous connu cette situation où, en lisant une phrase, nous avons initialement mal saisi son sens. Cependant, nous sommes également capable de nous corriger rapidement lorsque nous réalisons nos erreurs. Cette capacité d'auto-correction chez les humains soulève une question intéressante : les performances de systèmes du Traitement Automatique des Langues (TAL) pourraient-elles être améliorées par l'adoption de mécanismes d'auto-correction ?

Lee *et al.* (2018) et Lyu *et al.* (2019) ont montré que le raffinement itératif, un mécanisme d'auto-correction, pouvaient augmenter la performance de systèmes de traduction automatique et d'étiquetage des rôles sémantiques (*Semantic Role Labeling*, *SRL*), respectivement. Dary (2022) a étudié un mécanisme de *retour arrière* permettant une ré-analyse partielle dans le cadre d'un analyseur incrémental et a montré que ce mécanisme augmentait la performance du système en étiquetage morpho-syntaxique et analyse syntaxique en dépendances, et ce pour différentes langues. En revanche, dans un travail précédent (Zhao, 2022), nous avons montré que les corrections n'apportaient pas de gain de performance au système de transitions plus exotique de l'analyseur joint de Bernard (2021). Nous faisons

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 352–361.

ici l'hypothèse que la taille du réseau de neurones utilisé (c.-à-d. le nombre de paramètres du modèle) serait liée au fait que la capacité d'auto-correction n'apporte pas toujours de gain en performance ; la capacité d'auto-correction serait un atout pour un modèle de petite taille, atout dont l'impact s'amenuiserait avec la taille du modèle.

Dans ce contexte, notre étude se penche sur l'effet de différents facteurs sur la performance du système de Bernard (2021). Plus précisément, nous étudions non seulement la liberté laissée au système d'effectuer des corrections, mais aussi le type d'oracle utilisé à l'entraînement, en plus de la variation de la taille du réseau. Nous montrons que pour un petit modèle, alors qu'un oracle dynamique peut augmenter la performance dans la plupart des cas, la possibilité d'auto-correction est bénéfique ou nuisible selon les tâches. Nous constatons également qu'en augmentant la taille des modèles, les effets sont atténués voire disparaissent.

Nous souhaitons souligner le fait que le sujet étudié nous amène à travailler avec des modèles de taille réduite par rapport à l'état de l'art et ne pouvant donc pas rivaliser en termes de performance. Ce genre de modèles est cependant utile lorsque les ressources de calcul sont limitées.

2 Revue de la littérature

On désigne par *auto-correction* d'un analyseur linguistique toute action modifiant le résultat de ses décisions antérieures. Le raffinement itératif (Lee *et al.*, 2018; Lyu *et al.*, 2019) est un mécanisme d'auto-correction : un premier module est utilisé pour produire une prédiction initiale, puis un second module calcule une nouvelle prédiction fondée sur la prédiction initiale. Cette nouvelle prédiction peut ensuite être utilisée comme entrée pour une itération ultérieure du processus de raffinement. Cette technique a été utilisée notamment en traduction automatique Lee *et al.* (2018) et en étiquetage des rôles sémantiques Lyu *et al.* (2019).

Dary (2022) introduit un analyseur incrémental par transition. Cet analyseur lit le texte brut de gauche à droite, caractère par caractère, effectuant de manière jointe les tâches de segmentation (en mots et en phrase), d'étiquetage morpho-syntaxique, de lemmatisation et d'analyse syntaxique en dépendances. Un mécanisme de correction y est implémenté sous forme de retour arrière. Concrètement, en lisant le texte, lorsque le modèle reçoit une information qu'il interprète comme incompatible avec l'analyse actuelle, le modèle retourne à une certaine position précédemment traversée et effectue de nouveau l'analyse depuis cette dernière position — mais en gardant en mémoire certaines des informations acquises entre-temps.

Bernard (2021) introduit MTI-tagsynsem, un système d'analyse jointe de la morpho-syntaxe, de la syntaxe (en dépendances) et de la sémantique (en dépendances) intégrant des possibilités d'auto-correction directement dans son système de transition. Zhao (2022) montre cependant que ce système [1] n'effectue que très peu de corrections lorsqu'il est entraîné avec un *oracle statique* standard, c.-à-d. lorsque le système est entraîné à reproduire les annotations de référence sur des trajectoires sans erreur (et ne contenant donc aucune correction). Zhao s'intéresse alors à la possibilité de substituer l'oracle statique par un *oracle dynamique* (Goldberg & Nivre, 2012). Concrètement, un oracle dynamique calcule à chaque étape quelles sont les actions admissibles — dans notre cas, les actions qui introduisent des prédictions correctes et/ou corrigent des prédictions incorrectes — et si les

1. Ou plus précisément, une variante n'effectuant que les tâches d'étiquetage morpho-syntaxique et d'analyse sémantique en dépendances (donc, sans l'analyse syntaxique).

actions dont la probabilité est maximisée sont toujours des actions admissibles, les actions définissant la trajectoire explorée à l'entraînement peuvent être autant des actions admissibles (comme il est trivialement le cas avec un oracle statique) que des actions non-admissibles prédites par le modèle (ce qui introduit donc des erreurs dans la trajectoire, ensuite corrigées si le système de transition le permet — ce qui est le cas ici) [2]. Zhao (2022) montre que le système entraîné avec l'oracle dynamique apprend effectivement à effectuer des corrections, mais que la performance du système n'est pas significativement plus haute qu'avec un oracle statique.

Les biais inductifs d'un modèle d'apprentissage sont les hypothèses impliquées durant le processus d'apprentissage (Zhao *et al.*, 2018). Les biais inductifs ont été longtemps considérés comme nécessaires pour qu'un modèle puisse généraliser au-delà des exemples d'entraînement (Mitchell, 1980). Cependant, Bachmann *et al.* (2023) ont montré qu'un manque de biais inductifs pouvait être compensé par une augmentation de la taille du modèle. Ils montrent qu'à grandes tailles, les modèles construits autour d'un simple *perceptron multicouche* (*multilayer perceptron* ; MLP) peuvent atteindre sur certaines tâches classiques de vision (telles que CIFAR10 et CIFAR100 ; Krizhevsky, 2009) une performance similaire à ResNet18 (He *et al.*, 2016), un modèle fondé sur une architecture présentant de plus forts biais inductifs adaptés aux tâches de vision. Dans le même ordre d'idée, nous avançons l'hypothèse que l'absence de gain de performance observée en TAL par Zhao (2022) entre un système qui s'auto-corrige et une version sans auto-correction peut être attribuée à la taille du réseau utilisé, suffisamment importante pour atténuer les effets des possibilités d'auto-correction, et qu'un gain de performance pourrait être observé à taille plus faible.

3 Données

Nous travaillons avec une partie des données anglaises du jeu de données *SemEval 2015 Task 18* (Oepen *et al.*, 2015). Ce jeu de données contient des annotations (morpho-)syntaxiques et sémantiques notamment pour les textes du *Penn Treebank* (*PTB* ; Marcus *et al.*, 1993). Les annotations en morpho-syntaxe sont directement celles du du PTB. Pour la syntaxe, nous utilisons les arbres obtenus par conversion en dépendances *Stanford Basic* (De Marneffe & Manning, 2008) des arbres de constituants du PTB [3]. Les annotations sémantiques, le cœur de ce jeu de données centré sur l'analyse sémantique en dépendances (*Semantic Dependency Parsing*, *SDP*), sont des graphes acycliques orientés (*Directed Acyclic Graph*, *DAG*). Les noeuds de ces graphes correspondent aux tokens de la phrase et reçoivent non pas nécessairement un unique arc entrant, mais possiblement zéro ou plusieurs. En outre, chaque phrase peut avoir de zéro à plusieurs noeuds annotés comme *prédicat sommet* (*top predicate*) [4]. Alors que le jeu de données de SemEval 2015 Task 18 propose pour chaque phrase jusqu'à trois graphes sémantiques, issus de trois formalismes distincts, nous nous servons uniquement des annotations en *DELPH-IN MRS-Derived Bi-Lexical Dependencies* (*DM* ; Oepen *et al.*, 2014) dans nos expériences.

2. L'idée d'explorer ainsi des trajectoires imparfaites a ensuite été reprise pour définir l'*échantillonage programmé* (*scheduled sampling* ; Bengio *et al.*, 2015).

3. Ce jeu de données contient 33964 et 1692 phrases respectivement pour les ensembles d'entraînement et de développement. Nous réservons l'ensemble de test pour d'éventuelles études futures visant l'état de l'art, ce qui n'est pas l'objectif de cette étude.

4. La notion de prédicat sommet dans un graphe sémantique est comparable à celle de racine dans un arbre syntaxique. Sa définition précise varie d'un formalisme à l'autre (Oepen *et al.*, 2015).

4 Modèle

Nous décrivons ici la variante du système MTI-tagsynsem que nous utilisons dans cette étude. Contrairement à un système par transition usuel, ce système n'analyse pas son entrée de gauche à droite et ne repose pas sur une *pile* (*stack*) ni un *tampon* (*buffer*). À chaque étape de calcul (c.-à-d., à chaque transition), *pour chaque token* de la phrase, une action est choisie depuis un ensemble d'actions intégrant toutes les tâches (étiquetage morpho-syntaxique, analyse syntaxique en dépendances, analyse sémantique en dépendances) ; ces actions sont principalement des actions d'étiquetage ou de sélection de tête (Zhang *et al.*, 2017).

Nous détaillons ici le répertoire d'actions utilisé dans cette étude. Afin de faciliter la discussion, nous utilisons les notations $j \xrightarrow{l} i$ pour désigner une dépendance (syntaxique ou sémantique) de relation l depuis j (gouverneur) vers i (dépendant), et $j \rightarrow i$ pour désigner une dépendance de relation quelconque depuis j vers i. Pour un token (de position) i dans la phrase, les actions possibles sont :

- TAG-t, qui ajoute l'étiquette morpho-syntaxique t au token i, et TAG[ERASE], qui supprime l'étiquette de i si elle existe ;
- SYN-j-l, qui ajoute la dépendance syntaxique $j \xrightarrow{l} i$, et SYN-j[ERASE], qui supprime la dépendance syntaxique $j \rightarrow i$ si elle existe ;
- ROOT, qui détermine le token i comme étant la racine syntaxique, et ROOT[ERASE], qui détermine i comme n'étant pas la racine ;
- SEM-j-l, qui ajoute la dépendance sémantique $j \xrightarrow{l} i$, et SEM-j[ERASE], qui supprime la dépendance sémantique $j \rightarrow i$ si elle existe ;
- TOP_PRED, qui détermine le token i comme étant un *prédicat sommet* (*top predicate*), et TOP_PRED[ERASE], qui détermine i comme n'étant pas un prédicat sommet ;
- HALT qui n'a pas d'autre effet que de provoquer la fin de l'analyse lorsque cette action est simultanément choisie pour chaque token [5].

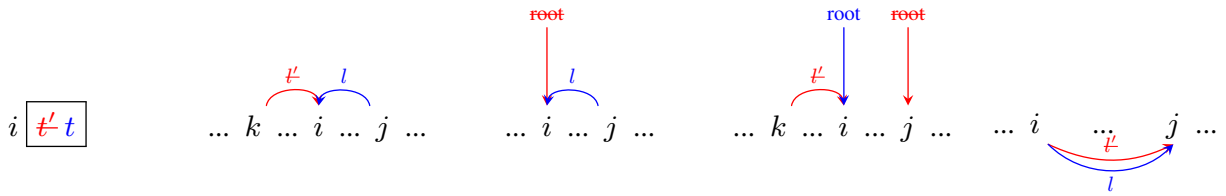
Si le système choisit pour un token une action ayant pour effet l'ajout d'une annotation alors que celle-ci existe déjà, ou la suppression d'une annotation alors que celle-ci n'existe pas, l'effet de cette action est nul. Et si une action choisie par le système ajoute une annotation incompatible avec des annotations antérieures, celles-ci sont effacées [6]. Nous détaillons ci-dessous toutes les incompatibilités dans le système utilisé. Pour un token (de position) i dans la phrase, lorsque :

a. une action TAG-t est choisie, si une étiquette morpho-syntaxique $t' \neq t$ a été prédite pour i, alors cette prédiction est supprimée ;
b. une action SYN-j-l est choisie, si 1) une dépendance syntaxique $k \rightarrow i$ a été prédite ou si 2) i a été prédit comme racine syntaxique, alors cette prédiction est supprimée ;
c. une action ROOT est choisie, si une dépendance syntaxique $k \rightarrow i$ a été prédite et/ou si un token $j \neq i$ a été prédit comme racine syntaxique, alors ces prédictions sont supprimées ;
d. une action SEM-j-l est choisie, si une dépendance sémantique $j \xrightarrow{l'} i$ avec $l' \neq l$ a été prédite, alors cette prédiction est supprimée.

Nous illustrons ces incompatibilités avec les schémas en figure 1.

5. L'analyse est aussi arrêtée automatiquement après un certain nombre d'étapes. Dans tous les cas, aucun mécanisme ne vérifie que les structures d'annotations prédites (séquence d'étiquettes morpho-syntaxiques et arbre de dépendances syntaxiques) sont complètes au moment de l'arrêt.

6. Cela est aussi le cas lorsque le système choisit des actions incompatibles entre elles au sein d'une même étape. En effet, les actions choisies pour les différents tokens lors d'une même étape ne sont pas exécutées en même temps, mais dans un ordre aléatoire. Si deux d'entre elles sont en conflit, la dernière effacera l'annotation ajoutée par la première.



a. TAG-t b.1) SYN-j-l b.2) SYN-j-l c. ROOT d. SEM-j-l

FIGURE 1 – **Illustrations des cas d'incompatibilité.** Une boîte à droite d'un token indique une étiquette morpho-syntaxique, les annotations au dessus du texte correspondent à la structure syntaxique et les annotations en dessous du texte à la structure sémantique. Les ~~annotations supprimées~~ durant l'étape illustrée sont marquées en rouge et avec une ligne de suppression. Les annotations ajoutées durant l'étape illustrée sont marquées en bleue.

Ce système nous intéresse parce qu'il intègre des possibilités d'auto-correction de manière particulièrement légère dans le sens où (i) il n'existe pas de distinctions entre deux types de phases de prédiction comme avec le raffinement itératif et (ii) la définition d'un oracle dynamique est simple, ce qui permet d'effectuer un apprentissage classique, sans nécessité de recourir à de l'apprentissage par renforcement comme avec le système de Dary *et al.* (2022). De plus, le fait que ce système effectue de manière jointe différentes tâches classiques du TAL permet d'étudier l'incidence de la capacité d'auto-correction avec plus de globalité.

5 Incidence de la capacité d'auto-correction et du type d'oracle

Comme mentionné précédemment, Zhao (2022) montre qu'un modèle MTI-tagsem (une variante de MTI-tagsynsem n'effectuant pas d'analyse syntaxique) avec auto-correction a une performance similaire à celle d'une version sans auto-correction, c.-à-d. ne différant que sur le fait que lorsqu'une action est en conflit avec des annotations déjà prédites, les effets de cette action sont ignorés. Se pourrait-il que la taille du réseau soit assez importante pour avoir compensé le gain en performance du système que l'auto-correction aurait pu apporter ? Afin de répondre à cette question, nous étudions deux séries de modèles, se distinguant par leur taille : GRAND avec $\approx 2{,}7 \times 10^6$ et PETIT avec $\approx 1{,}7 \times 10^6$ paramètres entraînables [7].

Comme Zhao (2022), nous étudions deux modèles : un premier modèle sans capacité d'auto-correction et entraîné avec oracle statique (OS-CORR) et un second modèle avec auto-correction et donc entraîné avec oracle dynamique (OD+CORR). De plus, afin de nous assurer que la différence en performance des deux modèles — si une telle différence apparaît — vient de l'auto-correction et non pas de l'oracle dynamique, nous ajoutons dans la comparaison un troisième modèle sans auto-correction mais entraîné avec oracle dynamique (OD-CORR) [8]. Notre hypothèse est qu'en dessous d'une certaine taille de modèle, les effets de la capacité d'auto-correction commencent à apparaître : le modèle avec auto-correction (OD+CORR) aura une performance plus élevée que celle des modèles sans

7. Ce comptage exclut les représentations lexicales vectorielles (*word embeddings*) pré-entraînées. Nous utilisons les vecteurs *GloVe* 6B (Pennington *et al.*, 2014) de dimension 100 et ne conservons que les vecteurs des mots apparaissant dans l'ensemble d'entraînement. Ces représentations lexicales représentent $\approx 3 \times 10^6$ paramètres entraînables supplémentaires.

8. L'oracle dynamique de ce modèle ne calcule comme actions admissibles que les actions non-correctives, c.-à-d., qui sont compatibles avec toutes les annotations actuelles.

auto-correction (OS-CORR et OD-CORR).

	Taille	OS-CORR	OD-CORR	OD+CORR
SEM	PETIT	$0{,}870_{(0{,}903\,/\,0{,}839)}$ / 0,929	$0{,}891_{(0{,}893\,/\,0{,}889)}$ / 0,995	$0{,}888_{(0{,}904\,/\,0{,}871)}$ / 0,964
	GRAND	$0{,}897_{(0{,}902\,/\,0{,}891)}$ / 0,988	$0{,}900_{(0{,}901\,/\,0{,}900)}$ / 0,998	$0{,}893_{(0{,}905\,/\,0{,}882)}$ / 0,974
SYN	PETIT	$0{,}899_{(0{,}899\,/\,0{,}899)}$ / 0,999	$0{,}899_{(0{,}899\,/\,0{,}899)}$ / 0,999	$0{,}906_{(0{,}913\,/\,0{,}900)}$ / 0,986
	GRAND	$0{,}909_{(0{,}909\,/\,0{,}909)}$ / 1,000	$0{,}907_{(0{,}907\,/\,0{,}907)}$ / 1,000	$0{,}910_{(0{,}915\,/\,0{,}905)}$ / 0,989
TAG	PETIT	$0{,}969_{(0{,}969\,/\,0{,}970)}$ / 1,000	$0{,}969_{(0{,}969\,/\,0{,}969)}$ / 1,000	$0{,}968_{(0{,}970\,/\,0{,}967)}$ / 0,997
	GRAND	$0{,}970_{(0{,}970\,/\,0{,}970)}$ / 1,000	$0{,}970_{(0{,}970\,/\,0{,}970)}$ / 1,000	$0{,}970_{(0{,}971\,/\,0{,}969)}$ / 0,997

TABLE 1 – **Performance en F1(précision / rappel) / taux d'analyse.** SEM : analyse sémantique en dépendances, SYN : analyse syntaxique en dépendances, TAG : étiquetage morpho-syntaxique. Les résultats sont calculés sur la moyenne de neuf exécutions pour chaque modèle, évalué sur l'ensemble de développement.

	Taille	1ères annotations	nombre de succès | raté | échec		
SEM	PETIT	$0{,}884_{(0{,}888\,/\,0{,}881)}$ / 0,991	$3193_{(56{,}2\%)}$	$7_{(0{,}1\%)}$	$2486_{(43{,}7\%)}$
	GRAND	$0{,}889_{(0{,}886\,/\,0{,}892)}$ / 1,006	$3086_{(56{,}6\%)}$	$9_{(0{,}2\%)}$	$2353_{(43{,}2\%)}$
SYN	PETIT	$0{,}899_{(0{,}901\,/\,0{,}897)}$ / 0,996	$3607_{(59{,}0\%)}$	$128_{(2{,}1\%)}$	$2382_{(38{,}9\%)}$
	GRAND	$0{,}903_{(0{,}904\,/\,0{,}901)}$ / 0,997	$2945_{(59{,}1\%)}$	$105_{(2{,}1\%)}$	$1929_{(38{,}7\%)}$
TAG	PETIT	$0{,}967_{(0{,}968\,/\,0{,}967)}$ / 0,999	$918_{(54{,}7\%)}$	$8_{(0{,}5\%)}$	$751_{(44{,}8\%)}$
	GRAND	$0{,}969_{(0{,}969\,/\,0{,}969)}$ / 0,999	$900_{(54{,}3\%)}$	$11_{(0{,}7\%)}$	$745_{(45{,}0\%)}$

TABLE 2 – **Analyse de corrections pour le modèle OD+CORR**. SEM : analyse sémantique en dépendances, SYN : analyse syntaxique en dépendances, TAG : étiquetage morpho-syntaxique. L'évaluation des 1ères annotations : F1(précision / rappel) / taux d'analyse. Les résultats sont calculés sur la moyenne de neuf exécutions, évalués sur l'ensemble de développement.

Le tableau 1 regroupe les performances des modèles OS-CORR, OD-CORR, OD+CORR en tailles PETIT et GRAND sur les trois tâches : analyse sémantique en dépendances, analyse syntaxique en dépendances et étiquetage morpho-syntaxique. Le système MTI-tagsynsem n'étant contraint à produire une analyse complète pour aucune des tâches (voir note 5), nous utilisons comme métriques la précision, le rappel et la F1. Nous calculons aussi, pour chaque tâche, un *taux d'analyse* : il s'agit du ratio entre le nombre d'annotations prédites et le nombre d'annotations de référence. Le taux d'analyse est toujours inférieur ou égale à 1,0 en syntaxe et morpho-syntaxe, mais peut aussi être supérieur à 1.0 en sémantique. Lorsque le taux d'analyse s'approche de 1.0, précision, rappel et F1 convergent — vers le score d'attachement étiqueté (*LAS* ou *labeled attachment score*) en syntaxe, et vers l'exactitude en morpho-syntaxe.

Nous regardons tout d'abord les résultats de l'analyse sémantique en dépendances, qui est choisie comme la tâche principale. Cela veut dire que nous utilisons la F1 en sémantique comme critère pour déterminer les meilleurs hyperparamètres et pour arrêter l'entraînement du système par arrêt précoce (*early-stopping*). En taille PETITE, comme attendu, le modèle avec oracle dynamique et auto-correction a une F1 sémantique plus élevée (0,888) que le modèle avec oracle statique et sans auto-correction (0,870). Cependant, contrairement à notre prédiction, c'est le modèle avec oracle dynamique mais sans auto-correction qui a la meilleure F1 des trois modèles (0,891). Le gain du modèle OD-CORR est principalement dû au fait que son rappel (0,893) et son taux d'analyse (0,995) sont plus élevés, bien que sa précision soit plus faible (0,893). Cela veut dire que OD-CORR a tendance à effectuer des analyses plus complètes mais au prix de la précision. Cependant, la différence de F1 en sémantique entre OS-CORR et OD-CORR est atténuée en taille GRAND.

La performance en syntaxe des modèles sans auto-correction, OS-CORR et OD-CORR, est similaire pour les deux tailles considérées. En revanche, le modèle avec auto-correction, OD+CORR, a une meilleure F1 en syntaxe que les modèles sans auto-correction quand les modèles sont petits. Cette différence s'atténue elle aussi quand les modèles sont en taille GRAND.

En termes d'étiquetage morpho-syntaxique, nous constatons que tous les modèles ont une performance similaire quelle que soit leur taille, sur toutes les métriques d'évaluation.

En résumé, l'utilisation de l'oracle dynamique augmente la performance en sémantique, alors que la capacité d'auto-correction la fait baisser. La capacité d'auto-correction fait par contre augmenter la performance en syntaxe, tandis que l'oracle dynamique seul n'a pas d'effet dans ce cas. Dans la plupart des cas, ces effets sont plus visibles lorsque les modèles ont une taille relativement petite et s'atténuent à plus grandes tailles.

Afin de comprendre comment les corrections auraient entraîné les effets en syntaxe et en sémantique mentionnés ci-dessus, nous nous proposons d'effectuer une analyse de corrections dans la section suivante.

6 Analyse de corrections

Nous classons les actions effectuées durant un épisode d'analyse de la manière suivante :

— une action est *corrective* si et seulement si elle efface au moins une annotation précédente ;
— sinon, elle est *non-corrective*.

Une action corrective peut être du type :

— *succès* si la ou les annotations supprimées étaient incorrectes, et que l'annotation éventuellement ajoutée est correcte ;
— *raté* si la ou les annotations supprimées étaient incorrectes, et que l'annotation ajoutée est aussi incorrecte ;
— *échec* si au moins une des annotations supprimées était correcte.

Nous définissons aussi : une *première annotation* est une annotation qui, lorsqu'elle est ajoutée, est compatible avec l'ensemble des premières annotations précédentes [9]. Il est possible d'évaluer les premières annotations d'un modèle avec les mêmes métriques que celles utilisées pour les prédictions globales du modèle (c.-à-d. à l'issue de l'analyse, en prenant en compte les corrections) : F1, précision, rappel et taux d'analyse.

Nous présentons les performances des premières annotations ainsi que le nombre des différents types d'actions correctives du modèle OD+CORR dans le tableau 2.

Globalement, nous constatons plus de succès que d'échecs pour toutes les tâches et tailles de modèle. En sémantique, la performance de OD+CORR en termes de premières annotations est inférieure à la performance de OD-CORR, pour toutes les métriques d'évaluation. Les actions correctives, bien que plus souvent des succès que des échecs, ne permettent pas à la performance de OD+CORR de dépasser celle de OD-CORR (indiquée en tableau 1). Nous constatons aussi que le taux d'analyse des premières annotations est proche de 1,0. Quand les corrections sont prises en compte, le taux

9. Toutes les actions non-correctives n'introduisent pas une première annotation. Considérons par exemple, pour un token de position i, que la séquence d'action choisie soit TAG-t (étape 0), TAG[ERASE] (étape 1), TAG-t' (étape 2) ; seule l'étiquetage de i avec t est considérée comme une première annotation, alors que TAG-t' en étape 2 est une action non-corrective (l'étiquetage précédent ayant été supprimée à l'étape 1).

d'analyse ainsi que le rappel baisse, mais la précision augmente.

En syntaxe, nous constatons que la performance de OD+CORR en termes de premières annotations est déjà au même niveau que la performance de OD-CORR. Compte tenu du fait qu'il y a plus d'actions correctives succès que d'échecs, la performance de OD+CORR dépasse celle de OD-CORR. Avec les corrections, par rapport aux premières annotations, la précision augmente sans que le rappel ne baisse et ce alors même que le taux d'analyse baisse.

Nous observons donc que les corrections, par rapport aux premières annotations, mènent à une baisse variable du taux d'analyse en syntaxe et en sémantique, et à une augmentation de la précision. Cela suggère que l'auto-correction a tendance à favoriser les annotations dont le modèle est plus « sûr » et de supprimer celles dont il est moins « sûr ». Cette différence entre syntaxe et sémantique est possiblement liée au fait que l'analyse sémantique est plus complexe, ne serait-ce que parce que le nombre de dépendances sémantiques, au contraire du nombre de dépendances syntaxiques, n'est pas exactement déterminé par le nombre de tokens.

Les modèles avec auto-correction effectuent beaucoup moins de corrections de l'étiquetage morpho-syntaxique que des structures syntaxiques et sémantiques. Les nombres de succès et d'échecs étant proches, la capacité d'auto-correction n'a donc quasiment aucun impact sur cette tâche.

7 Conclusion

Dans cette étude, nous avons exploré l'effet de différents facteurs — la capacité d'auto-correction, l'utilisation d'un oracle dynamique et la taille du modèle — sur la performance du système MTI-tagsynsem. Nous avons montré que les corrections étaient nuisibles pour la performance en sémantique mais étaient bénéfiques pour la performance en syntaxe. Selon notre analyse, l'auto-correction tend à favoriser les annotations les plus fiables d'après le modèle, ce qui se fait au détriment de l'analyse sémantique (qui est plus difficile). D'autre part, nous avons montré que l'utilisation d'un oracle dynamique augmentait la performance en sémantique. Nous avons constaté également que ces effets étaient atténués pour des modèles de taille plus importante, ce qui tend à conforter l'idée selon laquelle il y aurait plus à attendre sur le long terme d'une augmentation de la quantité de donnée d'entraînement et de la taille des modèles que de variations fines d'architecture ou d'algorithme (Sutton, 2019) ; une telle augmentation, cependant et lorsqu'elle est possible, n'est pas ni sans coût ni sans problème (Bender *et al.*, 2021).

Remerciements

Ces travaux ont été financés par une bourse Émergence 2021 (projet SYSNEULING) de l'IdEx *Université Paris Cité*.

Références

BACHMANN G., ANAGNOSTIDIS S. & HOFMANN T. (2023). Scaling mlps : A tale of inductive bias. *arXiv preprint arXiv :2306.13575*.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

BENDER E. M., GEBRU T., MCMILLAN-MAJOR A. & SHMITCHELL S. (2021). On the Dangers of Stochastic Parrots : Can Language Models Be Too Big ? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, p. 610–623, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3442188.3445922.

BENGIO S., VINYALS O., JAITLY N. & SHAZEER N. (2015). Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, p. 1171–1179, Cambridge, MA, USA : MIT Press.

BERNARD T. (2021). Multiple tasks integration : Tagging, syntactic and semantic parsing as a single task. In P. MERLO, J. TIEDEMANN & R. TSARFATY, Éds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 783–794, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.66.

DARY F. (2022). *Modèles incrémentaux pour le traitement automatique des langues*. Thèse de doctorat. Thèse de doctorat dirigée par Nasr, Alexis et Fourtassi, Abdellah Informatique Aix-Marseille 2022.

DARY F., PETIT M. & NASR A. (2022). Dependency Parsing with Backtracking using Deep Reinforcement Learning. *Transactions of the Association for Computational Linguistics*, **10**, 888–903. DOI : 10.1162/tacl_a_00496.

DE MARNEFFE M.-C. & MANNING C. D. (2008). *Stanford typed dependencies manual*. Rapport interne, Technical report, Stanford University.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

GOLDBERG Y. & NIVRE J. (2012). A dynamic oracle for arc-eager dependency parsing. In *International Conference on Computational Linguistics*.

HE K., ZHANG X., REN S. & SUN J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 770–778. DOI : 10.1109/CVPR.2016.90.

KRIZHEVSKY A. (2009). Learning multiple layers of features from tiny images. p. 32–33.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LEE J., MANSIMOV E. & CHO K. (2018). Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 1173–1182, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1149.

LYU C., COHEN S. B. & TITOV I. (2019). Semantic role labeling with iterative structure refinement. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 1071–1082, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1099.

MARCUS M. P., MARCINKIEWICZ M. A. & SANTORINI B. (1993). Building a large annotated corpus of english : The penn treebank. *Comput. Linguist.*, **19**(2), 313–330.

MITCHELL T. M. (1980). The need for biases in learning generalizations.

OEPEN S., KUHLMANN M., MIYAO Y., ZEMAN D., CINKOVÁ S., FLICKINGER D., HAJIČ J. & UREŠOVÁ Z. (2015). SemEval 2015 task 18 : Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 915–926, Denver, Colorado : Association for Computational Linguistics. DOI : 10.18653/v1/S15-2153.

OEPEN S., KUHLMANN M., MIYAO Y., ZEMAN D., FLICKINGER D., HAJIČ J., IVANOVA A. & ZHANG Y. (2014). SemEval 2014 task 8 : Broad-coverage semantic dependency parsing. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 63–72, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.3115/v1/S14-2008.

PENNINGTON J., SOCHER R. & MANNING C. (2014). GloVe : Global vectors for word representation. In A. MOSCHITTI, B. PANG & W. DAELEMANS, Éds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543, Doha, Qatar : Association for Computational Linguistics. DOI : 10.3115/v1/D14-1162.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SUTTON R. S. (2019). The Bitter Lesson.

ZHANG X., CHENG J. & LAPATA M. (2017). Dependency Parsing as Head Selection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, p. 665–676, Valencia, Spain : Association for Computational Linguistics.

ZHAO F. (2022). Auto-correction dans un analyseur neuronal par transitions : un comportement factice ?(self-correction in a transition-based neural parser : a spurious behaviour ?). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 : 24e Rencontres Etudiants Chercheurs en Informatique pour le TAL (RECITAL)*, p. 20–32.

ZHAO S., REN H., YUAN A., SONG J., GOODMAN N. & ERMON S. (2018). Bias and generalization in deep generative models : An empirical study. *Advances in Neural Information Processing Systems*, **31**.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Construction d'une mesure de similarité thématique non supervisée pour les conversations

Amandine Decker[1,2] Maxime Amblard[1]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
(2) University of Gothenburg, Suède
amandine.decker@loria.fr, maxime.amblard@univ-lorraine.fr

RÉSUMÉ

La structure thématique d'une conversation représente la manière dont l'interaction est organisée à un niveau plus global que le strict enchaînement des interventions. Elle permet de comprendre comment la cohérence est maintenue sur le temps de l'échange. La création d'une mesure de similarité thématique qui donne un score de similarité à deux énoncés du point de vue thématique pourrait nous permettre de produire et d'analyser ces structures. Nous entraînons une mesure non supervisée, basée sur le modèle *BERT* avec prédiction de la phrase suivante, sur des conversations *Reddit*. La structure de *Reddit* nous fournit différents niveaux de proximité de cohérence entre des paires de messages, ce qui nous permet d'entraîner notre modèle avec une fonction de perte basée sur des comparaisons plutôt que sur des valeurs numériques attendues *a priori*. Cette mesure nous permet de trouver des ensembles d'interventions localement cohérents dans nos conversations *Reddit*, mais aussi de mesurer la variabilité en termes de thème tout au long d'une conversation.

ABSTRACT

Building an Unsupervised Topical Similarity Measure for Conversation.

The topical structure of a conversation gives insight on the way interaction is organised at a more global level than utterance sequences. It enablles us to understand how coherence is maintained throughout a dialogue. Creating a topical similarity measure which would give a similarity score to two pieces of interaction in terms of topic could enable us to analyse this structure. We train an unsupervised measure based on the *Next Sentence Prediction* BERT model on Reddit conversations. The structure of Reddit provides us different coherence levels between pairs of messages which allows us to train our model with a marginal ranking loss rather than with numerical values. This measure enables us to find locally coherent pieces of interaction in our dataset but also to measure the variability in terms of topic throughout a conversation.

MOTS-CLÉS : topic modelling, apprentissage non supervisé, corpus, dialogue, Reddit, similarité.

KEYWORDS: topic modelling, unsupervised learning, corpus, dialogue, Reddit, similarity.

1 Introduction

L'enchaînement des thèmes d'une conversation donne un aperçu de la manière dont l'interaction est organisée à un niveau global. Nous nous intéressons à la production d'une structure permettant d'analyser la dynamique des échanges au delà des énoncés eux-même. Pour parvenir à ces dernières, nous nous intéressons à la cohérence thématique entre messages. La première étape est la segmentation

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 362–375.

en thème qui se définit comme la division d'un discours en morceaux localement cohérents. La plupart du temps, cette segmentation est linéaire, c'est à dire que les morceaux thématiquement cohérents se suivent linéairement, et aucune ou peu de structure entre les morceaux n'est mise en évidence. Si cette approche peut donner de bons résultats sur des textes bien organisés et, dans une certaine mesure, sur des dialogues structurés tels que des comptes rendus de réunions, elle n'est pas aussi adaptée aux conversations informelles. Par conséquent, la production d'une segmentation hiérarchique est essentielle pour modéliser des dialogues spontanés.

Étant donné la complexité de la modélisation hiérarchique des thèmes, en particulier pour ce type de conversations, qui plus est entre plusieurs intervenants, nous proposons de construire une mesure de similarité en thème qui fournit un score de similarité entre deux messages fondée sur leur proximité thématique. L'utilisation de cette mesure permet de trouver des ensembles localement cohérents dans le dialogue, et de manière plus large de mesurer la variabilité thématique tout au long de la conversation. Puisque nous ne pouvons pas donner manuellement un score de similarité à des paires de messages, nous proposons d'utiliser des méthodes non supervisées pour entraîner cette mesure à partir de comparaisons. L'utilisation d'un ensemble de données où les interactions sont déjà organisées de manière hiérarchique nous fournit différents niveaux de cohérence que nous utilisons comme base d'entraînement. Pour cela, nous nous basons sur le média social américain *Reddit* que nous considérons comme un grand ensemble de messages, organisés en structures arborescentes, discutant d'une diversité de sujets et présentant différents formats de relations entre les messages (questions/réponses, discussion informelle, argumentation, etc.)

Nos contributions dans cet article sont de deux ordres : (1) la création d'un corpus à partir d'une extraction des données de *Reddit* et (2) l'entraînement d'une mesure de similarité thématique de manière non supervisée à partir de ces conversations. Nous appliquons également la mesure à notre ensemble de données pour identifier les ensembles de commentaires thématiquement cohérents, et nous procédons à une évaluation de cette cohérence par rapport au sujet initial le long de ces fils de discussion. Nous donnons finalement des perspectives d'utilisation de cette métrique pour construire des représentations du dialogue.

2 Background

Les médias sociaux sont largement utilisés dans la recherche sur le discours comme sources de données (Hamilton *et al.*, 2017; Baumgartner *et al.*, 2020; Balouchzahi *et al.*, 2023). Jovanovic & Leeuwen (2018) analysent la structure et le genre des dialogues tenus sur diverses plateformes de médias sociaux, ou encore Misra & Walker (2013) travaillent sur l'identification de l'accord et du désaccord dans les conversations à partir des échanges publics sur les médias sociaux. Diverses études s'appuient sur ces dialogues pour des tâches variées, par exemple sur la santé mentale (Gaur *et al.*, 2018; Turcan & McKeown, 2019; Naseem *et al.*, 2022), ou la reconnaissance des émotions (Balouchzahi *et al.*, 2023). L'activité sur les réseaux sociaux est donc considérée comme représentative de la dynamique de l'interaction dialogique. Si cela reste un usage particulier de la langue, cela nous permet de palier les difficultés de la production de transcription d'échanges réels.

Si *Reddit* n'est pas le média le plus étudié (Kathie Treen & Coan, 2022), sa structure le rend particulièrement pertinent pour notre problématique. En effet, les messages sur *Reddit* sont organisés de manière arborescente, avec un message initial (*post*) sur un sujet donné et des réponses à ce message ainsi que des réponses aux différentes réponses. À partir d'un *post*, des conversations se

mettent en place, formant une extension arborescente en lien avec le thème de départ. Sur cette structure, nous étudions la manière dont le thème évolue à partir du premier message au travers d'une séquence de réponses. Nous souhaitons également comparer l'évolution de différentes séquences issues d'un même message initial. La Section 3 détaille la structure utilisée par *Reddit*.

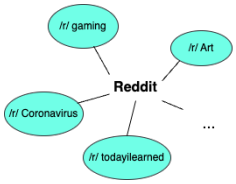
Afin d'étudier l'évolution des sujets dans notre ensemble de données, nous suivons (Wang *et al.*, 2017; Xu *et al.*, 2019; Wang *et al.*, 2020; Xing & Carenini, 2021) et construisons une mesure en nous appuyant sur la tâche de notation de la cohérence de paires de textes. La structure arborescente de *Reddit* nous permet de générer un corpus de paires de commentaires présentant trois niveaux de cohérence thématique. Pour cela, nous affinons le modèle *Next Sentence Prediction* (NSP) de BERT (Devlin *et al.*, 2019) qui utilise la fonction de perte fondée sur le *marginal ranking* pour obtenir une mesure de similarité thématique. Le modèle NSP BERT a l'avantage d'être entraîné sur un large corpus de texte non annoté, et en particulier sur une tâche de jugement cohérence. De plus, les grands modèles de langage (LLM), plus récents et plus puissants, basés sur les séquences, nécessitent des adaptations spécifiques afin d'être utilisés pour une tâche particulière comme l'évaluation de la cohérence. En effet, le *fine tuning* d'un LLM pour qu'il produise des scores de cohérence nécessiterait un ensemble de données incluant ces scores. Étant donné que l'attribution manuelle d'un score de cohérence à des paires de messages est une tâche complexe, s'appuyer sur différents niveaux de cohérence pour former un modèle est une solution de contournement qui nous parait plus prometteuse.

3 Données

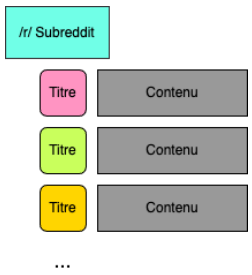
Reddit est un réseau social américain où les utilisateurs peuvent faire partie de communautés définies par des sujets particuliers comme un hobby, une culture, ou le format d'un *post* par exemple. La structure de *Reddit* est décrite dans la Figure 1. Dans les communautés, appelées *subreddits* (Figure 1a), un utilisateur peut proposer un *post* qu'il rédige (Figure 1b) et les autres utilisateurs peuvent réagir à ce *post* / interagir avec ce *post* en écrivant à leur tour un message, considéré comme un commentaire (Figure 1c). Un tel commentaire peut être une réponse directe à un *post* ou une réponse à un autre commentaire. La « discussion » ainsi créée à partir d'un *post* est organisée sous forme d'arbre dont la racine est le *post* initial et les noeuds sont les commentaires.

Nous utilisons cette structure pour former des paires de messages auxquelles nous attribuons plus ou moins de similarité thématique, et ce de façon automatique, comme nous l'expliquons dans la section suivante (Section 4). Bien que les messages récupérés ne soient pas à proprement parler du dialogue spontané oral, le format et l'organisation des commentaires nous permettent de considérer une dynamique d'interaction similaire à celle qui apparaît dans des conversations réelles. L'enjeu de la transcription est écarté, nous permettant de nous focaliser sur cette dynamique. Ces conversations sont plus ou moins longues selon le fil et impliquent un nombre variable de participants.

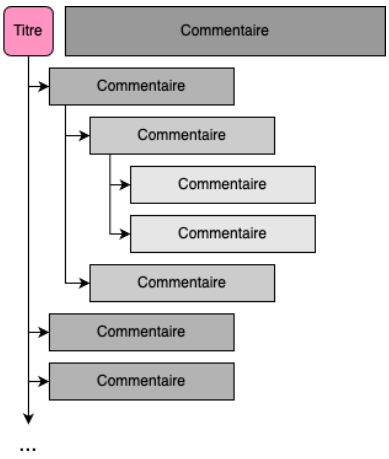
Le corpus est constitué du contenu de cinq thèmes différents (donc cinq *subreddits*, qui seront décrits en annexe dans la version finale) afin d'avoir de la variabilité des contenus et de la forme des messages. L'API *Reddit* nous permet de récupérer seulement une partie des données contenues dans un subreddit. Nous pouvons obtenir jusqu'à cent *threads* (un *post* et l'ensemble de ses commentaires) dans un *subreddit*, ces *threads* pouvant être ordonnés de différentes façons (par exemple en fonction du nombre de votes positifs qu'ils ont récemment reçu, ou en fonction de leurs votes positifs depuis la création du *subreddit*). Afin de récupérer des arbres de discussion de taille substantielle, nous avons décidé de récupérer les cent *posts* avec le plus de votes positifs depuis la création du *subreddit*. Ainsi,



(a) *Reddit* est divisé en *subreddits*.



(b) Chaque *subreddit* contient des *posts* formés d'un titre et de contenu.



(c) Chaque *post* est organisé sous forme d'arbre avec des commentaires répondant au *post* et des commentaires répondant aux autres commentaires.

FIGURE 1 – Organisation de Reddit.

pour chacun de nos cinq *subreddits*, nous avons récupéré jusqu'à cent *threads* [1] sous la forme d'arbres de discussion. Pour chaque *thread*, nous avons récupéré le titre et le contenu du *post* principal, son identifiant (ID), le nombre de votes positif, le pseudonyme de son auteur et la liste de ses 'enfants', c'est-à-dire l'ensemble des commentaires du *thread*. De même pour les commentaires nous avons récupéré leur contenu, leur ID, le nombre de votes positif, le pseudonyme de l'auteur ainsi que l'ID de leur père dans l'arbre de discussion afin de reconstruire la structure d'arbre. Nous n'utilisons ni le nombre de votes positif ni les pseudonymes des auteurs dans nos expériences mais les avons récupérés dans l'éventualité où ils s'avérerait utiles pour des analyses ultérieures. La licence de la *Data API* de *Reddit* ne nous autorise pas à partager notre dataset sans leur consentement explicite. Nous l'avons demandé mais n'avons pas encore obtenu de réponse (le délai minimum indiqué sur le formulaire est de 14 semaines). Un dataset similaire peut-être obtenu en utilisant notre code [2], bien que certains *threads* du top 100 pourraient avoir changé depuis que nous les avons récupérés. En cas d'accord d'ici publication, l'ensemble des données sera rendu accessible.

4 Modèle non supervisé

Notre objectif est donc d'entraîner un modèle pour construire une mesure de similarité thématique pour le dialogue. Comme nous ne pouvons pas attribuer manuellement une valeur de similarité thématique à des paires de messages, nous adaptons le modèle de Hearst (1997) à nos propres données. Ce modèle est entraîné grâce à une fonction de perte *marginal ranking*, ce qui signifie qu'au lieu d'apprendre à prédire un score précis pour une paire de commentaires donnée, il apprend à classer les paires de messages sur la base de leur similarité. Nous nous appuyons sur la structure hiérarchique

1. Certains *threads* nécessitaient plus de requête que l'*API* n'autorisait afin de récupérer tout l'arbre, nous avons donc décidé de les abandonner.
2. https://gitlab.inria.fr/adecker/reddittopicsimilarity.git

de *Reddit* pour construire trois niveaux de comparaisons, comme illustré dans la Figure 2. Nous considérons qu'étant donné un commentaire, un de ses fils est plus proche thématiquement qu'un de ses frères (donc l'oncle du fils), et un de ses neveux est moins proche thématiquement que ses fils et ses frères. Notre modèle utilise l'outil *Next Sentence Prediction* BERT (NSP-BERT) pour apprendre les scores de similarité. Lors de l'entraînement, le modèle reçoit les trois paires en entrée et produit un score de similarité pour chacune d'entre elles. Il est récompensé par la fonction de perte pour chaque score bien ordonné par rapport à l'un des deux autres.

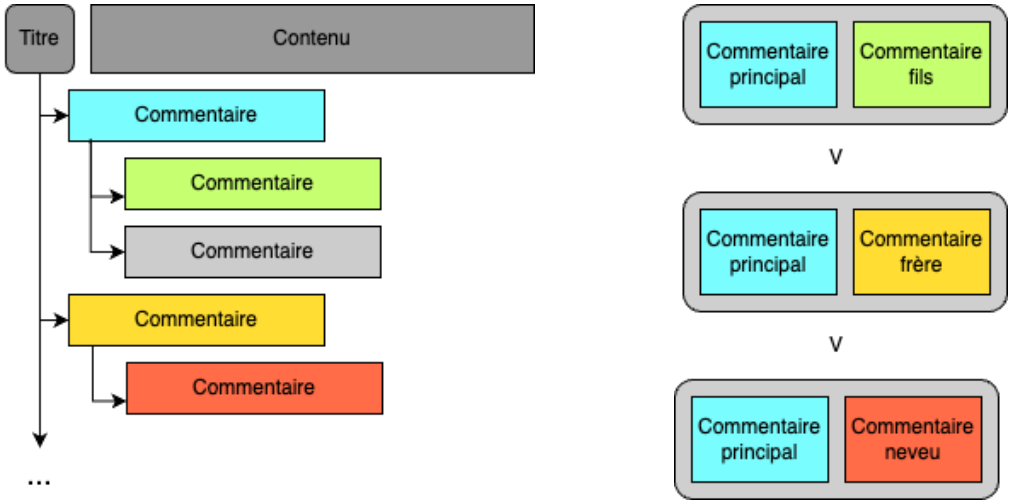


FIGURE 2 – Niveaux de similarité thématique entre les commentaires.

La structure de notre modèle est décrite dans Figure 3. Étant donné un commentaire C, nous récupérons un de ses fils, un de ses frères et un de ses neveux. Les quatre messages sont segmentés et les trois paires (C, fils), (C, frère) et (C, neveu) sont créées. Ces paires sont transmises à NSP-BERT dont la sortie est ensuite transmise à un perceptron multicouche. Nous utilisons la fonction d'activation linéaire rectifiée (ReLU) et un *dropout* de 10% entre les couches. La sortie finale est un score, ramené entre 0 et 1 avec une fonction sigmoïde, pour chacune des trois paires [3]

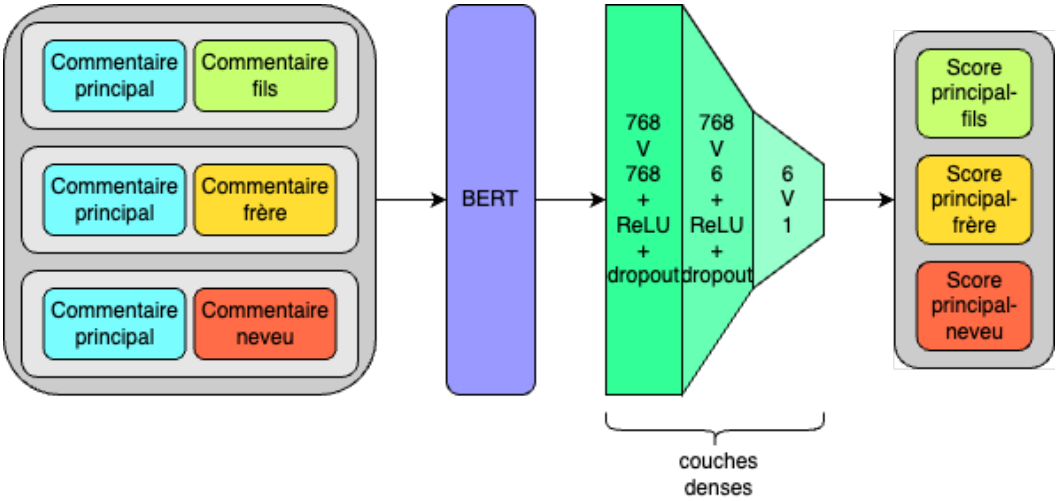


FIGURE 3 – Description du modèle.

3. L'ensemble des hyperparamètres est donné en Annexe B.

5 Expériences

Nous avons entraîné des modèles distincts pour chacun des cinq *subreddits* que nous avons récupérés. À chaque fois, l'ensemble des données est divisé en un ensemble d'entraînement et un ensemble de test dans les proportions 80%/20%. Afin de maintenir un équilibre concernant la taille des fils de discussion, nous les sélectionnons pour l'ensemble de test en les classant par taille, c'est-à-dire le nombre de commentaires qu'ils contiennent, et conservons un fil de discussion sur cinq. Pour la phase d'entraînement, nous extrayons également 10% de l'ensemble pour la validation.

5.1 Résultats

Comme expliqué ci-dessus, les données que nous utilisons pour former et évaluer notre modèle sont constituées de triplets de paires de messages (voir l'entrée du modèle sur la Figure 3). Pour construire ces triplets, nous sélectionnons au hasard un commentaire dans un fil de discussion et récupérons un de ses fils, un de ses frères et un de ses neveux. La plupart des commentaires ont plusieurs fils, frères et neveux. Par conséquent, pour maximiser la variabilité des données, nous ne générons pas tous les tuples possibles (commentaire, fils, frère, neveu), mais seulement l'un d'entre eux. Malgré cela, la quantité de commentaires dans notre ensemble de données nous permet de produire plus de triplets que nécessaire pour entraîner le modèle. Nous avons essayé trois tailles d'ensembles de données (1000, 7500 et 15000-triplets) pour l'entraînement afin de déterminer une quantité appropriée, en tenant compte du fait qu'un ensemble de données plus important nécessite plus de temps et de ressources pour entraîner notre modèle. Les meilleurs résultats ont été obtenus avec l'ensemble de données de 15000 triplets, mais l'amélioration par rapport à l'ensemble de données de 7500 triplets n'était pas suffisamment importante pour justifier l'utilisation d'un encore plus grand ensemble de données. Par conséquent, les résultats présentés dans la suite sont ceux que nous avons obtenus avec l'ensemble de données 15000-triplets.

Les résultats de nos expériences sont décrits dans le Tableau 1. Nous avons formé trois modèles par ensemble de données (c'est-à-dire 15 modèles au total). Nous les avons évalués sur tous les ensembles de données pour voir s'ils sont performants sur les messages d'autres *subreddits* qui diffèrent par leur forme (Choi *et al.*, 2015). Les résultats sont la moyenne et l'écart type des trois modèles. Le modèle utilisé est indiqué par le nom de la colonne et l'ensemble de données d'évaluation par celui de la ligne. Les résultats pour lesquels le modèle et l'ensemble de données sont congruents apparaissent sur la diagonale. Comme on pouvait s'y attendre, les résultats sont meilleurs dans ce cas (voir les résultats en gras dans Tableau 1), mais les résultats de l'application d'un modèle à un ensemble d'évaluation différent sont du même ordre (en restant inférieur).

La métrique obtenue est particulièrement sensible à la proximité thématique. Si elle varie entre 0 et 1, elle prend très fréquemment des valeurs proches des extremums se comportant de fait comme un bon classifieur thématique. Pour la suite, nous introduisons un seuil pour cette métrique que nous fixons à $0,5$ qui est une valeur classique et en adéquation avec nos observations.

5.2 Localiser les sous-*clusters* cohérents

Afin d'évaluer la qualité de la mesure obtenue, nous analysons la structure thématique des fils de discussion de *Reddit*. Nous avons donc appliqué notre métrique entraînée pour localiser des séquences

Data \ Model	Art	Coronavirus	gaming	OnePiece	sports
Art	**65.80 ± 0.17**	62.48 ± 0.68	63.72 ± 0.33	63.40 ± 0.51	63.71 ± 0.38
Coronavirus	66.01 ± 0.46	**73.13 ± 1.42**	67.62 ± 0.34	67.11 ± 0.70	67.42 ± 0.69
gaming	65.14 ± 0.59	64.42 ± 0.57	**66.65 ± 0.26**	65.04 ± 0.70	65.41 ± 0.39
OnePiece	66.01 ± 0.45	65.01 ± 0.24	66.06 ± 0.29	**69.95 ± 0.62**	65.77 ± 0.58
sports	64.91 ± 0.43	63.90 ± 0.57	65.07 ± 0.06	64.66 ± 0.45	**66.43 ± 0.42**

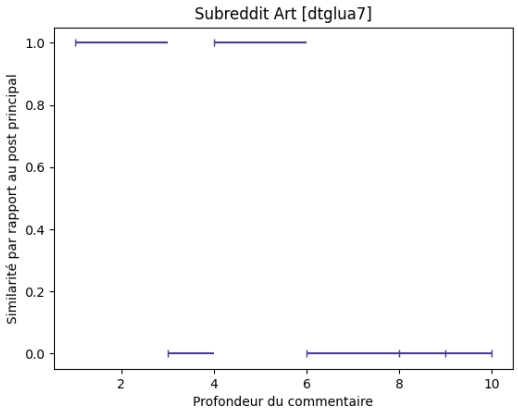
TABLE 1 – Résultats du transfert (moyenne ± écart type pour les trois versions de chaque modèle). *le gras indique les meilleurs modèles pour un ensemble d'évaluation.*

cohérentes de messages à l'intérieur d'un fil de discussion. Nous avons utilisé l'algorithme suivant :

— Dans une séquence donnée de commentaires, (sans regarder la structure horizontale des arbres de conversation,) nous calculons la similarité de chaque message avec son message suivant ;
— Chaque fois que la similarité est inférieure à un seuil $(0, 5)$, nous considérons que le message suivant est le début d'un nouveau sujet et qu'il contient le sujet dans son contenu ;
— Nous calculons la similarité entre le *post* initial et le premier commentaire de chaque nouveau *cluster* afin d'analyser l'évolution de la conversation.

Nous avons d'abord extrait toutes les séquences de commentaires de dix messages ou plus afin d'avoir des séquences assez longues pour observer des variations thématiques. La Figure 4 est une représentation de ce processus. Sur la Figure 4a, chaque ligne horizontale représente un *cluster* cohérent de commentaires et leur position verticale indique la similarité avec le *post* principal. Leur longueur dépend du nombre de messages. La Figure 4b reprend des exemples de commentaires ainsi que le *post* initial de la conversation. Nous voyons que le deuxième *cluster* est réduit au commentaire de profondeur 3 et est considéré comme thématiquement différent du *post* principal (*"They Don't Even Taste That Good Anymore (I), oil on canvas, 24x30""*). Cela a du sens car le commentaire ne contient que le mot *"Arrow"*. En revanche le *cluster* suivant est à nouveau proche du sujet d'origine, ce que nous confirmons avec le commentaire de profondeur 4 : *"It has to be good in the first place to stop being good, though."*. D'autres exemples seront donnés en annexe dans la version finale.

Le Tableau 2 présente différentes statistiques sur les séquences de messages et les *clusters*. Nous pouvons voir que le *subreddit Coronavirus* a peu de séquences que nous avons considérées comme suffisamment longues pour le regroupement. Tous les *subreddits* ont en moyenne un nombre similaire de messages par séquence, mais le *subreddit gaming* a quelques séquences très longues (au plus 275 messages) et les *subreddits Art* et *sports* ont également des séquences assez longues (environ 50 messages). En ce qui concerne l'évolution des thèmes, nous constatons qu'environ un tiers des premiers commentaires suivant un *post* sont proches de celui-ci, mais que près de 90% des têtes des



(a) Longueur et similarité au *post* initial des *clusters* identifiés.

Post principal : They Don't Even Taste That Good Anymore (I), oil on canvas, 24x30"
Commentaire de profondeur 3 : Arrow.
Commentaire de profondeur 4 : It has to be good in the first place to stop being good, though.
Commentaire de profondeur 9 : Fine, as long as you have a beard, like a proper dwarf.

(b) Commentaires en tête des *clusters* thématiques identifiés.

FIGURE 4 – Exemple de représentation d'une séquence de messages

derniers *clusters* sont thématiquement éloignées des *posts* d'origine. Cela semble soutenir l'idée que la conversation évolue naturellement au fur-et-à-mesure. Nous avons également calculé le nombre de fois où la similarité entre le *post* initial et la tête d'un *cluster* de commentaires est suffisamment différente de celle entre le *post* initial et la tête du *cluster* suivant. Ainsi nous mesurons la présence d'un changement thématique pour revenir au thème de départ ou s'en éloigner. Il y a en moyenne entre zéro et un changement de thème, ce qui nous indique que la conversation est soit considérée comme hors sujet par notre modèle dès le début, soit commence par être sur le sujet et évolue en s'éloignant du thème sans y revenir. Toutefois, ces structures ne sont pas les plus intéressantes. Les données contiennent également des séquences de commentaires comme ceux présentés dans la Figure 4 où un groupe proche du sujet principal est situé entre des groupes hors sujet. La structure de ces dialogues nous intéresse particulièrement car nous voulons suivre comment un sujet évolue dans une conversation. Pour mieux comprendre la structure thématique de nos séquences de messages, nous avons procédé à la modélisation globale de la conversation.

5.3 Hiérarchie de clusters

Pour aller plus loin dans notre analyse, nous avons regroupé les séquences cohérentes dans un fils de discussion. Pour cela, nous avons appliqué récursivement la méthode décrite précédemment mais sur les têtes des *clusters* au lieu de le faire sur chaque commentaire. L'algorithme est donc le suivant :

— Identifier les séquences cohérentes dans un fil avec l'algorithme de la Section 5.2 ;
— Calculer la similarité du premier message de chaque groupe avec celui du groupe suivant ;
— Si la similarité est inférieure à un seuil, le deuxième groupe est considéré comme le début d'un nouveau thème ;
— Application de la même stratégie pour le nouveau groupement jusqu'à stabilisation.

La Figure 5 est une illustration de ce processus. Nous pouvons mettre en évidence une hiérarchie entre les *clusters* thématiques. En effet, nous localisons d'abord les sujets avec une granularité élevée et chaque couche supplémentaire de regroupement est liée à des sujets avec une granularité plus générale, ce qui constitue une première étape vers une modélisation hiérarchique des thèmes.

Subreddit	Nb de séquences ($\geq$ 10 msgs)	Nb max	Nb moyen	Sim. initiale (<0,1 / >0,9)	Sim. finale (<0,1 / >0,9)	Nb de changements brutaux
Art	1556	51	11,88 (±2,96)	63% / 34%	87% / 11%	1,07 (±1,67)
Coronavirus	58	18	11,62 (±1,95)	69% / 28%	84% / 12%	0,55 (±0,90)
gaming	9409	275	11,88 (±5,44)	64% / 34%	88% / 10%	0,70 (±1,14)
OnePiece	908	28	11,84 (±2,48)	61% / 37%	91% / 07%	0,54 (±1,01)
sports	5930	52	11,71 (±2,58)	46% / 52%	85% / 14%	0,75 (±1,14)

TABLE 2 – Statistiques sur les séquences de messages et les *clusters* thématiques formés.

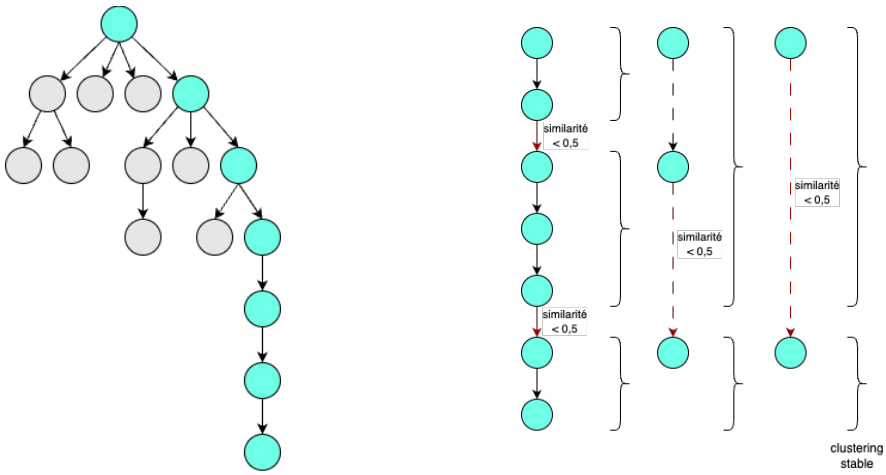
En suivant la méthode décrite précédemment, nous avons travaillé avec les séquences d'au moins 10 messages et calculé les différents niveaux de *clusters*. différentes statistiques sont reprises dans le Tableau 3. En ce qui concerne le nombre de niveaux de *clustering*, tous les *subreddits* nécessitent en moyenne trois niveaux. Le *cluster* de niveau le plus élevé contient en moyenne environ quatre *clusters* avec un écart-type élevé pour tous les *subreddits*, ce qui montre plus de variabilité sur cette question.

Subreddit	Nb moyen de niveaux de *clustering*	Nb moyen de *clusters* finaux
Art	3,03 (±1,34)	4,59 (±2,97)
Coronavirus	2,97 (±1,04)	3,86 (±2,54)
gaming	3,09 (±1,90)	4,41 (±3,23)
OnePiece	3,15 (±1,30)	4,72 (±2,90)
sports	3,14 (±1,33)	4,30 (±2,91)

TABLE 3 – Statistiques sur les hiérarchies de *clusters* formées.

6 Conclusion et Perspectives

Dans cet article, nous avons présenté une mesure de similarité thématique non supervisée. Nous avons montré qu'elle nous permet d'identifier des *clusters* de messages thématiquement cohérents au sein d'une conversation. Ce résultat correspond à la tâche de segmentation thématique linéaire.



(a) Extraction d'une séquence de commentaires.

(b) Regroupement selon la cohérence thématique à plusieurs niveaux.

FIGURE 5 – Processus de *clustering* hiérarchique selon les thèmes dans un fil de discussion.

Mais cette mesure nous permet également de construire une structure à plusieurs niveaux qui est une représentation hiérarchique des *clusters* thématiquement cohérents.

Notre mesure peut être utilisée pour une segmentation thématique précise puisque les morceaux de conversation seraient organisés logiquement en plus d'être séparés en fonction de leur sujet. À l'avenir, nous voulons enrichir notre représentation en étendant la représentation avec des relations rhétoriques de coordination et subordination entre *clusters*. Nous produirions une structure semblable à celles de la *Segmented Discourse Representation Theory* (SDRT, Lascarides & Asher (2007)) qui se focaliserait d'abord sur les liens thématiques. La comparaison de ces structures avec celles produites par des analyseurs de la SDRT (Li *et al.*, 2023) pourrait nous aider à comprendre l'influence des thèmes sur les représentations SDRT, et donner du sens aux changements de thème non adéquats.

Le regroupement que nous avons effectué était axé sur les séquences verticales de messages, en ce sens que nous avons extrait des séquences de messages plutôt que des sous-arbres complets de la conversation. La dimension horizontale doit également être explorée pour comparer l'évolution des sujets dans des fils de discussions parallèles et, par exemple, voir dans quelle mesure les conversations ont tendance à converger ou à diverger, et si ce comportement varie en fonction du *subreddit*. Cette application permettrait de mettre en avant dans l'ensemble d'un fil de conversation l'importance d'un thème donné, même si celui-ci apparaît peu dans chaque conversation. La récurrence transverse d'un thème permet d'identifier l'apparition de nouveaux sujets. Nous proposons également une structure d'analyse des échanges qui permet de mettre en avant des stratégies d'échange d'informations non explicite à la lecture des données. L'ajout d'un plus grand nombre de *subreddits* à notre ensemble de données nous permettrait d'effectuer des analyses sur une plus grande diversité de domaines, ce qui renforcerait la robustesse des résultats. Nous envisageons aussi d'entraîner de manière non supervisée et itérative la mesure pour gagner en précision. La sélection des messages pour l'entraînement ne serait pas laissée au hasard mais serait informée par la version de la métrique à l'itération précédente.

Enfin, contrairement aux modèles de classification binaires qui permettent seulement d'indiquer si deux segments sont thématiquement cohérents ou pas, notre modèle donne un score de proximité thématique. Cela pourrait permettre de repérer différents types de changements tels que des changements graduels qui ne sont pas repérables par classification binaire.

Références

BALOUCHZAHI F., BUTT S., SIDOROV G. & GELBUKH A. (2023). Reddit : Regret detection and domain identification from text. *Expert Systems with Applications*, **225**, 120099. DOI : https ://doi.org/10.1016/j.eswa.2023.120099.

BAUMGARTNER J., ZANNETTOU S., KEEGAN B., SQUIRE M. & BLACKBURN J. (2020). The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, **14**(1), 830–839. DOI : 10.1609/icwsm.v14i1.7347.

CHOI D., HAN J., CHUNG T., AHN Y.-Y., CHUN B.-G. & KWON T. T. (2015). Characterizing conversation patterns in reddit : From the perspectives of content properties and user participation behaviors. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*, COSN '15, p. 233–243, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/2817946.2817959.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

GAUR M., KURSUNCU U., ALAMBO A., SHETH A., DANIULAITYTE R., THIRUNARAYAN K. & PATHAK J. (2018). "let me tell you about your mental health !" contextualized classification of reddit posts to dsm-5 for web-based intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, p. 753–762.

HAMILTON W. L., YING R. & LESKOVEC J. (2017). Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 1025–1035, Red Hook, NY, USA : Curran Associates Inc.

HEARST M. A. (1997). Text tiling : Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, **23**(1), 33–64.

JOVANOVIC D. & LEEUWEN T. V. (2018). Multimodal dialogue on social media. *Social Semiotics*, **28**(5), 683–699. DOI : 10.1080/10350330.2018.1504732.

KATHIE TREEN, HYWEL WILLIAMS S. O. & COAN T. G. (2022). Discussion of climate change on reddit : Polarized discourse or deliberative debate ? *Environmental Communication*, **16**(5), 680–698. DOI : 10.1080/17524032.2022.2050776.

LASCARIDES A. & ASHER N. (2007). Segmented discourse representation theory : Dynamic semantics with discourse structure. In *Computing meaning*, p. 87–124. Springer.

LI C., HUBER P., XIAO W., AMBLARD M., BRAUD C. & CARENINI G. (2023). Discourse structure extraction from pre-trained and fine-tuned language models in dialogues. In A. VLACHOS & I. AUGENSTEIN, Éds., *Findings of the Association for Computational Linguistics : EACL 2023*, p. 2562–2579, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-eacl.194.

MISRA A. & WALKER M. (2013). Topic independent identification of agreement and disagreement in social media dialogue. In M. ESKENAZI, M. STRUBE, B. DI EUGENIO & J. D. WILLIAMS, Éds., *Proceedings of the SIGDIAL 2013 Conference*, p. 41–50, Metz, France : Association for Computational Linguistics.

NASEEM U., DUNN A. G., KIM J. & KHUSHI M. (2022). Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM Web Conference 2022*, p. 2563–2572.

TURCAN E. & MCKEOWN K. (2019). Dreaddit : A Reddit dataset for stress analysis in social media. In E. HOLDERNESS, A. JIMENO YEPES, A. LAVELLI, A.-L. MINARD, J. PUSTEJOVSKY & F. RINALDI, Éds., *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, p. 97–107, Hong Kong : Association for Computational Linguistics. DOI : 10.18653/v1/D19-6213.

WANG L., LI S., LV Y. & WANG H. (2017). Learning to rank semantic coherence for topic segmentation. In M. PALMER, R. HWA & S. RIEDEL, Éds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 1340–1344, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1139.

WANG W., HOI S. C. & JOTY S. (2020). Response selection for multi-party conversations with dynamic topic tracking. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6581–6591, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.533.

XING L. & CARENINI G. (2021). Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. In H. LI, G.-A. LEVOW, Z. YU, C. GUPTA, B. SISMAN, S. CAI, D. VANDYKE, N. DETHLEFS, Y. WU & J. J. LI, Éds., *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 167–177, Singapore and Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.sigdial-1.18.

XU P., SAGHIR H., KANG J. S., LONG T., BOSE A. J., CAO Y. & CHEUNG J. C. K. (2019). A cross-domain transferable neural coherence model. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 678–687, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1067.

A Description des subreddits utilisés dans notre dataset

Les descriptions ci-dessous sont toutes basées sur la description présente sur la page d'accueil des subreddits.

- Art : Une communauté où les utilisateurs parlent d'art et d'artistes, la description de la communauté insiste sur le fait de discuter d'art de façon mature et substantielle ;
- Coronavirus : Une communauté pour parler de la Covid-19, la description de la communauté demande des *posts* et des discussions de haute qualité ;
- gaming : Une communauté pour parler de jeux vidéos ;
- OnePiece : Une communauté pour discuter de toutes choses liées au manga One Piece de Eiichiro Oda et l'adaptation en anime ;
- sports : Une communauté pour parler de l'actualité sportive et de différentes ligues dans le monde comme la NBA.

B Hyperparamètres du model

— *Architecture* :
  — NSP-BERT
  — *Multi-layer Perceptron* ($768 \rightarrow 768$, $768 \rightarrow 6$, $6 \rightarrow 1$)
  — ReLU
  — *Dropout* 10% ;
— *Optimizer* : Adam ;
— *Learning rate* : 2e-5 ;
— *Epsilon* : 1e-8 ;
— *Batch size* : 16 ;
— *Epochs* : 5 ;
— *Output* : fonction sigmoïde (valeur entre 0 et1).

C Entraînement des modèles sur différentes tailles d'ensemble de données

TABLE 4 – Résultats (en %) pour différents nombres de triplets lors de l'entraînement

Subreddit	**1 000 triplets**	**7 500 triplets**	**15 000 triplets**
Art	59,25 ± 6,24	64,89 ± 0,23	65,80 ± 0,17
Coronavirus	67,42 ± 0,66	72,26 ± 0,66	73,13 ± 1,42
gaming	64,40 ± 0,04	65,34 ± 0,13	66,65 ± 0,26
OnePiece	65,39 ± 0,28	69,08 ± 0,43	69,95 ± 0,62
sports	64,22 ± 0,32	64,98 ± 0,31	66,43 ± 0,42

D Exemples de représentations de séquences de messages

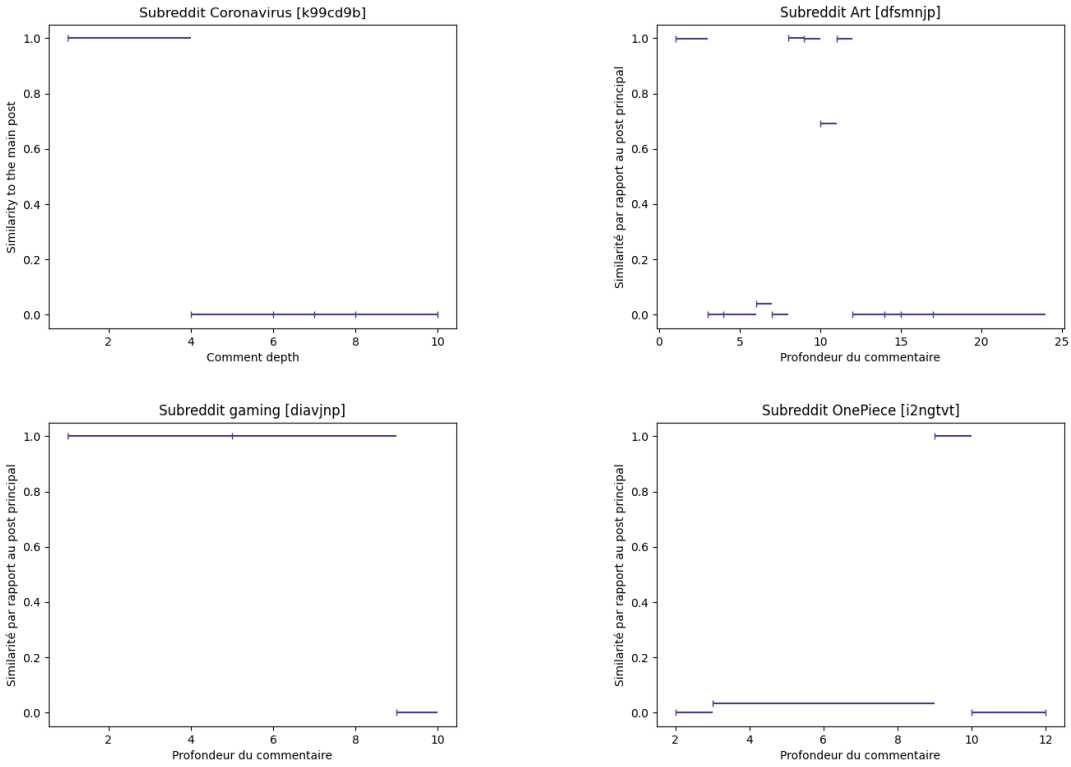


FIGURE 6 – Différentes structures thématiques.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

De nouvelles méthodes pour l'exploration de l'interface syntaxe-prosodie : un treebank intonosyntaxique et un système de synthèse pour le pidgin nigérian

Emmett Strickland[1, 2] Anne Lacheret-Dujour[1] Marc Evrard[2] Sylvain Kahane[1]
Dana Aubakirova[2] Dorin Doncenco[2] Diego Torres[2] Perrine Quennehen[1]
Bruno Guillaume[3]

(1) MoDyCo, 200 Avenue de la République, 92001 Nanterre, France
(2) LISN, Campus Universitaire, bât.507 Rue du Belvédère, 91405 Orsay, France
(3) LORIA, Campus Scientifique, 615 Rue du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy, France
{emmett.strickland, anne.lacheret, sylvain.kahane, perrine.quennehen}@parisnanterre.fr, marc.evrard@lisn.paris-saclay.fr, bruno.guillaume@inria.fr, {dana.aubakirova, dorin.doncenco, diego-andres.torres-guarin}@universite-paris-saclay.fr

RÉSUMÉ

Cet article présente deux ressources récemment développées pour explorer l'interface prosodie-syntaxe en pidgin nigérian, une langue à faibles ressources d'Afrique de l'Ouest. La première est un treebank intonosyntaxique dans laquelle chaque token est associé à une série de traits prosodiques au niveau de la syllabe, ce qui permet d'analyser diverses structures syntaxiques et prosodiques en utilisant une même interface. La seconde est un système de synthèse de la parole entraîné sur le même ensemble de données, conçu pour permettre un contrôle direct sur les contours intonatifs de la parole générée. Cet outil a été développé pour nous permettre de tester les hypothèses formulées à partir de l'exploration du treebank. Cet article est largement une adaptation de deux publications récentes présentant chaque outil, avec un accent sur leur interconnexion dans notre recherche en cours.

ABSTRACT

New methods for exploring the prosodic-syntactic interface : an intonosyntactic treebank and speech synthesis system for Nigerian Pidgin.

This paper presents two newly-developed resources for exploring the prosody-syntax interface in Nigerian Pidgin, a low-resource language of West Africa. The first is an intonosyntactic treebank in which tokens are associated with a series of syllable-level prosodic features, allowing for analyses of various syntactic and prosodic structures within the same interface. The second is a speech synthesis system trained on the same dataset which is designed to allow for direct control over the prosodic contours of generated speech. This tool was developed to allow us to test hypotheses made from the exploration of this treebank. This paper is largely an adaptation of two recent publications presenting these tools with an emphasis on their interconnectedness in our ongoing research.

MOTS-CLÉS : pidgin nigérian, linguistique de corpus, prosodie, treebank, synthèse de la parole.

KEYWORDS: Nigerian Pidgin, corpus linguistics, prosody, treebanks, speech synthesis.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 376–383.

1 Introduction et contexte

Le pidgin nigérian, ou naijá, est une langue créole peu dotée pourtant parlée par quelque 100 millions de locuteurs en Afrique de l'Ouest. Si l'essentiel de son lexique a été hérité de l'anglais, il a développé un ensemble de caracteristiques prosodiques et grammaticales qui le distingue de son lexifieur. Notamment, plusieurs analyses de cette langue postulent l'existence d'un système de ton lexical, même si ses caracteristiques exactes diffèrent selon l'analyse (Mafeni, 1971; Elugbe & Omamor, 1991; Faraclas, 1996).

Cette publication s'inscrit dans un projet de recherche en cours visant à mieux comprendre la prosodie du pidgin nigérien en utilisant une méthode novatrice qui combine la linguistique de corpus et la linguistique expérimentale. Concrètement, notre objectif est d'exploîter deux outils qui ont été produits parallèlement et qui permettent, respectivement, d'explorer des structures intonosyntaxiques dans un corpus de pidgin nigérian parlé, et de tester des hypothèses extraites de ce corpus dans un contexte expérimental. Cette approche permettra donc de découvrir de nouvelles phénomènes dans un corpus de parole spontannée, et de les valider grâce à des expériences perceptives contrôlées. Ces deux outils, le corpus NaijaSynCor-Prosody (Strickland *et al.*, 2024), et le système de synthèse NaijaTTS (Strickland *et al.*, 2023a), ont été développés simultanément comme des ressources complémentaires produites à partir des mêmes données. Les deux ressources ont déjà fait l'objet de deux publications, que cet article prolonge pour mettre en évidence leurs rôles complémentaires dans une méthodologie partagée.

2 NaijaSynCor-Prosody : Un treebank intonosyntaxique

Cette section présente le corpus NaijaSynCor-Prosody, une extension récente du corpus NaijaSynCor. Elle adapte une publication antérieure (Strickland *et al.*, 2024) qui a introduit ce corpus, et explique son rôle dans notre flux de travail expérimental plus large.

Le corpus NaijaSynCor-Prosody s'inscrit dans la continuité du projet ANR NaijaSynCor (Manfredi *et al.*, 2021), ainsi que du projet Rhapsodie sur le français parlé (Lacheret-Dujour *et al.*, 2019). Le projet NaijaSynCor a piloté le développement d'un corpus de 500 000 tokens transcrits à partir de 321 enregistrements de monologues et de dialogues se déroulant dans divers contextes sociaux. Un ensemble de 88 sessions d'enregistrement correspondant à environ 150 000 tokens répartis sur huit dialogues et 80 monologues ont été annotées manuellement en arbres de dépendance selon le schéma des *Surface Syntactic Universal Dependencies* (SUD) avant d'être converties en *Universal Dependencies* (UD, Gerdes *et al.*, 2018). Ce *gold standard* a ensuite été utilisé pour entraîner un analyseur syntaxique afin d'annoter automatiquement les fichiers restants (Guiller, 2020). Chaque session d'enregistrement transcrite est segmentée en unités illocutoires (UI) (Pietrandrea *et al.*, 2014) représentées sous forme d'arbres de dépendance encodés sous le format de données tabulaires CoNLL-U. Les UI sont également associées à différentes métadonnées, dont un identifiant numérique du locuteur, ce qui permet aux utilisateurs d'accéder à diverses informations sociolinguistiques telles que l'âge, le sexe, la profession et le niveau d'éducation. Se référer à Kahane *et al.* (2021) pour description plus détaillée de ce corpus et de son format.

En construisant le corpus NaijaSynCor-Prosody, notre objectif était de préserver les informations morphosyntaxiques du treebank NaijaSynCor original et d'ajouter une couche détaillée d'informations segmentales et suprasegmentales. Comme un treebank traditionnel, ce corpus augmenté permet aux

utilisateurs d'effectuer des études quantitatives de divers phénomènes morphosyntaxiques tels que ceux présentés par Courtin *et al.* (2018). Cependant, les utilisateurs auront désormais accès à diverses informations segmentales et suprasegmentales en plus des annotations originales.

Pour construire cette ressource, nous avons utilisé le logiciel d'alignement SPPAS (Bigi *et al.*, 2020) afin de produire un alignement phonétique `.TextGrid` des 80 monologues du corpus de référence, dont les transcriptions orthographiques avaient été soigneusement vérifiées au cours du projet NaijaSynCor. Ces alignements ont ensuite fait l'objet d'une segmentation en syllabes, l'unité porteuse de ton dans plusieurs analyses du pidgin nigérian. Les alignements syllabiques et les transcriptions phonétiques ont ensuite été vérifiées et corrigées manuellement par des annotateurs. Dans le cadre de notre système d'annotation, les phonèmes sont représentés à l'aide du format X-SAMPA (Wells, 1995). Ces transcriptions syllabiques constituent la base de l'annotation segmentale de ce corpus.

Nos annotations suprasegmentales consistent en une mesure de la fréquence fondamentale (F0) appliquée aux enregistrements audio des monologues et conservée au format `.PitchTier`. Les erreurs de pitchtracking ont été corrigées manuellement à l'aide du logiciel Analor (Lacheret & Victorri, 2002). Les alignements de chaque syllabe ainsi que leurs transcriptions phonétiques sont conservées au format `.TextGrid`. Ces deux formats de fichier ont ensuite été utilisés comme données d'entrée pour le logiciel de modélisation prosodique SLAM3 (Strickland *et al.*, 2023b), la version la plus récente du modèle SLAM (Obin *et al.*, 2014; Liu *et al.*, 2019). Pour chaque syllabe, SLAM3 a produit des étiquettes catégorielles décrivant les valeurs de hauteur de début et de fin de chaque syllabe, ainsi que les maxima de F0.

Plusieurs autres traits prosodiques ont également été extraits des étiquettes SLAM, notamment sa hauteur moyenne catégorielle et sa pente. Nous avons également produit un ensemble d'étiquettes continues en utilisant les fichiers `.PitchTier` et `.TextGrid`, notamment la durée (en ms) de chaque syllabe ainsi que sa F0 moyenne (en Hz)

Pour visualiser les relations entre les tokens et les syllabes, nous utilisons l'outil GREW-Match (Guillaume, 2021) avec un encodage CoNLL-U modifié pour prendre en compte les différents traits décrivant les syllabes. Cela permet d'observer les données syntaxiques et syllabiques via la même interface pour faire des observations ou tester des hypothèses linguistiques.

La figure 1 est un exemple de l'encodage utilisé. Chaque énoncé est représenté sous la forme d'un graphe avec deux types de nœuds : les nœuds de mots (en noir) et les nœuds de syllabes (en violet). Des arêtes spécifiques (bleues) sont utilisées pour relier les mots aux syllabes qui les composent. Ces arêtes sont étiquetées avec un numéro correspondant à la position de la syllabe correspondante dans le mot.

Les utilisateurs peuvent désormais quantifier un large éventail de phénomènes à l'interface de la syntaxe et de la prosodie à partir du treebank en utilisant la syntaxe GREW. À titre d'illustration, nous présenterons le cas de GO, qui peut fonctionner soit comme un auxiliaire marquant le futur, soit comme un verbe de mouvement. Les descriptions du pidgin nigérian les présentent généralement comme une paire minimale tonale, avec un ton bas marquant la forme auxiliaire et un ton haut marquant le verbe. Pour vérifier si cette distinction est représentée dans le corpus, nous utilisons la requête GREW suivante. [1]

```
pattern {
```

1. cette requête et les résultats peuvent également être visualisés ici : https ://naija.grew.fr/ ?custom=664205b96cefe

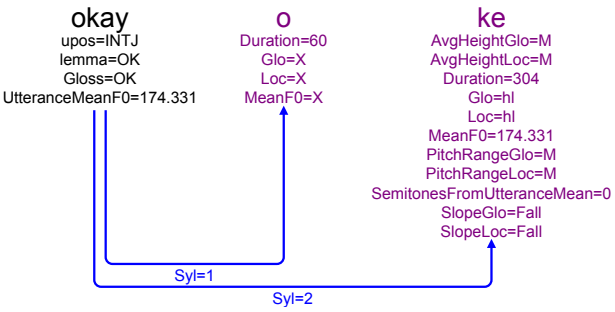


FIGURE 1 – Un token bisyllabique du corpus NaijaSynCor-Prosody

```
GO1 -[comp:aux]-> GO2;
GO1 [form="go"]; GO2 [form="go"];
GO1 -[Syl=1]-> S1; GO2 -[Syl=1]-> S2 }
```

Les deux premières lignes de cette requête localisent deux tokens (`GO1` et `GO2`) reliés par une relation `comp:aux` (complément auxiliaire), qui portent tous deux la forme GO. Le reste de la requête garantit que les deux tokens sont associés à des de syllabes distinctes (`S1` et `S2`), ce qui exclut les cas où les deux mots sont prononcés comme une seule syllabe fusionnée.

Les constructions résultantes peuvent ensuite être regroupées selon qu'elles remplissent ou non la condition `S2.MeanF0 > S1.MeanF0`, ce qui signifie une F0 plus élevée sur la syllabe associée au second GO. Étant donné que chaque paire représente deux mots prononcés par le même locuteur dans le même énoncé, nous pouvons comparer directement leurs valeurs moyennes de F0 en Hz. Ce test vise à identifier d'éventuelles différences de hauteur entre les deux parties du discours dans ce contexte particulier. Si, comme le décrit la littérature, le marqueur de futur doit avoir une hauteur inférieure au verbe de mouvement, on s'attendrait à ce que la forme auxiliaire soit plus basse que la forme verbale dans les contextes où les deux formes coïncident. Sur les 89 occurrences d'un GO verbal lié à un GO auxiliaire, 80 instances (89.9%) présentent une F0 plus élevée sur le verbe. Les résultats corroborent donc partiellement la littérature décrivant une différence de hauteur entre les deux utilisations.

Il est toutefois intéressant de noter qu'en modifiant cette requête pour comparer la durée relative de chaque instance de GO, on constate qu'en plus d'avoir une F0 plus basse, les auxiliaires sont majoritairement plus courts. Dans 73 cas sur 89 (82%), l'auxiliaire est plus court que le verbe. À notre connaissance, cette différence de durée n'a jamais été documentée dans les études sur cette langue. Une interprétation possible est que les syllabes qui portent un ton bas sont phonétiquement plus courts que celles qui portent un ton haut, mais que cette différence joue un rôle minimal dans la perception. Une autre interprétation est que la paire minimale examinée dans cet article n'est pas distinguée par le ton, mais par une combinaison d'une hauteur et d'une durée élevées. Dans une langue à accent lexical, comme l'anglais, les auxiliaires et d'autres grammèmes monosyllabiques ne seraient pas accentués, alors que les verbes et d'autres mots lexicaux portent toujours un accent. Si le pidgin nigérian est aussi une langue à accent lexical, les différences observées entre le GO grammatical et le verbe de mouvement seraient équivalentes à celles qui existent entre *butt* et le grammème non-accentué *but* en anglais.

Le simple test présenté dans cet article est loin d'être suffisant pour fournir des preuves solides pour ou contre cette analyse. Cependant, cela ouvre la voie à de nouvelles questions sur le rôle de la hauteur et de la durée dans la désambiguïsation de certains mots et constructions. Au fur et à mesure que nous explorons ce corpus, de telles hypothèses peuvent être validées par l'utilisation de l'outil présenté dans la section suivante, un système de synthèse vocale conçu pour des expériences perceptives.

3 NaijaTTS : vers un système de synthèse pour valider les hypothèses

Dans la section précédente, nous avons vu comment l'exploration de cette ressource peut mener à de nouvelles hypothèses sur le rôle de la durée et de la F0 en pidgin nigérian, et de la typologie prosodique de cette langue. Si d'autres tests révèlent une tendance généralisée et statistiquement significative dans l'ensemble du corpus, elle mérite d'être confirmée dans un cadre expérimental contrôlé. Cette section présente NaijaTTS, un système de synthèse (TTS) conçu pour tester les hypothèses faites pendant les explorations du treebank NaijaSynCor-Prosody. Cette section adapte Strickland *et al.* (2023a) et explique sa relation avec le corpus décrit dans la section précédente.

NaijaTTS a été entraîné sur les fichiers *gold* du corpus NaijaSynCor, correspondant à environ 7,5 heures de parole et 80 locuteurs. Pour garantir la fiabilité de nos données d'entraînement, nous avons limité ce projet à ces fichiers, car leur transcription et alignement syllabique ont été vérifiés par des annotateurs. Les données d'entraînement sont donc basées sur les mêmes fichiers sons et alignements utilisés pour produire le corpus NaijaSynCor-Prosody.

Nous avons basé notre système de synthèse vocale sur la plateforme FastSpeech 2 (FS2) (Ren *et al.*, 2022), un système de TTS neuronal de bout en bout non autorégressif entraîné directement sur des entrées `.wav`, dont il extrait des descripteurs de durée, de hauteur et d'énergie pour les utiliser lors de l'entraînement. L'architecture de FS2 inclut un encodeur qui convertit la séquence d'encastrement des phonèmes (*phoneme embedding sequence*) en une séquence cachée des phonèmes. Ensuite, un *variance adaptor* ajoute des informations relatives à la durée, la hauteur, et l'intensité prédits pour chaque phonème. Ensuite, cette séquence cachée est convertie en mel-spectrogrammes par un décodeur. FS2 a été choisi principalement pour le *variance adapter*. Nous intervenons dans cette partie du pipeline pour remplacer les valeurs prédites par nos propres vecteurs contenant une valeur de hauteur, de durée, et d'intensité pour chaque phonème. Cela permet un contrôle direct et précis de la prosodie au niveau phonémique.

Des efforts ont été déployés pour s'assurer que NaijaTTS était capable de produire un discours naturel et compréhensible dans un cadre expérimental. Cela s'est avéré être l'un des plus grands défis dans le développement de NaijaTTS, car les fichiers audio n'ont pas été enregistrés dans l'optique de la création d'un système de TTS. Les fichiers ont souvent été enregistrés dans de mauvaises conditions acoustiques, contiennent des bruits de fond importants, et les éconcés présentent de nombreuses dysfluences. En raison de ces facteurs, les premières itérations ont produit des sorties de mauvaise qualité, difficiles à comprendre et contenant de nombreux artefacts. De gains significatifs ont été faits grâce à un processus de filtrage de données visant à entraîner le modèle sur les fichiers enregistrés dans les meilleures conditions. Pour ce faire, nous avons brièvement écouté chaque session d'enregistrement et évalué leur qualité sonore sur une base informelle est purement perceptuelle. Nous avons ensuite ré-entraîné le modèle avc les enregistrements que nous avons jugés les meilleures.

En ce qui concerne les dysfluences, nous avons exploité les informations annotées dans le treebank pour exclure tous les énoncés contenant des pauses remplies, des réparations, des mots abandonnés ou des segments jugés incompréhensibles par les transcripteurs.

Ces contraintes ont réduit la taille de notre jeu de données à 3,7 heures de parole et 52 locuteurs. Cependant, malgré la taille réduite du corpus d'apprentissage, nous avons constaté une nette amélioration de la qualité de la parole générée. L'exclusion des dysfluences et des fichiers enregistrés dans de mauvaises conditions semblent avoir contribué de manière comparable à l'amélioration de la qualité de la parole synthétisée. Des gains significatifs en termes de qualité de la parole ont également été obtenus en affinant le vocodeur HiFi-GAN (Kong *et al.*, 2020) à l'aide des fichiers `.wav` utilisés pour l'entrainement du modèle.

Nous avons également veillé à ce que la parole de sortie ne corresponde pas aux voix des locuteurs figurant dans le corpus NaijaSynCor. Nous avons utilisé deux approches distinctes pour anonymiser notre modèle. La première a consisté à séparer les locuteurs par genre, afin de produire 2 modèles voix selon leur genre. Cette approche a permis de produire de voix masculines et féminines de haute qualité, bien que nous ayons remarqué que les caractéristiques vocales perçues changeaient parfois en fonction de l'énoncé généré. Nous soupçonnons que certaines séquences rares de phonèmes ont amené le modèle à adopter les caractéristiques des fichiers sonores dans lesquels ces séquences apparaissent de manière disproportionnée. Une approche prometteuse, inspirée par les travaux de Meyer *et al.* (2023) consistait à enregistrer brièvement un membre du projet et à extraire un embedding anonymisé en utilisant une méthode d'extraction de xvector fournie par Speechbrain (Ravanelli *et al.*, 2021), puis à le fournir directement au modèle en tant qu'entrée. Cela a produit une parole de haute qualité avec des caractéristiques vocales stables.

Dans les sections précédentes, nous avons vu que la F0 et la durée semblaient être des indicateurs fiables pour déterminer si une occurrence donnée de GO fonctionnait comme un marqueur de futur ou comme un verbe de mouvement. Cependant, il n'est pas clair si c'est la F0 seule ou une combinaison de F0 et de durée qui permet aux locuteurs du pidgin nigérian de distinguer ces deux utilisations. C'est un domaine dans lequel NaijaTTS peut être utilisé pour compléter les techniques exploratoires présentées dans la section précédente. Dans un contexte expérimental contrôlé, NaijaTTS nous permettrait de générer une série d'énoncés différant seulement par la hauteur et la durée assignées à GO. Nous pourrons donc voir si la hauteur est bien le seul paramètre qui permet aux locuteurs de distinguer les deux usages, ce qui suggérerait l'existence d'un système de ton lexical, ou si la durée est aussi perceptivement saillante, ce qui suggérerait plutôt un système d'accent lexical.

Plusieurs variantes de cette expérience peuvent être réalisées. L'une d'entre elles consisterait à utiliser des constructions syntaxiquement non ambiguës telles que *my father go go house* « mon père ira à la maison », qui sont clairement composées d'un verbe de mouvement précédé d'un marqueur de futur. Dans de tels cas, nous testerons si la hauteur et la durée contribuent de manière similaire à la perception du caractère naturel de l'énoncé. Si une hauteur plus basse mais une durée plus élevée sur le GO grammatical sont perçus comme naturelles, mais pas un hauteur plus élevée et une durée plus courte, cela indiquerait que la durée n'est pas aussi perceptivement saillante et que le pidgin nigérian pourrait être une langue à tons comme l'ont supposé des analyses antérieures. Cependant, si les deux contribuent de manière significative au caractère naturel perçu de l'énoncé, il se peut que la langue n'ait pas de ton lexical mais plutôt un système d'accent lexical. D'autres tests pourraient impliquer l'interprétation sémantique de constructions syntaxiquement ambiguës comme *my father go eat* qui peut signifier « mon père mangera » ou « mon père est parti manger » en fonction de la forme de GO utilisée. Lors de ces tests, les participants seraient interrogés sur leur interprétation de la phrase plutôt

que sur leur opinion quant à son caractère naturel. Dans ces cas, nous verrons si c'est la durée, la hauteur ou une combinaison des deux qui contribuent à l'interprétation du sens. Comme l'autre test, celui-ci permettrait de mieux comprendre si le pidgin nigérian utilise un système d'accent lexical, ce que les données du corpus suggèrent mais ne prouvent pas de manière définitive.

4 Conclusion

Dans cet article, nous avons présenté deux outils produits à partir des mêmes données, conçus pour remplir deux rôles dans une méthodologie scientifique partagée. NaijaSynCor-Prosody est un corpus novateur qui permet de faire des hypothèses sur l'intonosyntaxe en pidgin nigérian. NaijaTTS représente un système de synthèse conçu pour tester ces hypothèses dans un environnement expérimental contrôlé. Même si ces deux outils ont été spécifiquement développés pour le pidgin nigérian, cette méthodologie peut également être appliquée à d'autres langues si une quantité de données adéquate est disponible.

Références

BIGI B., ABIOLA O. S. & CARON B. (2020). Resources and tools for automated speech segmentation of the african language Naija (Nigerian Pidgin). In *Human Language Technology. Challenges for Computer Science and Linguistics : 8th Language and Technology Conference, LTC 2017, Poznań, Poland, November 17–19, 2017, Revised Selected Papers 8*, p. 164–173 : Springer.

COURTIN M., CARON B., GERDES K. & KAHANE S. (2018). Establishing a language by annotating a corpus. In *annDH 2018 Annotation in Digital Humanities*, volume 2155, p. 7–11 : CEUR.

ELUGBE B. O. & OMAMOR A. P. (1991). *Nigerian Pidgin : Background and Prospects*.

FARACLAS N. (1996). *Nigerian Pidgin*. Routledge.

GERDES K., GUILLAUME B., KAHANE S. & PERRIER G. (2018). SUD or Surface-Syntactic Universal Dependencies : An annotation scheme near-isomorphic to UD. In *Universal dependencies workshop 2018*.

GUILLAUME B. (2021). Graph matching and graph rewriting : Grew tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : System Demonstrations*, p. 168–175.

GUILLER K. (2020). Analyse syntaxique automatique du pidgin-créole du nigeria à l'aide d'untransformer (bert) : Méthodes et résultats.

KAHANE S., CARON B., STRICKLAND E. & GERDES K. (2021). Annotation guidelines of UD and SUD treebanks for spoken corpora. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, p. pp–35 : Association for Computational Linguistics.

KONG J., KIM J. & BAE J. (2020). Hifi-gan : Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, **33**, 17022–17033.

LACHERET A. & VICTORRI B. (2002). La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques. *Verbum : Analecta Neolatina*, **1**(24), 55–72.

LACHERET-DUJOUR A., KAHANE S. & PIETRANDREA P. (2019). *Rhapsodie : A prosodic and syntactic treebank for spoken French*, volume 89. John Benjamins Publishing Company.

LIU L., LACHERET-DUJOUR A. & OBIN N. (2019). Automatic Modelling and Labelling of Speech Prosody : What's New with SLAM+ ? In S. CALHOUN, P. ESCUDERO, M. TABAIN & P. WARREN, Éds., *International Congress of Phonetic Sciences (ICPhS)*, Melbourne, Australia : International Phonetic Association and Australasian Speech Science and Technology Association Australasian Speech Science and Technology Association Inc. HAL : hal-02119926.

MAFENI B. (1971). Nigerian pidgin. *The English Language in West Africa*, p. 95–112.

MANFREDI S., CARON B., GERDES K. & COURTIN M. (2021). Naijasyncor : a syntactic treebank, a parser and a wiktionary for naija. In *Summer Conference of the Society of Pidgin and Creole Linguistics*.

MEYER S., TILLI P., DENISOV P., LUX F., KOCH J. & VU N. T. (2023). Anonymizing speech with generative adversarial networks to preserve speaker privacy. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, p. 912–919. DOI : 10.1109/SLT54892.2023.10022601.

OBIN N., BELIAO J., VEAUX C. & LACHERET A. (2014). SLAM : Automatic stylization and labelling of speech melody. In *Speech prosody*, p. 246.

PIETRANDREA P., KAHANE S., LACHERET-DUJOUR A. & SABIO F. (2014). The notion of sentence and other discourse units in corpus annotation. *Spoken corpora and linguistic studies*, p. 331–364.

RAVANELLI M., PARCOLLET T., PLANTINGA P., ROUHE A., CORNELL S., LUGOSCH L., SUBAKAN C., DAWALATABAD N., HEBA A., ZHONG J., CHOU J.-C., YEH S.-L., FU S.-W., LIAO C.-F., RASTORGUEVA E., GRONDIN F., ARIS W., NA H., GAO Y., MORI R. D. & BENGIO Y. (2021). SpeechBrain : A general-purpose speech toolkit. arXiv :2106.04624.

REN Y., HU C., TAN X., QIN T., ZHAO S., ZHAO Z. & LIU T.-Y. (2022). Fastspeech 2 : Fast and high-quality end-to-end text to speech.

STRICKLAND E., AUBAKIROVA D., DONCENCO D., TORRES D. & EVRARD M. (2023a). NaijaTTS : A pitch-controllable TTS model for Nigerian Pidgin. In *ISCA Speech Synthesis Workshop*, Grenoble, France. HAL : hal-04183972.

STRICKLAND E., EVRARD M. & LACHERET-DUJOUR A. (2023b). SLAM 3 : An updated stylization model for speech melody. In *International Congress of Phonetic Sciences*, Prague, Czech Republic. HAL : hal-04171671.

STRICKLAND E., LACHERET-DUJOUR A., EVRARD M., KAHANE S., QUENNHEN P., EGBOKHARE F. & GUILLAUME B. (2024). New Methods for Exploring Intonosyntax : Introducing an Intonosyntactic Treebank for Nigerian Pidgin. In *LREC-Coling*, Turin, Italy.

WELLS J. C. (1995). Computer-coding the ipa : a proposed extension of sampa. *Revised draft*, **4**(28), 1995.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Étude des facteurs de complexité des modèles de langage dans une tâche de compréhension de lecture à l'aide d'une expérience contrôlée sémantiquement

Elie Antoine[1] Frédéric Béchet[1,4] Géraldine Damnati[2] Philippe Langlais[3]
(1) Aix-Marseille Université, CNRS, LIS
(2) Orange Innovation, DATA&AI, Lannion
(3) RALI/DIRO, Université de Montréal, Canada
(4) International Laboratory on Learning Systems (ILLS - IRL CNRS), Montreal
{first.last}@lis-lab.fr , {first.last}@orange.com,
felipe@iro.umontreal.ca

RÉSUMÉ

Cet article propose une méthodologie pour identifier des facteurs de complexité inhérents aux tâches de traitement automatique du langage (TAL), indépendamment de la dimension des modèles. Il montre que la performance inférieure de certains exemples est attribuable à des facteurs de complexités spécifiques. Plutôt que de procéder à des évaluations générales, nous préconisons des évaluations restreintes portant sur des tâches, des ensembles de données et des langues spécifiques, décrites de manière linguistique. Appliquée à une tâche de compréhension de texte via un corpus de questions-réponses, notre méthode met en évidence des facteurs de complexité sémantique affectant divers modèles de tailles et d'architectures différentes. En outre, nous proposons plusieurs corpus de complexité sémantique croissante dérivés de ces facteurs, avançant que l'optimisation de leur traitement dépasse la simple augmentation de la taille des modèles.

ABSTRACT

Investigating the complexity factors of language models in a reading comprehension task using a semantically controlled experiment

The paper introduces a methodology focusing on identifying complexity factors specific to natural language processing (NLP) tasks, independent of model size. It suggests that certain examples consistently yield lower scores regardless of model size due to specific complexity factors. Rather than broad evaluations, the methodology advocates for constrained evaluations on specific tasks, datasets, and languages, described linguistically. The study demonstrates this approach on a reading comprehension task using a corpus of question-answer pairs. It proposes and validates semantic complexity factors affecting models of different sizes and architectures. Additionally, it defines multiple corpora of increasing semantic complexity derived from these factors. The study argues that improving the processing of these corpora requires more than just increasing model parameters.

MOTS-CLÉS : Compréhension de texte, Question/Réponse, Grand modèle de langage, Annotation sémantique, FrameNet.

KEYWORDS: Text Understanding, Question Answering, LLM, semantic annotation, FrameNet.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 384–396.

1 Introduction

Les modèles de langage génératifs constituent actuellement l'état de l'art pour presque toutes les tâches de traitement du langage naturel, en particulier grâce au développement de grands modèles de langage, dont le nombre et la taille ne cessent de croître. Toutefois, malgré l'intense activité de recherche les concernant, de nombreuses zones d'ombre demeurent quant à leurs capacités, limites, et risques. L'étude académique de ces modèles est entravée par des défis significatifs : les modèles "ouverts" nécessitent des ressources considérables non accessibles à tous, tandis que l'analyse des modèles "fermés", accessibles via des API, est limitée par des coûts potentiellement élevés et un manque de transparence sur leur fonctionnement et apprentissage.

Cette réalité encourage l'examen de modèles plus petits, dont la gestion, le développement, et l'analyse sont simplifiés. Cependant, cette approche est freinée par le risque que les améliorations apportées ne soient pas transférables à des modèles possédant plus de paramètres, posant ainsi la question de leur impact réel sur le développement futur.

Pour remédier à cette limitation, nous présentons dans cet article une approche axée sur l'identification de facteurs de complexité inhérents à des tâches spécifiques, indépendamment de la taille des modèles développés. Cette méthode postule que peu importe leur capacité individuelle, ces modèles sont uniformément influencés par ces facteurs de complexité.

Pour obtenir ces facteurs, au lieu d'effectuer des évaluations "boîte noire" couvrant de nombreuses tâches, jeux de données et langues (Liang *et al.*, 2023; Srivastava *et al.*, 2023) nous préconisons des analyses précises, confinées à des contextes spécifiquement définis et décrits sous un angle linguistique. Cette méthode permet de comparer des modèles de différentes tailles et cherche à révéler les facteurs de complexité linguistique inhérents aux capacités de résolution de tâches de chaque modèle.

Cet article présente un cas d'application de notre méthode à une tâche de compréhension de texte, en utilisant le corpus français Calor (Béchet *et al.*, 2019), qui comprend des paires Question-Réponse (QR) enrichies d'annotations sémantiques fines détaillant la relation entre questions et réponses. En analysant les résultats de différents modèles de tailles diverses sur ces paires et en tenant compte des annotations sémantiques, nous visons à identifier des facteurs de complexité pertinents pour tous les modèles.

Les principales contributions de cette étude consistent à proposer et à valider des facteurs de complexité sémantique qui ont un impact négatif sur des modèles d'architectures et de tailles diverses. En outre, l'étude définit des corpus de complexité sémantique croissante dérivés de ces facteurs, obtenus par des partitions du corpus Calor . Nous soutenons que l'amélioration du traitement de ces corpus nécessite plus qu'une simple augmentation du nombre de paramètres du modèle.

2 Un corpus de compréhension de texte sémantiquement contrôlé

L'étude de la tâche de question-réponse à partir de documents a été largement étudiée avec l'émergence des modèles de réseaux neuronaux profonds, stimulée par l'accès à d'importants corpus d'évaluation tels que SQuAD (Rajpurkar *et al.*, 2016) ou MultiRC (Khashabi *et al.*, 2018), qui fait partie du benchmark SuperGLUE (Wang *et al.*, 2019). Dans ces classements, les modèles de langage basés sur des Transformers sont aujourd'hui systématiquement plus performants que les humains. Cela

souligne la difficulté d'évaluer les modèles en raison de la nature subjective de la génération des réponses, tout en démontrant la capacité des modèles à capturer et à reproduire les biais inhérents aux données sur lesquelles ils sont entraînés.

Cette tâche présente plusieurs avantages malgré ses contraintes : (1) Elle facilite la comparaison directe entre modèles classificateurs et génératifs dans un contexte unifié, où les classificateurs, bien que moins complexes, parviennent à égaler les performances des génératifs ; (2) Elle peut être partiellement décrite par un modèle linguistique formel détaillant les liens syntaxiques et sémantiques entre questions et réponses, contrairement aux tâches exclusivement génératives comme le résumé ou la traduction, plus difficiles à formaliser linguistiquement.

Dans cette étude, nous utilisons le corpus Calor contenant des paires question/réponse enrichies d'annotations sémantiques selon le modèle Berkeley Framenet (Baker *et al.*, 1998). Ce corpus, destiné à l'origine à l'évaluation de la génération de questions, rassemble des textes en français issus de Wikipedia et de ClioTexte [1], un ensemble de documents historiques utilisés dans l'enseignement en France, couvrant principalement trois sujets : la Première Guerre mondiale, l'archéologie, et l'antiquité, avec des sources variées, depuis les documents historiques de ClioTexte jusqu'aux articles encyclopédiques de Wikipedia.

Les annotations s'articulent autour de *cadres sémantiques* (***Frames***), définissant des scénarios prototypiques (tels que *décider*, *perdre*, *attaquer*, *vaincre*, etc.). Le processus d'annotation consiste d'abord à identifier l'*Unité Lexicale ou Lexical Unit (**LU**)* déclenchant le Cadre, suivie par celle des éléments constitutifs du cadre, ou *Frame Elements (**FE**)*.

Un exemple est donné dans la figure 1 pour une phrase annotée avec les deux *Frames*, `Losing` déclenché par le mot `perdu` et `Attack` déclenché par `assauts`.

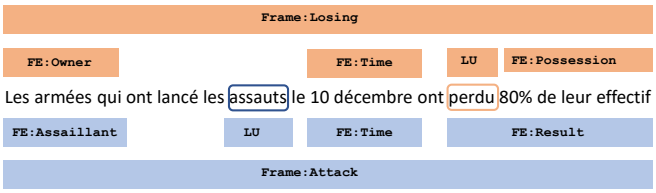


FIGURE 1 – Exemple de phrase annotée avec les cadres sémantiques FrameNet

La création d'un corpus de *questions/réponses* structuré sémantiquement a été réalisée selon la méthode suivante : initialement, une sélection aléatoire d'un *cadre* sémantique F et d'un *élément de cadre* E est effectuée dans une phrase spécifique. Des annotateurs humains sont ensuite chargés de formuler une question dont E constitue la réponse, exploitant pour cela la relation sémantique existant entre F et E. Les autres *FE* présents dans la phrase servent de contexte C à la question, offrant ainsi une contextualisation supplémentaire. Par la variation des éléments F, E, et C au sein d'une même phrase, il est possible de générer un large éventail de questions, accompagnées de leurs réponses et catégorisées selon leur classe sémantique correspondant au type du *FE*.

Prenons pour exemple la phrase mentionnée dans la Figure 1. Avec F = `Perdre`, E = `Propriétaire`, et C = `Temps, Possession`, il est possible pour les annotateurs de formuler des questions comme :

— *Qui a perdu la majorité de ses troupes le 10 décembre ?*

1. https://clio-texte.clionautes.org/

— *Quelles armées ont été décimées lors des attaques du 10 décembre ?*

Pour encourager une diversité lexicale dans la formulation des questions, la phrase source n'était pas révélée aux annotateurs. Cette approche leur donnait la latitude de choisir librement les mots lorsqu'ils construisaient les questions et de décider des éléments contextuels C à intégrer. Lorsque les réponses exigeaient une analyse de coréférences, les chaînes de coréférence menant à la réponse appropriée étaient annotées. Dans l'exemple mentionné précédemment, la réponses (`les armées`) nécessitaient une résolution de coréférence pour déterminer la réponse exacte, bien que l'annotation des cadres se fasse au niveau de la phrase. Il en résulte que répondre correctement pourrait demander d'extraire des informations au-delà de la phrase donnée, dans d'autres sections du texte. Cette méthodologie a permis d'enrichir le corpus Calor de **1821** questions issues de 54 cadres sémantiques différents.

3 Modèles de langage pour la compréhension de la lecture

Dans cette étude, nous comparons six modèles de langage pré-entraînés sur notre tâche de question-réponse en utilisant le corpus Calor . Parmi ces modèles, l'un est un modèle de classification basé sur une architecture BERT développée pour la langue française, CamemBERT (Devlin *et al.*, 2019; Martin *et al.*, 2020). Trois d'entre eux sont des modèles génératifs multilingues basés sur **T5** (T5-LARGE, FLAN-T5-LARGE (Wei *et al.*, 2021), MT5-LARGE (Xue *et al.*, 2021)), et les deux autres sont des grand modèle de langage (LLM) actuels : LLAMA2 (Touvron *et al.*, 2023), Mixtral 8x7B (Jiang *et al.*, 2024) et chatGPT-3.5 [2].

Tous ces modèles pré-entraînés, à l'exception de GPT3.5, ont été affinés sur notre tâche de question-réponse (*QR*) en utilisant le corpus français FQuAD (d'Hoffschmidt *et al.*, 2020). Ce corpus, construit de manière similaire à SQuAD (Rajpurkar *et al.*, 2016), contient des questions basées sur des documents Wikipédia en français. Il convient de noter que si la plupart des textes du corpus Calor proviennent également de Wikipédia, Calor est nettement plus difficile que FQuAD. Cette différence s'explique par le fait que les textes de Calor sont spécialisés et que les réponses annotées sont significativement plus longues, correspondant aux éléments de cadre de nos annotations sémantiques.

Pour l'évaluation sur le corpus Calor , nous avons harmonisé les formats des corpus FQuAD et Calor , facilitant ainsi l'évaluation directe des systèmes affinés avec FQuAD sur Calor . Cette unification a permis d'appliquer un affinage spécifique à chaque modèle en vue de leur évaluation. Pour CamemBERT et les variantes de T5, nous avons procédé à un affinage sur le corpus FQuAD pendant deux époques. Concernant LLAMA2, nous avons employé la méthode Low-Rank Adaptation (LoRA), tandis que pour GPT-3.5 et Mixtral 8x7B, une approche d'amorçage à respectivement un et deux exemples a été utilisé. Cette dernière consiste à fournir au modèle un exemple d'entrée et de sortie dans le format désiré, lui indiquant explicitement de se concentrer sur l'extraction de contenu à partir du document source.

Les performances de ces sept modèles ont été évaluées sur le corpus Calor . Pour comparer l'efficacité des modèles extractifs (comme CamemBERT) et abstractifs (les autres modèles considérés), la métrique ROUGE-L, via l'outil ROUGE [3] a été utilisé. Une évaluation humaine a également été menée, nous n'indiquons ici que le pourcentage de réponses annotées comme correctes. Les résultats

2. API de `https://chat.openai.com`
3. Nous utilisons l'implémentation de google research disponible ici

obtenus, illustrés dans le tableau 1, montrent les scores moyens en ROUGE-L et en pourcentage de réponses correctes via l'annotation manuelle pour l'ensemble des 1821 questions composant le corpus.

Model	type	adapt	#param	Rouge-L	% réponses correctes
CamemBERT	classif.	FT	335M	0.82	78
T5-LARGE	gene.	FT	738M	0.81	77
FLAN-T5-LARGE	gene.	FT	783M	0.80	79
MT5-LARGE	gene.	FT	1.2B	0.80	77
LLAMA-2	gene.	LoRA	7B	0.69	72
Mixtral-8x7b	gene.	prompt	47B	0.80	82
GPT 3.5	gene.	prompt	175B	0.72	82

TABLE 1 – Description des 7 modèles de langage utilisés dans nos expériences avec le score ROUGE-L moyen et le pourcentage de réponses annotées comme correctes sur le corpus d'évaluation

Dans l'ensemble, les performances des différents modèles sur notre corpus sont considérablement plus basses par rapport à celles rapportées dans des tâches analogues telles que SQuAD [4] ou MultiRC dans SuperGLUE [5]. Cet écart peut être attribué en partie aux caractéristiques du corpus Calor , aux différences entre le corpus d'affinage FQuAD et le corpus d'évaluation Calor comme discuté précédemment, et aussi à l'absence d'optimisation du modèle via hyperparamétrage.

Les scores ROUGE-L des modèles de génération basés sur T5 et du modèle de classification basé sur CamemBERT sont relativement proches, alors que ceux de deux des grands modèles de langage, LLAMA-2 et GPT3.5, sont significativement en retard par rapport aux autres modèles. Ces résultats ne sont pas inattendus, étant donné que la tâche de question-réponse employée dans cette étude favorise les modèles extractifs par rapport aux modèles génératifs. Ce biais provient du fait que les références dans le corpus d'évaluation Calor sont extractives, comprenant des segments du texte original. Par conséquent, les métriques ROUGE favorisent intrinsèquement les modèles qui ne font que reproduire les segments sans introduire de mots supplémentaires.

Il est donc essentiel d'interpréter avec prudence les comparaisons directes de scores entre modèles basées sur cette métrique. L'intention de cette étude est moins de juger de la supériorité d'un modèle sur un autre en termes de performances brutes que d'explorer les facteurs linguistiques qui affectent les résultats de chaque modèle, indépendamment de leur efficacité absolue.

L'étude initiale que nous avons menée visait à examiner la corrélation potentielle entre le type de relation sémantique liant une question et sa réponse, et la performance des modèles. Pour discerner ces relations sémantiques, nous avons utilisé les cadres sémantiques (**Frames**) qui ont formé la base de l'élaboration des questions, comme détaillé dans le paragraphe 2. En segmentant le corpus Calor en 54 sous-corpus correspondant au nombre de cadres présents au total dans notre corpus, nous avons pu évaluer la performance de chaque modèle en fonction du cadre considéré.

S'il n'y avait pas de corrélation entre la performance et la relation sémantique, la variance des scores entre les sous-corpus devrait être minimale. Cependant, nos résultats indiquent le contraire.

Ceci est illustré dans la figure 2 où la distribution des scores de GPT-3.5 pour chaque sous-corpus de Frame est représentée. Comme on peut le voir, cette distribution n'est pas uniforme, validant l'hypothèse que la performance du modèle est liée aux relations sémantiques sous-jacentes. Ce même

4. https://rajpurkar.github.io/SQuAD-explorer
5. https://super.gluebenchmark.com/leaderboard

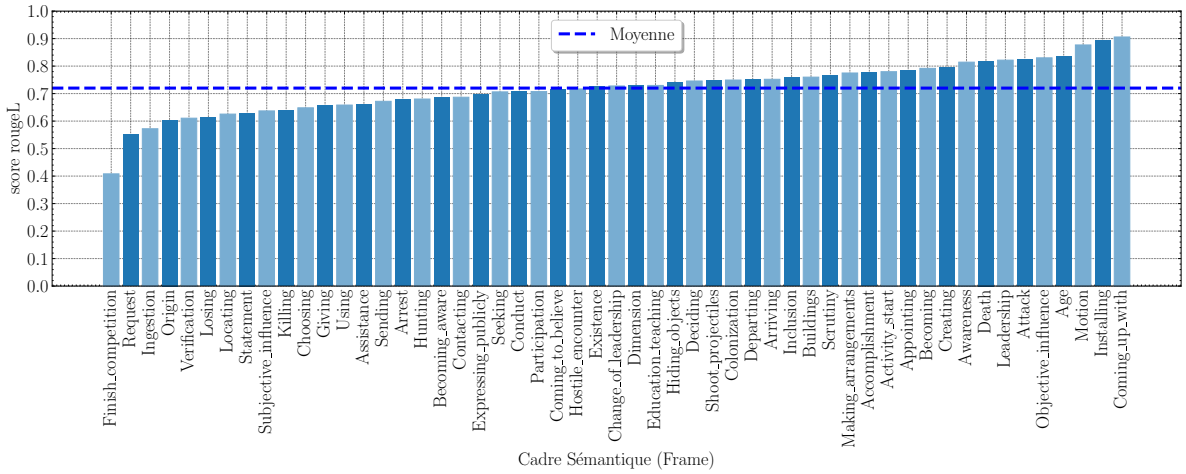


FIGURE 2 – Performance du modèle GPT-3.5 pour chaque Frame triée par la mesure rouge-L

type de distribution se retrouve pour tous les modèles, même si les mêmes Frames ne se retrouvent pas toujours dans les mêmes positions. Cette distribution se retrouve également dans l'évaluation humaine des réponses produites par les modèles. Dans tous les cas, le type de relations sémantiques a une influence significative sur la performance de tous les modèles. Dans la section suivante, nous proposerons différents facteurs de complexité pouvant expliquer ce phénomène.

4 Facteurs de complexité sémantique

Nous avons vu dans la section précédente que les performances des différents modèles de question-réponse n'étaient pas uniformes sur l'ensemble des cadres sémantiques. Cette section introduit une méthodologie destinée à identifier des facteurs de complexité sémantique susceptibles de justifier ces variations de performances. Nous étudierons aussi dans quelle mesure ces facteurs sont applicables de manière générale aux différents modèles analysés.

Méthode

1. Nous formulons plusieurs hypothèses concernant les facteurs de complexité sémantique ($F = f_1, f_2, \ldots$) sous forme de questions binaires applicables aux exemples du corpus d'évaluation. Par exemple : *La recherche de la réponse nécessite-t-elle de résoudre une chaîne de coréférence ?*
2. Pour chaque facteur f, nous divisons le corpus d'évaluation en deux sous-ensembles E_f et $\bar{E}f$ contenant respectivement les exemples répondant "oui" (supposés "difficiles") et "non" (supposés "plus faciles") à la question posée par le facteur f. Lorsqu'un seuil numérique est nécessaire, nous sélectionnons une valeur pour répartir le plus équitablement le corpus.
3. Pour chaque facteur f et modèle m, nous calculons la performance ROUGE-L du modèle m sur les partitions Ef et $\bar{E}f$: $RL(m, Ef)$ et $RL(m, \bar{E}f)$. Ensuite, nous calculons : $\delta = \lfloor (RL(m, Ef) - RL(m, \bar{E}f)) * 100 \rfloor$ représentant la dégradation en termes de points de ROUGE-L due au facteur de complexité f.

4. Enfin, nous calculons une mesure de significativité statistique pour δ avec le test U de Mann-Whitney avec un niveau de risque de 5% entre les deux partitions Ef et $\bar{E}_f$. Ce test prend en compte la valeur de δ et les caractéristiques de chaque ensemble dans la partition.

Facteurs de complexité sémantique. Pour cette étude, nous avons classé les facteurs en deux catégories : les facteurs génériques indépendants des relations sémantiques ($f0$ et $f1$) et ceux qui dépendent des relations sémantiques ($f2$ à $f6$). Pour ces derniers, nous avons été guidés par certains facteurs de complexité proposés pour l'analyse syntaxique automatique dans Frames dans (Marzinotto *et al.*, 2018).

$f0$: **biais dans le corpus d'adaptation.** La première explication pourrait être l'association entre la distribution des cadres sémantiques dans le corpus d'adaptation et les scores du modèle dans le corpus d'évaluation. Malgré l'absence de corrélation directe entre ces corpus, une grande quantité de certaines relations sémantiques dans le corpus FQuAD pourrait être corrélée avec la performance du modèle sur des questions/réponses présentant des relations similaires dans le corpus d'évaluation. Pour ce faire, nous avons automatiquement annoté le corpus FQuAD avec des cadres et classé les cadres en fonction de leur fréquence. Ensuite, nous divisons les exemples FQuAD en deux sous-ensembles, afin d'obtenir une représentation équilibrée en termes de fréquence des cadres : E_f comprend les exemples correspondant aux cadres les moins fréquents, tandis que $\bar{E}_f$ comprend le reste.

$f1$: **coréférence.** La nécessité de résoudre une coréférence constitue un facteur de complexité potentiel. Comme mentionné dans la section 2, les chaînes de coréférence sont annotées pour les arguments des relations sémantiques liant les questions et les réponses, ce qui nous permet de diviser le corpus en deux : les exemples nécessitant la résolution d'une chaîne de coréférence pour trouver la réponse E_f, et les autres $\bar{E}_f$.

$f2$: **nature de la relation sémantique du déclencheur.** Les déclencheurs d'un cadre sémantique dans le modèle FrameNet, appelés *Unité lexicale - LU*, peuvent être verbaux ou nominaux. Il a été démontré dans (Marzinotto *et al.*, 2018) que les relations déclenchées par une LU nominale sont plus difficiles à traiter. Nous avons donc divisé les exemples du corpus d'évaluation en fonction de la nature de la LU : soit nominale E_f, soit verbale $\bar{E}_f$.

$f3$: **présence du déclencheur du cadre sémantique dans la question.** Lorsque le même terme déclenche la relation sémantique dans le texte et apparaît dans la question, l'établissement d'un lien entre la question et la réponse est évidemment plus simple. Pour tenir compte de ce facteur, nous classons les exemples dans le sous-ensemble E_f lorsque le déclencheur est absent de la question et de la réponse, et dans $\bar{E}_f$ lorsque l'unité lexicale (LU) (ou l'une de ses inflexions) est présente dans les deux.

$f4$: **Distance entre le déclencheur et la réponse en termes d'arcs de dépendance.** La distance entre le déclencheur du cadre et la réponse dans le texte peut constituer un facteur de complexité, étant donné qu'une plus grande distance tend à être corrélée à un plus grand nombre d'attracteurs. Pour quantifier ce facteur, nous calculons la distance en termes d'arcs de dépendance à l'aide d'une

	Facteur de complexité						
modèles/facteurs	f0	f1	f2	f3	f4	f5	f6
proportion de E_f (%)	42%	6%	37%	45%	12%	59%	46%
CamemBERT	-1	-4	-1	-2	-7	-3	-1
T5	-1	-9	-2	-1	-7	-5	-2
FLAN	-2	-4	-3	-2	-4	-5	-3
MT5	0	-13	-1	-1	-10	-4	-2
llama-2	0	-3	-1	3	-3	-7	-2
mixtral-8x7b	0	1	-2	-1	-5	-6	0
GPT-3.5	0	4	-1	0	-4	-4	-3

TABLE 2 – Validation des facteurs de complexité : chaque cellule montre δ pour chaque modèle et facteur, avec des marquages jaunes pour différences significatives. La ligne "proportion" indique le pourcentage de chaque partition E_f dans le corpus total.

analyse syntaxique du corpus. Nous regroupons les exemples pour lesquels il existe au moins deux arcs de dépendance entre l'unité lexicale (LU) et la réponse dans le sous-ensemble E_f, et ceux pour lesquels il n'existe qu'un seul arc dans $\bar{E}_f$.

$f5$: **Nombre d'arguments dans le cadre.** Certaines relations sémantiques ont un plus grand nombre d'arguments (Frame Elements - FEs) que d'autres. La quantité d'éléments de cadre dans la relation sémantique sous-jacente à la paire question-réponse est également un facteur influençant la difficulté : un plus grand nombre d'éléments de cadre implique que le processus de liaison dispose de plus de contexte pour identifier avec précision la réponse. À l'inverse, un nombre réduit d'arguments rend la tâche plus ambiguë. Nous regroupons les exemples ne comportant pas plus de deux FE dans le sous-ensemble E_f, et ceux comportant plus de deux arguments dans $\bar{E}_f$.

$f6$: **Mesure de l'entropie de la distribution des LUs pour un cadre donné.** Certains cadres sont systématiquement déclenchés par les mêmes termes, tandis que d'autres présentent une diversité beaucoup plus grande, ce qui entraîne une ambiguïté dans leurs déclenchements. Cette mesure de la "surprise" peut être quantifiée par l'entropie de la distribution des LU dans le corpus d'entraînement. Une entropie plus élevée indique une plus grande ambiguïté dans le déclenchement des cadres. Nous incluons des exemples dans le sous-ensemble E_f pour les cas dont la valeur d'entropie est supérieure à un seuil α, et dans $\bar{E}_f$ pour les cas inférieurs au même seuil. Le seuil α est calculé comme la valeur moyenne de l'entropie sur l'ensemble du corpus.

5 Validation expérimentale

Le tableau 2 présente les résultats pour ces 7 facteurs de complexité. Dans chaque case, pour un modèle m et un facteur f, la valeur correspond à l'impact de f sur m exprimé par la différence en termes de points ROUGE-L δ présentée précédemment. Les cases colorées en jaune correspondent aux facteurs qui ont validé le test U de Mann-Whitney pour la significativité statistique avec un risque de 5

Comme on peut le voir, le facteur générique $f0$ correspondant au lien entre la fréquence d'une Frame

dans le corpus d'adaptation et dans le corpus d'évaluation a très peu d'influence sur les résultats.

En ce qui concerne $f1$, on constate que le fait de devoir résoudre une chaîne de coréférence est un facteur de complexité pour tous les modèles sans être significatif, mais qu'il n'a principalement un impact que pour les "petits" modèles. Dans les faits, même s'il existe une certaine perte de performance avec les LLMs, celle-ci est moins importante qu'avec les autres modèles, voir même un gain significatif dans le cas de GPT-3.5. Ceci suggère que les LLMs, par leur taille, ont une bien meilleure capacité à gérer les coréférences, au moins celles considérées dans cette tâche de question-réponse.

Parmi tous les autres facteurs, nous pouvons observer que la nature du déclencheur du cadre ($f2$) est effectivement un facteur de complexité pour tous les modèles, bien qu'il ne soit statistiquement significatif pour aucun d'entre eux. Le facteur $f3$ est également validé pour tous les modèles sauf LLAMA et GPT-3.5, mais il n'est significatif que pour CamemBERT. Concernant la distance entre le déclencheur et la réponse ($f4$), elle affecte négativement de façon plus importante les "petits" modèles, ce qui peut orienter vers l'hypothèse que les LLMs encodent mieux la structure syntaxique des phrases et modélisent ainsi la structure profonde des énoncés.

Les deux facteurs de complexité qui sont validés pour tous les modèles et qui sont pour la plupart statistiquement significatifs sont le nombre d'éléments du cadre dans la relation sémantique ($f5$) et la mesure de l'entropie dans la distribution des déclencheurs de ces mêmes relations ($f6$).

Il est intéressant de noter que les facteurs les plus fiables sont ceux les plus étroitement liés à la définition des cadres (nombre d'arguments pour $f5$ et entropie dans la distribution des déclencheurs pour $f6$) plutôt qu'à leur utilisation dans un contexte particulier (choix de la forme syntaxique du déclencheur dans $f2$, répétition du déclencheur dans la question dans $f3$, et complexité de l'arbre syntaxique dans $f4$). Par conséquent, les facteurs $f5$ et $f6$ peuvent être assimilés à une mesure de l'ambiguïté sémantique intrinsèque dans les relations question/réponse.

Cela peut être illustré par quelques exemples de notre corpus. Par exemple, le Frame *Request* peut avoir plus de 20 déclencheurs dans le lexique de Berkeley Framenet [6]. Dans notre corpus d'entraînement, il compte 118 occurrences avec 18 déclencheurs différents, résultant en une des mesures d'entropie les plus élevées, et un score ROUGE-L variant de 0,55 à 0,84 selon le modèle.

En contraste, la Frame *Installing*, défini comme "Un *Agent* place un *Composant* dans un *Emplacement Fixe* de sorte que le *Composant* est attaché et interconnecté et par là fonctionnel", a seulement deux déclencheurs dans le dictionnaire Framenet, *installer* et *installation*. Il compte 58 occurrences dans notre corpus d'entraînement avec 2 déclencheurs principaux et est l'un des cadres *faciles* avec une entropie faible et un score ROUGE-L variant de 0,79 à 0,90 selon le modèle.

De même, ($f5$) démontre que certains cadres présentent un nombre moyen inhabituellement bas d'éléments de cadre dans leurs exemples (≤ 2). Par exemple, le Frame *Origin* contient seulement deux FEs essentiel s(*Origin* et *Entity*), sans FEs non-essentiels présents dans Berkeley Framenet. Par contraste, les Frames *Giving* et *Contacting* comportent respectivement trois et cinq FEs essentiel, ainsi que de nombreux FEs non-essentiels dans FrameNet. Ce schéma reflète le phénomène observé avec $f6$, où le Frame *Origin* obtient un score en dessous de la moyenne tandis que *Contacting* et *Giving* sont classifiés comme Frames 'faciles'.

6. https://framenet.icsi.berkeley.edu/frameIndex

6 Travaux connexes

Notre travail se situe dans le domaine de l'évaluation des modèles. Notre approche contraste avec les évaluations à grande échelle qui couvrent plusieurs tâches, corpus et langues (Laskar *et al.*, 2023; Liang *et al.*, 2023; Srivastava *et al.*, 2023; Brown *et al.*, 2020; Wang *et al.*, 2019). Il se rapporte à des études ciblées traitant de phénomènes linguistiques spécifiques tels que les négations (Truong *et al.*, 2022, 2023; Zhang *et al.*, 2023; Ravichander *et al.*, 2022), l'ambiguïté dans les tâches d'inférence (Liu *et al.*, 2023), et l'extraction d'informations ouverte (Lechelle *et al.*, 2019), qui utilisent des ensembles de données petits et méticuleusement organisés pour évaluer avec précision les capacités des modèles pour la tâche. Notre étude fait écho à cette dernière, en explorant des évaluations linguistiques ciblées.

Cette étude est également liée à d'autres efforts de recherche qui ont été dirigés vers l'évaluation de la fiabilité des LLMs "fermés" accessibles uniquement via une API comme ChatGPT sur des benchmarks dans le domaine de la question-réponse basée sur la connaissance (KBQA) (Tan *et al.*, 2023), ainsi que son applicabilité générale à travers un large éventail de tâches NLP (Kocoń *et al.*, 2023; Laskar *et al.*, 2023). Ces études suggèrent que bien que ChatGPT présente une performance robuste sur un ensemble de tâches très large et diversifié, il peut également être devancé par des modèles spécialisés spécifiques à la tâche. Nous avons trouvé des résultats similaires dans notre étude mais montrons également que malgré ses forces, un modèle tel que ChatGPT peut être sensible aux mêmes facteurs de complexité qu'un modèle beaucoup plus petit comme T5 ou même CamemBert.

Dans l'ensemble, cette étude promeut l'idée que nous avons besoin d'un cadre d'évaluation plus précis et peut être reliée à d'autres études telles que (Ribeiro *et al.*, 2020) qui identifient des *échecs critiques* dans les modèles commerciaux et à la pointe de la technologie en proposant une méthodologie de test agnostique au modèle et à la tâche ou (Gehrmann *et al.*, 2023) insistant sur le fait que pour comparer les modèles, nous avons besoin d'un processus d'annotation plus "*soigné [...] pour caractériser leur qualité de sortie et les distinguer*".

7 Conclusion

Dans cette étude, nous avons mené une expérience utilisant un corpus de questions-réponses annoté sémantiquement pour identifier des facteurs de complexité sémantique inhérents à cette tâche, indépendamment de l'architecture et de la taille des modèles. Cette investigation est cruciale car elle souligne les avantages potentiels de se concentrer sur des modèles avec moins de paramètres pour gérer des ensembles de données difficiles. Il convient de noter que les gains de performance réalisés avec des modèles plus petits pourraient être éclipsés par des modèles plus grands. En partitionnant les ensembles de données basés sur des phénomènes linguistiques d'une complexité équivalente, indépendamment du modèle, nous pouvons nous attendre à ce que les améliorations de performance se généralisent à travers les modèles. Nos résultats démontrent que les facteurs de complexité sémantique identifiés catégorisent efficacement le corpus en sous-corpus de niveaux de difficulté variables, indépendamment du modèle employé.

Références

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The berkeley framenet project. In *COLING 1998 Volume 1 : The 17th International Conference on Computational Linguistics*.

BÉCHET F., ALOUI C., CHARLET D., DAMNATI G., HEINECKE J., NASR A. & HERLEDAN F. (2019). Calor-quest : generating a training corpus for machine reading comprehension models from shallow semantic annotations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, p. 19–26.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

D'HOFFSCHMIDT M., BELBLIDIA W., HEINRICH Q., BRENDLÉ T. & VIDAL M. (2020). FQuAD : French question answering dataset. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 1193–1208, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.107.

GEHRMANN S., CLARK E. & SELLAM T. (2023). Repairing the cracked foundation : A survey of obstacles in evaluation practices for generated text. *J. Artif. Int. Res.*, **77**. DOI : 10.1613/jair.1.13715.

JIANG A. Q., SABLAYROLLES A., ROUX A., MENSCH A., SAVARY B., BAMFORD C., CHAPLOT D. S., CASAS D. D. L., HANNA E. B., BRESSAND F. *et al.* (2024). Mixtral of experts. *arXiv preprint arXiv :2401.04088*.

KHASHABI D., CHATURVEDI S., ROTH M., UPADHYAY S. & ROTH D. (2018). Looking beyond the surface : A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 252–262.

KOCOŃ J., CICHECKI I., KASZYCA O., KOCHANEK M., SZYDŁO D., BARAN J., BIELANIEWICZ J., GRUZA M., JANZ A., KANCLERZ K. *et al.* (2023). Chatgpt : Jack of all trades, master of none. *Information Fusion*, p. 101861.

LASKAR M. T. R., BARI M. S., RAHMAN M., BHUIYAN M. A. H., JOTY S. & HUANG J. (2023). A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Findings of the Association for Computational Linguistics : ACL 2023*, p. 431–469, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.29.

LECHELLE W., GOTTI F. & LANGLAIS P. (2019). WiRe57 : A fine-grained benchmark for open information extraction. In A. FRIEDRICH, D. ZEYREK & J. HOEK, Éds., *Proceedings of the 13th Linguistic Annotation Workshop*, p. 6–15, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-4002.

LIANG P., BOMMASANI R., LEE T., TSIPRAS D., SOYLU D., YASUNAGA M., ZHANG Y., NARAYANAN D., WU Y., KUMAR A. & BENJAMIN NEWMAN E. A. (2023). Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

LIU A., WU Z., MICHAEL J., SUHR A., WEST P., KOLLER A., SWAYAMDIPTA S., SMITH N. & CHOI Y. (2023). We're afraid language models aren't modeling ambiguity. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 790–807, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.51.

MARTIN L., MULLER B., SUÁREZ P. J. O., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D., SAGOT B. *et al.* (2020). Camembert : a tasty french language model. In *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics*.

MARZINOTTO G., BÉCHET F., DAMNATI G. & NASR A. (2018). Sources of Complexity in Semantic Frame Parsing for Information Extraction. In *International FrameNet Workshop 2018*, Miyazaki, Japan. HAL : hal-01731385.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). Squad : 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 2383–2392 : Association for Computational Linguistics. DOI : 10.18653/v1/D16-1264.

RAVICHANDER A., GARDNER M. & MARASOVIC A. (2022). CONDAQA : A contrastive reading comprehension dataset for reasoning about negation. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 8729–8755, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.598.

RIBEIRO M. T., WU T., GUESTRIN C. & SINGH S. (2020). Beyond accuracy : Behavioral testing of NLP models with CheckList. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 4902–4912, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.442.

SRIVASTAVA A., RASTOGI A., RAO A., SHOEB A. A. M., ABID A., FISCH A., BROWN A. R., SANTORO A., GUPTA A. & ADRIÀ GARRIGA-ALONSO E. A. (2023). Beyond the imitation game : Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

TAN Y., MIN D., LI Y., LI W., HU N., CHEN Y. & QI G. (2023). Can chatgpt replace traditional kbqa models ? an in-depth analysis of the question answering performance of the gpt llm family. In *International Semantic Web Conference*, p. 348–367 : Springer.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S. *et al.* (2023). Llama 2 : Open foundation and fine-tuned chat models. *arXiv preprint arXiv :2307.09288*.

TRUONG T. H., BALDWIN T., VERSPOOR K. & COHN T. (2023). Language models are not naysayers : an analysis of language models on negation benchmarks. In A. PALMER & J. CAMACHO-COLLADOS, Éds., *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, p. 101–114, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.starsem-1.10.

TRUONG T. H., OTMAKHOVA Y., BALDWIN T., COHN T., LAU J. H. & VERSPOOR K. (2022). Not another negation benchmark : The NaN-NLI test suite for sub-clausal negation. In Y. HE, H. JI, S. LI, Y. LIU & C.-H. CHANG, Éds., *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 883–894, Online only : Association for Computational Linguistics.

WANG A., PRUKSACHATKUN Y., NANGIA N., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. (2019). Superglue : A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, **32**.

WEI J., BOSMA M., ZHAO V. Y., GUU K., YU A. W., LESTER B., DU N., DAI A. M. & LE Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv :2109.01652*.

XUE L., CONSTANT N., ROBERTS A., KALE M., AL-RFOU R., SIDDHANT A., BARUA A. & RAFFEL C. (2021). mt5 : A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 483–498.

ZHANG Y., YASUNAGA M., ZHOU Z., HAOCHEN J. Z., ZOU J., LIANG P. & YEUNG S. (2023). Beyond positive scaling : How negation impacts scaling trends of language models. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Findings of the Association for Computational Linguistics : ACL 2023*, p. 7479–7498, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.472.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Évaluation de l'apport des chaînes de coréférences pour le liage d'entités

Léo Labat[1, 2] Lauriane Aufrant[2]
(1) ENS, Paris, France
(2) Inria, Paris, France
prenom.nom@inria.fr

RÉSUMÉ

Ce travail propose de revisiter les approches de liage d'entités au regard de la tâche très proche qu'est la résolution de coréférence. Nous observons en effet différentes configurations (appuyées par l'exemple) où le reste de la chaîne de coréférence peut fournir des indices utiles pour améliorer la désambiguïsation. Guidés par ces motivations théoriques, nous menons une analyse d'erreurs accompagnée d'expériences oracles qui confirment le potentiel de stratégies de combinaison de prédictions au sein de la chaîne de coréférence (jusqu'à 4.3 F1 sur les mentions coréférentes, en anglais). Nous esquissons alors une première preuve de concept de combinaison par vote, en explorant différentes heuristiques de pondération, qui apporte des gains modestes mais interprétables.

ABSTRACT

Evaluating the benefits of coreference chains for entity linking.

This work proposes to revisit entity linking approaches in light of the closely related task of coreference resolution. Indeed, we observe several configurations (supported by examples) in which the rest of the coreference chain can yield valuable cues to improve disambiguation. Driven by these theoretical motivations, we conduct an error analysis as well as oracle experiments, which confirm the potential of combination strategies for predictions within the coreference chain (up to 4.3 F1 on coreferring mentions, in English). We then sketch a first proof of concept of voting-based combination, by exploring various weighting heuristics, which yields limited but interpretable gains.

MOTS-CLÉS : liage d'entités, résolution de coréférence, approches jointes.

KEYWORDS: entity linking, coreference resolution, joint approaches.

1 Introduction

Reconnaître les entités mentionnées dans un texte demeure un défi pour le traitement automatique des langues, qui approche cette compétence linguistique par deux tâches : le liage d'entité et la résolution de coréférence. Le liage d'entité (*entity linking* en anglais) consiste à associer toute chaîne de caractères se référant à une entité dans un texte, que l'on appelle une mention, à une entrée d'une base de connaissances, tandis que la résolution de coréférence regroupe les mentions se référant à des entités en ensembles de sorte que toutes les mentions d'un même ensemble se réfèrent à la même entité. Néanmoins, à chacune de ces tâches est souvent associé un concept de mentions différent : en liage, une mention est le plus souvent un nom propre [1], alors qu'en résolution de coréférence, tout

1. Le domaine biomédical constitue à cet égard une exception notable.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 397–409.

empan (*span*) se référant à une entité peut être coréférent, pourvu qu'un autre empan dans le même document se réfère à la même entité. Ces différences rendent bien souvent les performances de liage et de résolution incomparables car réalisées sur des corpus distincts. Ainsi, le corpus de référence pour le liage d'entité AIDA-CONLL (Hoffart *et al.*, 2011), qui comporte des liens entre les mentions de CONLL 2003 (Sang & Meulder, 2003) et la base YAGO 2 (Hoffart *et al.*, 2013) n'inclut pas de pronoms parmi ses mentions. Quant à la coréférence, le jeu de données le plus utilisé, ONTONOTES 5.0 (Pradhan *et al.*, 2013), inclut les pronoms et des mentions moins spécifiées qu'en liage dans ses chaînes de coréférence, mais ne contient pas d'annotation pour les singletons, c'est-à-dire que lorsqu'une entité n'est mentionnée qu'une fois, cette mention solitaire n'est pas annotée.

Nous faisons l'hypothèse que les informations de coréférence peuvent améliorer les décisions prises par un algorithme de liage. En effet, les mentions de coréférence peuvent contenir des indices utiles à la désambiguïsation d'entités, car tant les pronoms que les périphrases sont autant d'indices qui permettent de savoir de qui ou de quoi il est question. La plupart des algorithmes de liage prennent déjà en compte le contexte pour assigner un lien à une mention donnée, mais la particularité de la coréférence, en tant que dépendance à longue distance, est d'agréger au fil du texte des mentions qui elles-mêmes ou avec leur contexte, peuvent être directement désambiguïsantes, quelle que soit leur répartition dans le document. L'exemple présenté en Figure 1, inspiré de (Hajishirzi *et al.*, 2013) montre comment une chaîne de coréférence peut contenir des informations utiles à la désambiguïsation.

[*Michael Eisner*]$_1$ and [*Donald Tsang*]$_2$ announced the grand opening of [*Disneyland*]$_3$ yesterday.
[*Eisner*]$_1$ thanked [*the President*]$_2$ and welcomed fans to [*the Hong Kong park*]$_3$.

FIGURE 1 – Exemple de partage d'information entre résolution de coréférence et liage d'entités

Les mentions *Disneyland* et *the Hong Kong park* sont ambiguës lorsqu'elles sont considérées isolément, même dans leur contexte à l'échelle de la phrase. Mais si on les identifie comme coréférentes, déterminer l'entité à laquelle elles se réfèrent est aisé : il s'agit de Disneyland Hong Kong, et non de Disneyland Paris ou de Hong Kong Park. Nous proposons donc dans cet article d'explorer comment ces indices potentiellement utiles fournis par la coréférence peuvent être exploités pour améliorer les décisions de liage d'entités.

Après un aperçu de ces deux tâches et des approches associées (§2), nous présentons dans cet article notre cadre expérimental reposant sur le corpus AIDA-CONLL (§3), qui nous permettra d'affiner nos motivations (§4) et d'explorer une première preuve de concept pour l'exploitation de ces indices (§5) en estimant les gains potentiels à insérer ces informations dans des modèles de l'état de l'art.

2 Etat de l'art

2.1 Liage d'entités

Formellement, le liage d'entités consiste à prendre en entrée un document textuel D contenant un ensemble de mentions se référant à des entités ($M = \{m_i\}_i$), ainsi qu'une base de connaissances KB dont les entrées sont des entités nommées ($E = \{e_i\}_i$), et à retourner un ensemble de paires (m_i, e_i) où e_i est l'entité référencée par la mention m_i dans E. L'ensemble E peut parfois contenir une étiquette spéciale "NIL" pour dénoter qu'il s'agit d'une entité inconnue dans la base. Lorsqu'il est réalisé de bout en bout, les mentions ne sont pas données en entrée et doivent être détectées par

l'algorithme.

Approche modulaire L'approche modulaire peut être implémentée de diverses façons. Une décomposition proposée par Piccinno & Ferragina (2014) consiste en trois étapes : un module de détection associe d'abord à chaque mention possible un certain nombre de candidats, puis un module de désambiguïsation en sélectionne un pour chaque mention et un module d'élagage (*pruning*) élimine les liens les moins probables. D'autres divisions modulaires sont possibles, comme par exemple REL (van Hulst *et al.*, 2020), qui réalise successivement une détection de mentions, une génération de candidats pour chaque mention et enfin une désambiguïsation, c'est-à-dire la sélection d'un seul candidat pour chaque mention.

Modèles de langue pré-entraînés Comme pour de nombreuses autres tâches de TALN, le liage d'entités a bénéficié de l'introduction de représentations pré-entraînées à l'aide de Transformers (Vaswani *et al.*, 2017). Appliquer un encodeur de type BERT (Devlin *et al.*, 2019) à des mots, empans et entités permet d'intégrer le contexte dans le calcul de similarité mention/entité. Kolitsas *et al.* (2018) complète cette approche avec une table de correspondance probabiliste entre mentions et entités construite à partir d'hyperliens Wikipédia (Ganea & Hofmann, 2017). Les modèles autorégressifs peuvent aussi être employés pour reformuler le liage comme une tâche de génération de texte : par exemple, GENRE (Cao *et al.*, 2021) génère des séquences associant mentions tirées du texte et des identifiants d'entités, en contraignant la génération de ces derniers à partir d'une liste prédéfinie. Mrini *et al.* (2022) utilisent ce type de génération pour combiner détection de mentions, désambiguïsation et reclassement en mode multi-tâches.

Questions-Réponses En tant que tâche consistant en une forme de classification d'empans, le liage d'entité peut également être formulé comme une tâche de questions-réponses. C'est le cas de Zhang *et al.* (2021), qui obtiennent des performances comparables aux approches précédentes, mais au prix de ressources de mémoire et de calcul plus importantes, nécessitant de surcroît des canvas élaborés à la main.

Classification de tokens Une approche distincte consiste à poser le liage comme une tâche de classification de tokens. La décision de liage est alors prise à l'échelle du token sur l'ensemble du vocabulaire, c'est-à-dire sur l'ensemble des entités possibles : Broscheit (2019) mobilise un modèle BERT sur lequel est superposée un simple réseau à propagation avant pour obtenir une distribution de probabilités sur les 700 000 entités possibles pour chaque token. A la classifications de tokens peut également s'ajouter des classifications à des échelles suprasegmentales : le liage revient alors à réaliser une prédiction structurée. C'est le cas de SpEL (Shavarani & Sarkar, 2023), qui agrège de façon structurée des décisions de classification prises sur trois niveaux : au niveau des tokens, au niveau des mots puis au niveau des syntagmes (*phrase level*). Ces approches présentent l'avantage d'exploiter la capacité des encodeurs de modèles de langue à prendre en compte le contexte dans leur stratégie d'agrégation de prédictions pour former des annotations d'empans cohérentes.

2.2 Résolution de coréférences

Formellement, la résolution de coréférence consiste à prendre en entrée un document D constitué de texte et à identifier toutes les mentions qui se réfèrent à la même entité dans le monde réel. Contrairement au liage d'entités, où les mentions sont explicitement identifiées et liées à une base d'entités nommées, la résolution de coréférence vise à regrouper les mentions $M = \{m_i\}$ qui font référence à la même entité sans base de connaissances. Le résultat est un ensemble de chaînes de coréférence $C = \{c_i\}$, où chaque chaîne c_i est un sous-ensemble de M tel que toutes les mentions dans c_i se réfèrent à la même entité.

Approches par combinaisons de n-grams Le paradigme qui s'est imposé à partir de Lee *et al.* (2017) consiste à réaliser la résolution de coréférence sans détection de mentions préalable en attribuant des scores à des empans ainsi qu'à des paires de mentions. Cette approche surgénère des mentions en considérant tout empan comme potentielle mention. Des statégies d'élagage de mentions ont été affinées avec Lee *et al.* (2018), mais les gains de performances tiennent principalement de l'amélioration de la qualité des représentations vectorielles : d'abord avec des LSTMs (Hochreiter & Schmidhuber, 1997) pour Lee *et al.* (2017) puis avec ELMo (Peters *et al.*, 2018) pour Lee *et al.* (2018). L'avènement de modèles de langue tels que BERT ont donc également donné lieu à d'importants gains de performances en résolution de coréférence. Une variante de BERT spécifiquement dédiée à l'encodage d'empans, SpanBERT, a été élaborée par Joshi *et al.* (2019) et a démontré l'intérêt d'un objectif d'entraînement consistant à prédire des empans plutôt que des seuls mots dans le cadre d'une modélisation de langue masquée pour améliorer les performances de résolution de coréférence. Néanmoins, la méthode *Start-to-End*, ou S2E (Kirstain *et al.*, 2021), montre qu'une approche semblable basée sur les Longformer (Beltagy *et al.*, 2020) qui se contente de représentations contextuelles du premier et du dernier token de chaque mention réduit considérablement la complexité du problème et obtient des résultats comparables plus rapidement et avec moins de ressources de calcul.

Formulation générative Il existe également une approche générative du problème de la résolution de coréférence. En se fondant sur un modèle multilingue T5 (Raffel *et al.*, 2019), d'aucuns ont proposé un système de transitions pour faire réaliser des décisions d'association de mentions à des chaînes de coréférence par un modèle génératif (Bohnet *et al.*, 2023). Les modèles génératifs peuvent également être utilisés pour générer des prédictions de coréférence en reproduisant directement le texte fourni en entrée avec des annotations (Zhang *et al.*, 2023).

Questions-réponses D'autres (Wu *et al.*, 2020) ont reformulé la tâche de coréférence comme une tâche de réponse aux questions (*Question Answering*) en s'appuyant notamment sur SpanBERT. Les empans proposés au système de questions-réponses sont générés de façon similaire à Lee *et al.* (2017) et sont introduits dans des modèles de question prédéfinis contenant des encarts à remplir avec une mention donnée.

2.3 Approches jointes

Une proposition de modélisation jointe du liage d'entités et de la résolution de coréférence élaborée par Hajishirzi *et al.* (2013) montre comment la réalisation simultanée d'une tâche peut contribuer à améliorer les performances de l'autre et réciproquement. L'implémentation est modulaire et construit de façon incrémentale des chaînes de coréférence par application de règles déterministes de constitution d'ensembles de mentions coréférences qui satisfassent des contraintes d'entités, c'est-à-dire qui garantissent la cohérence des décisions de liage.

D'autres formulations jointes des deux tâches explorent le potentiel des informations de coréférence pour améliorer les performances de liage. En effet, les implémentations du liage d'entité sus-mentionnées considèrent tantôt une liste finie de candidats tantôt une liste générée *ad hoc* pour la mention courante. Si la bonne entité n'est pas présente dans la liste prédéfinie ou générée lors de l'inférence, alors l'algorithme de liage n'a aucune chance de prendre une décision de liage exacte. C'est le problème que des auteurs comme Zaporojets *et al.* (2022) cherchent à résoudre en s'inspirant de Angell *et al.* (2021) qui, dans leur implémentation de liage d'entités sur des données biomédicales, rendent possibles les décisions de liage entre les mentions plutôt qu'uniquement entre mentions et sommets de la base de connaissances. De la sorte, une mention pour laquelle la bonne entité

n'est pas présente dans la liste de candidats générée pour elle peut néanmoins être liée à une autre mention, elle-même susceptible d'avoir été bien liée. Néanmoins, cette approche telle qu'elle a été implémentée s'en tient à une conception des mentions référentes ancrée du côté du liage, du fait de la nature-même des données d'entraînement : AIDA-CONLL et le corpus DWIE (Zaporojets *et al.*, 2020) ne contiennent pas d'annotation de liage pour les pronoms, par exemple. La notion de coréférence est donc réduite au fait que les mentions liées à la même entité sont coréférentes dans le sens où elle se réfèrent à la même entité, mais ignore donc le phénomène linguistique de coréférence dans ses manifestations les plus ordinaires (pronoms, périphrases, etc.).

3 Conditions expérimentales

3.1 Données

Toutes nos expériences reposent sur AIDA-CONLL (Hoffart *et al.*, 2011), corpus anglais journalistique de 20 744 phrases annotées en liage d'entités avec la base de connaissances YAGO 2 (Hoffart *et al.*, 2013) (28813 mentions liées à 5586 entités, parmi les 50 millions de YAGO 2). Nous utilisons la version du corpus fournie par Zaporojets *et al.* (2022), qui inclut des liens supplémentaires (Kolitsas *et al.* (2018) ayant identifié la version originale comme incomplète à l'échelle du document). Nous conservons leur partition en données d'entraînement (946 documents, 19190 mentions, 4085 entités), de développement (TEST-A, 216 documents, 4960 mentions, 1649 entités) et de test (TEST-B, 231 documents, 4663 mentions, 1547 entités), issue du corpus sous-jacent CONLL 2003 (Sang & Meulder, 2003). Plus de la moitié des entités apparaissant en test (57%, 882 entités parmi 1547) sont nouvelles, c'est-à-dire jamais rencontrées à l'entraînement.

3.2 Métriques

On considère la précision, le rappel et la métrique F1 (micro-F1) appliqués à la détection de mentions ($F1^m$), au liage d'entités lui-même ($F1^e$, pour les cas où les mentions correctes sont déjà données), et à la combinaison de la détection et du liage ($F1^{m+e}$). Nous employons la variante stricte du score F1 : une mention est bien liée uniquement si ses frontières sont correctement prédites et si le lien prédit correspond au lien de référence dans le corpus.

Nous suivons la pratique usuelle de l'état de l'art de restreindre l'évaluation aux mentions dont l'entité (prédite ou de référence) est effectivement présente en base (configuration "*InKB*" de Röder *et al.* (2018)).

À des fins d'analyse, nous mesurons aussi le score F1 sur des sous-ensembles spécifiques de mentions : mentions de référence ($F1_R$), mentions détectées lors de la détection de mention du liage ($F1_L$), mentions détectées lors de la détection de mention de la coréférence ($F1_C$), mentions de référence qui ont été détectées pour le liage ($F1_{R+L}$), mentions détectées à la fois pour le liage et la coréférence ($F1_{L+C}$), etc.

3.3 Modèles

Liage d'entités Nous expérimentons avec trois modèles de liage d'entités couvrant des approches variées :

— **Un liage modulaire classique :** REL (van Hulst *et al.*, 2020) combine des modules de détection de mentions (reconnaissance d'entités nommées par FLAIR (Akbik *et al.*, 2018)), de génération d'entités candidates (suivant (Ganea & Hofmann, 2017) : 4 candidats les plus probables d'après

la mention et 3 d'après leur similarité au contexte, en sommant les plongements sur 50 tokens) et de désambiguïsation d'entités (sélection d'un candidat par mention, avec prise en compte des relations latentes entre mentions pour la cohérence de leurs liages (Le & Titov, 2018), sur une fenêtre de 200 tokens). Nous utilisons le modèle *REL 2019*, où les plongements Wikipedia2Vec (Yamada *et al.*, 2016) utilisés pour la génération de candidats sont calculés sur l'export Wikipédia de juillet 2019. Le module de désambiguïsation utilise les plongements GloVe (Pennington *et al.*, 2014) et est entraîné sur AIDA-CONLL.

— **Une approche par prédiction structurée :** SPEL (Shavarani & Sarkar, 2023) utilise RoBERTa (Liu *et al.*, 2019) pour estimer pour chaque sous-mot les probabilités d'un ensemble (fixe) d'entités candidates, puis combine les 20 candidats les plus probables à l'échelle du mot, et refait encore une prédiction à l'échelle du syntagme. Nous utilisons les deux versions `SpEL-base` et `SpEL-large` (suivant la version de RoBERTa), avec les mêmes hyperparamètres que Shavarani & Sarkar (2023) : une fenêtre de contexte de 284 tokens avec chevauchement de 20 tokens, et un ensemble fixe de 5601 candidats (dont "null").

— **Une approche générative :** GENRE (Cao *et al.*, 2020) affine un modèle de langue pré-entraîné BART (Lewis *et al.*, 2019) pour générer des identifiants d'entités (avec un décodage contraint pour qu'ils soient valides) afin de désambiguïser les mentions pré-identifiées. Dans sa version de bout en bout, les prédictions sont réalisées en insérant dans le texte des crochets aux limites de mentions en plus des identifiants d'entité. Nous utilisons les mêmes hyperparamètres que Cao *et al.* (2020) : décodage avec faisceau de taille 10 à l'apprentissage (6 au test) et au plus 15 étapes, 384 tokens de contexte maximal.

Résolution de coréférence Notre approche est applicable à n'importe quel algorithme de coréférence qui génère ses propres mentions. Pour nos expériences nous utilisons S2E (Kirstain *et al.*, 2021), que nous avons réentraîné avec les mêmes hyperparamètres et les mêmes données OntoNotes 5.0 (Pradhan *et al.*, 2013). Les auteurs rapportent un score $F1^m$ moyen de 80.3.

4 Motivations théoriques et empiriques

Les interactions fortes entre la tâche de liage d'entités et celle de résolution de coréférences amènent naturellement la question de leur combinaison à émerger. Afin de préciser cette intuition, nous évaluons les modèles de liage comparativement sur l'ensemble des mentions et uniquement sur les mentions coréférentes. Il en ressort (Table 1) que les mentions coréférentes sont typiquement plus difficiles à lier, en particulier pour les modèles les moins performants.

	R^{m+e}	R^{m+e}_{C}	$R^{m+e}_{C(\text{ref})}$
REL	75.5	71.4	74.8
SPEL-BASE	88.9	87.6	88.3
SPEL-LARGE	89.8	90.5	89.8

TABLE 1 – Impact de l'appartenance à une chaîne de coréférence sur le rappel

Une analyse manuelle sur des exemples du jeu de développement TEST-A a révélé 3 configurations théoriques d'intérêt :

1. Suite à une mention non-ambiguë (ex : prénom et nom), les suivantes y font référence d'une manière ambiguë (ex : prénom) et leur contexte contient trop peu d'indices pour les désambiguïser individuellement.

2. La première mention est ambiguë, mais la suite du texte ajoute de l'information en la centrant sur des mentions anaphoriques (ex : "*Il habite à Paris. Cette petite ville du Texas a été fondée en 1845.*").
3. Chacune des mentions présente une ambiguïté lorsque prise individuellement, mais en combinant tous leurs indices l'entité est claire. C'est le cas de l'exemple illustré en Figure 1.

Afin de quantifier ces effets, du moins les configurations 1 et 2 (la configuration 3 étant plus dure à objectiver), on mesure séparément les précisions, rappels et F1 suivant la position des mentions dans la chaîne de coréférence. Plus précisément, on affecte chaque mention à un quintile suivant la valeur $\frac{position-1}{taille_{\text{chaîne}}-1}$. Les résultats rapportés en Table 2 montrent une nette baisse de performance du deuxième jusqu'au dernier quintile, ce qui est cohérent avec la configuration 1. Le résultat bas sur le premier quintile est moins intuitif, mais peut facilement s'interpréter au regard du style du corpus AIDA-CONLL (articles de presse), où chaque document commence par un titre. Comme on peut l'observer en Figure 2, les premières mentions présentes dans un titre d'article sont souvent concises et ambiguës (configuration 2), et c'est le début de l'article qui apporte des informations substantielles. Dans cet exemple particulier on observe bien une combinaison des configurations 1 et 2.

		Quintile de la position dans la chaîne				
		0-19%	20-39%	40-59%	60-79%	80-100%
REL	$F1^{m+e}_{C(\text{ref})}$	73.6	84.2	82.5	83.5	73.7
SPEL-BASE	$F1^{m+e}_{C(\text{ref})}$	91.4	94.7	93.1	93.0	91.2
SPEL-LARGE	$F1^{m+e}_{C(\text{ref})}$	93.4	94.9	94.2	95.5	91.9
GENRE	$F1^{e}_{C(\text{ref})}$	81.4	82.2	83.2	78.2	80.2

TABLE 2 – Score F1 de liage en fonction de la position de la mention dans sa chaîne de coréférence

BASEBALL-**GONZALEZ** HOMERS TWICE AS RANGERS BEAT INDIANS.
ARLINGTON, Texas 1996-08-31
Juan Gonzalez homered twice and Ivan Rodriguez added a two-run shot as the Texas Rangers defeated the Cleveland Indians 5-3 in a matchup of division leaders Friday.
Rodriguez's 18th homer, off Chad Ogea (7-5) in the first, gave Texas a 2-0 lead. One out later, **Gonzalez** smacked his 40th homer, extending his hitting streak to 20 games.
Gonzalez, who hit in 21 straight games earlier this season, joined Mickey Rivers as the only players in Texas history with two 20-game streaks in the same year.
Gonzalez hit his second homer in the third for his fifth multi-homer game of the season. **Gonzalez** has three 40-homer seasons and his 121 RBI broke Ruben Sierra's team record of 119 set in 1989. (...)

FIGURE 2 – Répartition de l'ambiguïté le long d'une chaîne de coréférence (document 1059, TEST-A)

5 Amélioration du liage d'entités via les coréférences

Les résultats de la section précédente soulignent un certain potentiel à exploiter l'information de coréférence pour enrichir le liage, d'une part car les autres mentions de la chaîne peuvent apporter des informations utiles et d'autre part car les mentions coréférentes présentent souvent de plus grandes difficultés pour le liage.

Nous suggérons ici une approche où pour chaque mention à lier (tour à tour), on commence par désambiguïser les autres mentions de la chaîne (y compris si non-détectées comme mentions à lier : on force une désambiguïsation) puis on combine ces éléments pour informer le liage en cours.

Il est à noter que cette stratégie n'est applicable que pour une partie des mentions à lier : celles pour lesquelles des mentions coréférentes existent bien sûr, mais surtout celles qui sont détectées comme telles par le modèle de coréférence. En effet, outre les inévitables erreurs de ce dernier, les divergences de conventions d'annotation mènent aussi à des décalages entre mentions détectées de part et d'autre (par exemple "vendor Valery Ivankov" côté coréférence et "Valery Ivankov" côté liage), et donc une impossibilité d'exploiter les indices de coréférence, faute d'adéquation des mentions. En l'occurrence, sur les 5616 mentions de référence dans TEST-B, seulement 1804 (32%) d'entre elles appartiennent à une chaîne de coréférence prédite par S2E. Il serait envisageable d'explorer des heuristiques plus permissives pour utiliser les indices de coréférences fournis par des mentions non-identiques, mais avec un risque important d'introduire du bruit, et cela sort du cadre du présent travail.

Nous commençons par des expériences oracles pour estimer les gains envisageables, puis nous évaluons quelques heuristiques de vote, permettant d'esquisser des pistes de recherche future. L'ensemble de ces expériences est réalisé sur le jeu de test TEST-B. Nous expérimentons avec GENRE uniquement, dont les performances sont davantage interprétables, dans la mesure où forcer la désambiguïsation de mentions supplémentaires a en soi un impact sur la performance globale de REL (qui considère l'interaction entre prédictions), et la notion de mention n'existe que partiellement dans SPEL étant donné qu'elles sont construites hiérarchiquement à partir de sous-mots.

5.1 Expériences oracles

Afin d'estimer le gain potentiel que peuvent apporter les indices issus du reste de la chaîne de coréférence, on regarde si l'entité de référence est au moins présente parmi toutes les prédictions faites au sein de la chaîne de coréférence. Le cas échéant, une stratégie optimale de combinaison de ces prédictions permettrait de faire ressortir la bonne entité et l'attribuer à l'ensemble de la chaîne.

Les résultats présentés en Table 3 révèlent une marge de progression de +1.2 F1 dans l'ensemble, qui monte à +4.3 F1 lorsqu'on ne regarde que les mentions détectées par la coréférence (les seules pour lesquelles l'approche est donc applicable), la prédiction étant mécaniquement inchangée pour les autres.

	$F1^e_R$	$F1^e_R$ avec oracle	$F1^e_{R+C}$	$F1^e_{R+C}$ avec oracle
GENRE	72.4	73.6	75.7	79.4

TABLE 3 – Mesure oracle du gain de performance accessible avec une sélection parfaite de l'entité parmi les prédictions de la chaîne de coréférence

Ces résultats expérimentaux confirment le caractère prometteur de l'approche. En effet dans l'exemple "*Il habite à Paris. Cette petite ville du Texas a été fondée en 1845.*", même si les indices de la première occurrence ne sont pas suffisant pour désambiguïser "Paris", des indices supplémentaires sont disponibles (et suffisants) pour désambiguïser la mention "Cette petite ville" ("Texas", "1845") et cette information peut alors être remontée pour désambiguïser "Paris". Cela suggère qu'en exploitant bien ces autres prédictions au sein de la chaîne, il serait possible de corriger une part substantielle des erreurs de liage.

Une limite de cette approche est toutefois sa dépendance à une bonne génération de candidats, surtout pour les mentions non-nommées : en ne considérant que la mention "Cette petite ville", il est en effet peu probable que l'entité "Paris (Texas)" soit identifiée comme candidate, auquel cas la désambiguïsation échouerait et ne permettrait pas de contribuer à la désambiguïsation de "Paris". Ce

point particulier dépend de la méthode utilisée pour le liage d'entités, car toutes n'incluent pas une étape de génération de candidats, ou pas sous la même forme. Pour les méthodes de liage où cela a du sens, une approche possible pour pallier cette difficulté serait de fusionner au préalable les candidats générés pour toutes les mentions de la chaîne de coréférence, augmentant ainsi les options aussi bien pour "Paris" que pour "Cette petite ville". Avec cette stratégie, la marge de progression est en fait plus importante encore.

5.2 Combinaison de prédictions

Nous évaluons ici empiriquement l'intuition suivant laquelle les autres mentions de la chaîne de coréférence peuvent apporter des indices utiles pour désambiguïser la mention d'intérêt. On teste donc plusieurs heuristiques de combinaison des prédictions au sein de la chaîne. Plus précisément, pour chaque mention M à lier pour laquelle une chaîne de coréférence C est trouvée par S2E (pour les autres nous gardons la prédiction initiale du module de désambiguïsation), nous évaluons 4 approches de combinaison :

— **Indépendant :** L'entité retenue pour M est celle prédite par le module de désambiguïsation.
— **Vote simple :** L'entité est obtenue par vote entre les entités prédites pour chaque mention de C (en forçant le module de désambiguïsation sur ces mentions). Les égalités sont résolues en faveur de l'entité prédite pour M, ou à défaut priorité est donnée aux premières mentions du texte.
— **Vote pondéré (longueur) :** Similaire au vote simple, mais en pondérant chaque vote par le nombre de tokens de la mention : par exemple "Gonzalez" aura un poids de 1 et "Juan Gonzalez" de 2. L'intuition sous-jacente est que plus l'entité est longue plus elle est informative, donc la prédiction est de base plus faible sans même considérer le contexte. Ce n'est toutefois pas systématique : par exemple "Paris" aura un poids de 1 et "Cette petite ville" de 3.
— **Vote pondéré (position) :** Similaire au vote simple, mais en donnant un poids supérieur (1.5) aux mentions de la première moitié (arrondie à l'inférieur) de la chaîne, où l'information est a priori plus riche. Ainsi, dans l'exemple de la Figure 2, les mentions 1-3 auront un poids 1.5 et les mentions 4-6 un poids 1.

Les résultats présentés en Table 4 montrent un gain modeste (+0.1 en global, +0.3 sur les mentions coréférentes). Dans l'ensemble le fait de combiner ainsi les prédictions apporte donc bien de l'information utile, même si cela pose des questions de significativité statistique, qui mériteraient une exploration empirique plus approfondie. Les heuristiques de pondération expérimentées apparaissent toutefois trop simples pour permettre une réelle émergence de la bonne entité.

	$F1^e$	$F1^e_C$	# Dégradations	# Corrections
Indépendant	72.4	75.7	–	–
Vote simple	72.5	76.0	-21	+25
Vote pondéré (longueur)	72.5	76.0	-21	+25
Vote pondéré (position)	72.5	75.8	-33	+39

TABLE 4 – Effet de plusieurs heuristiques de vote sur la performance de GENRE

Afin de mieux caractériser l'apport des indices issus du reste de la chaîne de coréférence, on mesure également le nombre de prédictions correctes devenues erronées du fait de la combinaison (# Dégradations), et le nombre de prédictions erronées corrigées par cette combinaison (# Corrections). Cette mesure confirme que la pondération par longueur de mention n'a pas d'effet concret, alors que

la priorité au début de la chaîne modifie effectivement le comportement du vote. Il en ressort aussi que les gains modestes résultent d'un nombre important de dégradations qui fait perdre le bénéfice des corrections. Il y a donc là encore une marge importante d'amélioration dans la recherche d'une heuristique de pondération plus fine pour éliminer ces dégradations.

5.3 Discussion

Bien que la méthode montre déjà des gains prometteurs, appuyés autant théoriquement qu'empiriquement, il semble toutefois y avoir un important manque à gagner en se limitant aux prédictions déjà fournies par le module de désambiguïsation. L'une des marges supplémentaires d'amélioration, déjà évoquée ci-dessus, est de renforcer la génération de candidats (pour les modèles qui en utilisent), en tirant parti du fait que chaque mention peut contribuer son lot de candidats au profit de l'ensemble des mentions de la chaîne. Cette amélioration peut aussi être appliquée au module de génération de candidats de manière indépendante (y compris sans combinaison des prédictions), et constitue une piste intéressante pour de futures explorations.

Par ailleurs, il arrive aussi que les indices offerts dans le reste de la chaîne ne soient pas suffisants pris isolément (tous les liages sont erronés) mais qu'une fois combinés ils permettent une désambiguïsation complète de la mention : c'est la configuration 3 de la section précédente. Par exemple dans le cas de la Figure 1, les prédictions pour "Disneyland" et "the Hong Kong park" sont toutes deux fausses, donc une stratégie de sélection parmi les prédictions ne corrigerait pas le liage, pourtant l'apport des indices issus de la chaîne est clair en théorie. Cette intuition oriente vers l'idée d'une combinaison précoce de ces indices, par exemple en combinant directement les représentations contextuelles des mentions (moyenne, somme, max-pooling...). Cette stratégie n'est toutefois applicable qu'à certains types de modèles de désambiguïsation (où les représentations de mentions jouent un rôle prépondérant), or SpEL par exemple ne fait intervenir que les représentations de sous-mots et pas directement des représentations de mentions.

6 Conclusion

Nous avons exploré dans ce travail les interactions possibles entre liage d'entités et résolution de coréférences. Notre analyse à la fois théorique et empirique nous a permis d'identifier plusieurs configurations d'intérêt au sein des chaînes de coréférence, soulignant des difficultés spécifiques que peuvent rencontrer les modules de désambiguïsation sur les mentions coréférentes, mais aussi le bénéfice potentiel à exploiter des indices associés à d'autres mentions de la chaîne, potentiellement moins ambiguës. Sur la base d'expériences oracles confirmant ce bénéfice, nous avons proposé une première preuve de concept pour exploiter ces informations issues des mentions coréférentes, par un mécanisme de vote entre entités prédites dans la chaîne. Nous identifions plusieurs perspectives de prolongement de ces travaux : tout d'abord le recours à la chaîne de coréférence pour compléter la génération d'entités candidates (plutôt qu'améliorer la désambiguïsation uniquement), mais aussi l'option d'une fusion plus précoce de l'information offerte au sein de la chaîne, permettant de couvrir également la configuration où toutes les mentions de la chaîne sont ambiguës mais le tout est univoque. Enfin, les divergences de conventions d'annotation entre liage et coréférence limitent l'applicabilité de la méthode à un tiers des mentions environ (où les deux prédictions coïncident), et il serait intéressant d'explorer des stratégies d'appariement entre mentions des deux modules.

Remerciements

Ce travail a été partiellement financé par l'Agence de l'Innovation de défense, dans le cadre du projet CapiTAL.

Références

AKBIK A., BLYTHE D. A. J. & VOLLGRAF R. (2018). Contextual string embeddings for sequence labeling. In *International Conference on Computational Linguistics*.

ANGELL R., MONATH N., MOHAN S., YADAV N. & MCCALLUM A. (2021). Clustering-based inference for biomedical entity linking. In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 2598–2608, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.205.

BELTAGY I., PETERS M. E. & COHAN A. (2020). Longformer : The long-document transformer. *CoRR*, **abs/2004.05150**.

BOHNET B., ALBERTI C. & COLLINS M. (2023). Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, **11**, 212–226. DOI : 10.1162/tacl_a_00543.

BROSCHEIT S. (2019). Investigating entity knowledge in BERT with simple neural end-to-end entity linking. In M. BANSAL & A. VILLAVICENCIO, Éds., *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, p. 677–685, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/K19-1063.

CAO N. D., IZACARD G., RIEDEL S. & PETRONI F. (2020). Autoregressive entity retrieval. *CoRR*, **abs/2010.00904**.

CAO N. D., IZACARD G., RIEDEL S. & PETRONI F. (2021). Autoregressive entity retrieval. In *International Conference on Learning Representations*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

GANEA O. & HOFMANN T. (2017). Deep joint entity disambiguation with local neural attention. *CoRR*, **abs/1704.04920**.

HAJISHIRZI H., ZILLES L., WELD D. S. & ZETTLEMOYER L. (2013). Joint coreference resolution and named-entity linking with multi-pass sieves. In D. YAROWSKY, T. BALDWIN, A. KORHONEN, K. LIVESCU & S. BETHARD, Éds., *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 289–299, Seattle, Washington, USA : Association for Computational Linguistics.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long Short-Term Memory. *Neural Computation*, **9**(8), 1735–1780. DOI : 10.1162/neco.1997.9.8.1735.

HOFFART J., SUCHANEK F. M., BERBERICH K. & WEIKUM G. (2013). Yago2 : A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelli-*

gence, **194**, 28–61. Artificial Intelligence, Wikipedia and Semi-Structured Resources, DOI : https ://doi.org/10.1016/j.artint.2012.06.001.

HOFFART J., YOSEF M. A., BORDINO I., FÜRSTENAU H., PINKAL M., SPANIOL M., TANEVA B., THATER S. & WEIKUM G. (2011). Robust disambiguation of named entities in text. In R. BARZILAY & M. JOHNSON, Éds., *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, p. 782–792, Edinburgh, Scotland, UK. : Association for Computational Linguistics.

JOSHI M., CHEN D., LIU Y., WELD D. S., ZETTLEMOYER L. & LEVY O. (2019). Spanbert : Improving pre-training by representing and predicting spans. *CoRR*, **abs/1907.10529**.

KIRSTAIN Y., RAM O. & LEVY O. (2021). Coreference resolution without span representations. In *Annual Meeting of the Association for Computational Linguistics*.

KOLITSAS N., GANEA O.-E. & HOFMANN T. (2018). End-to-end neural entity linking. In A. KORHONEN & I. TITOV, Éds., *Proceedings of the 22nd Conference on Computational Natural Language Learning*, p. 519–529, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/K18-1050.

LE P. & TITOV I. (2018). Improving entity linking by modeling latent relations between mentions. *CoRR*, **abs/1804.10637**.

LEE K., HE L., LEWIS M. & ZETTLEMOYER L. (2017). End-to-end neural coreference resolution. In M. PALMER, R. HWA & S. RIEDEL, Éds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 188–197, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1018.

LEE K., HE L. & ZETTLEMOYER L. (2018). Higher-order coreference resolution with coarse-to-fine inference. In M. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, p. 687–692, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-2108.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2019). BART : denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, **abs/1910.13461**.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized BERT pretraining approach. *CoRR*, **abs/1907.11692**.

MRINI K., NIE S., GU J., WANG S., SANJABI M. & FIROOZ H. (2022). Detection, disambiguation, re-ranking : Autoregressive entity linking as a multi-task problem. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Findings of the Association for Computational Linguistics : ACL 2022*, p. 1972–1983, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-acl.156.

PENNINGTON J., SOCHER R. & MANNING C. (2014). GloVe : Global vectors for word representation. In A. MOSCHITTI, B. PANG & W. DAELEMANS, Éds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543, Doha, Qatar : Association for Computational Linguistics. DOI : 10.3115/v1/D14-1162.

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. *CoRR*, **abs/1802.05365**.

PICCINNO F. & FERRAGINA P. (2014). From TagME to WAT : a new entity annotator. In *Proceedings of the first international workshop on Entity recognition & disambiguation - ERD '14*, p. 55–62, Gold Coast, Queensland, Australia : ACM Press. DOI : 10.1145/2633211.2634350.

PRADHAN S., MOSCHITTI A., XUE N., NG H. T., BJÖRKELUND A., URYUPINA O., ZHANG Y. & ZHONG Z. (2013). Towards robust linguistic analysis using OntoNotes. In J. HOCKENMAIER & S. RIEDEL, Éds., *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, p. 143–152, Sofia, Bulgaria : Association for Computational Linguistics.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, **abs/1910.10683**.

RÖDER M., USBECK R. & NGONGA NGOMO A.-C. (2018). Gerbil–benchmarking named entity recognition and linking consistently. *Semantic Web*, **9**(5), 605–625.

SANG E. F. T. K. & MEULDER F. D. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. *CoRR*, **cs.CL/0306050**.

SHAVARANI H. & SARKAR A. (2023). SpEL : Structured prediction for entity linking. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 11123–11137, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.686.

VAN HULST J. M., HASIBI F., DERCKSEN K., BALOG K. & DE VRIES A. P. (2020). REL : an entity linker standing on the shoulders of giants. *CoRR*, **abs/2006.01969**.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. *CoRR*, **abs/1706.03762**.

WU W., WANG F., YUAN A., WU F. & LI J. (2020). CorefQA : Coreference resolution as query-based span prediction. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 6953–6963, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.622.

YAMADA I., SHINDO H., TAKEDA H. & TAKEFUJI Y. (2016). Joint learning of the embedding of words and entities for named entity disambiguation. In S. RIEZLER & Y. GOLDBERG, Éds., *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, p. 250–259, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/K16-1025.

ZAPOROJETS K., DELEU J., DEVELDER C. & DEMEESTER T. (2020). DWIE : an entity-centric dataset for multi-task document-level information extraction. *CoRR*, **abs/2009.12626**.

ZAPOROJETS K., DELEU J., JIANG Y., DEMEESTER T. & DEVELDER C. (2022). Towards consistent document-level entity linking : Joint models for entity linking and coreference resolution. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 778–784, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-short.88.

ZHANG W., HUA W. & STRATOS K. (2021). Entqa : Entity linking as question answering. *CoRR*, **abs/2110.02369**.

ZHANG W., WISEMAN S. & STRATOS K. (2023). Seq2seq is all you need for coreference resolution.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Extension d'AZee avec des règles de production concernant les gestes non-manuels pour la langue des signes française

Camille Challant, Michael Filhol
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay, France
camille.challant@lisn.upsaclay.fr, michael.filhol@cnrs.fr

RÉSUMÉ

Cet article présente une étude sur les gestes non-manuels (GNM) en utilisant AZee, une approche qui permet de représenter formellement des discours en langue des signes (LS) et de les animer avec un signeur virtuel. Les GNM étant essentiels en LS et donc nécessaires à une synthèse de qualité, notre objectif est d'augmenter l'ensemble de règles de production AZee avec des règles concernant les GNM. Pour cela, nous avons appliqué la méthodologie permettant de trouver de nouvelles règles de production sur un corpus de langue des signes française, *40 brèves*. 23 règles concernant les GNM ont été identifiées. Nous avons profité de cette étude pour insérer ces règles dans le premier corpus d'expressions AZee, qui décrivent avec AZee les productions en LS du corpus *40 brèves*. Notre étude donne lieu à une nouvelle version du corpus d'expressions AZee, qui comporte 533 occurrences de règles relatives aux GNM.

ABSTRACT

Extending AZee with Non-manual Gesture Rules for French Sign Language

This paper presents a study on non-manual gestures using AZee, an approach which allows to formally represent Sign Language (SL) discourses and to animate them with a virtual signer. As the non-manual gestures are essential in SL and therefore necessary for a quality synthesis, we wanted to extend the AZee production set with non-manual production rules. For this purpose, we applied the methodology which allows to find new production rules on a French Sign Language corpus, *40 brèves*. We identified 23 production rules for non-manual gestures. We took advantage of this study to insert these new rules in the first corpus of AZee discourses expressions, which describe with AZee the productions in SL of the *40 brèves* corpus. Our study results in a new version of the AZee expressions corpus, which includes 533 occurrences of non-manual rules.

MOTS-CLÉS : Modélisation, Langue des signes, AZee, Gestes non-manuels.

KEYWORDS: Modelling, Sign Language, AZee, Non-manual gestures.

1 Introduction

Aujourd'hui encore, les langues des signes (LS) sont considérées comme des langues peu dotées. L'outillage informatique de ces langues est plus complexe que pour les langues vocales, principalement parce qu'il est nécessaire de travailler avec des vidéos, les LS ne possédant pas de système d'écriture officiellement reconnu et adopté par la communauté Sourde. Cela explique également, en partie,

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 410–421.

pourquoi les LS sont considérablement moins décrites que les langues vocales.

De plus, en LS, les gestes non-manuels (désormais GNM), c'est-à-dire l'activité de toutes les parties du corps excepté les mains (comme les mouvements des yeux, des épaules, de la tête, etc.) sont relativement peu étudiés. Bien que la plupart des études reconnaisse leur importance, l'accent est encore souvent placé sur l'activité des mains uniquement. Les GNM font pourtant entièrement partie de la langue et permettent de transmettre des informations au même titre que ce qui est signé avec les mains. Les travaux sur les GNM sont donc cruciaux, notamment pour la synthèse avec des signeurs virtuels : il a été prouvé que les expressions faciales aident considérablement les personnes Sourdes à comprendre les avatars signants (Huenerfauth *et al.*, 2011). Malgré cela, les avatars actuels animés à partir de règles [1] ne sont pas encore assez performants dans ce domaine, car les GNM ne sont pas ou peu décrits. Un aperçu de ce problème a été présenté par Wolfe & McDonald (2021).

Par ailleurs, peu de travaux en linguistique synthétisent les différents GNM que l'on peut rencontrer dans un discours en LS. Les articulateurs (parties du corps impliquées dans la production d'un discours) sont plutôt étudiés séparément, comme la bouche (Lewin & Schembri, 2011), les sourcils (de Vos *et al.*, 2009) ou encore la tête (Pfau, 2008). Les GNM sont également souvent liés à un phénomène grammatical précis (Pfau & Quer, 2010 ; Benitez-Quiroz *et al.*, 2014 ; Kimmelman *et al.*, 2020).

Afin de perfectionner les GNM des avatars pour un rendu plus naturel et compréhensible, et puisque les GNM ne sont pas encore décrits avec une approche formelle qui permettrait d'améliorer cela, nous avons choisi de travailler avec AZee, le seul modèle formel permettant la synthèse avec les avatars. Le sujet des GNM avec AZee a commencé à être abordé dans les premiers temps d'AZee (Filhol *et al.*, 2014). Depuis, quelques règles de production concernant les GNM (que nous nommerons dorénavant RGNM) ont été suggérées : on peut par exemple mentionner `inter-subjectivity`, `intensity` ou `long` qui produisent différentes expressions faciales. Cependant, aucune recherche approfondie n'a été menée sur ce sujet. L'objectif de notre étude est donc identifier de nouvelles RGNM et les ajouter au premier corpus d'expressions AZee (Challant & Filhol, 2022).

Pour cela, nous devons suivre la méthodologie permettant de trouver de nouvelles règles de production, que nous présentons dans la première section, après avoir exposé l'approche AZee. Dans la section 3, nous montrons l'application de cette méthodologie sur un corpus de langue des signes française (LSF), appelé *40 brèves*. Puis, nous présentons dans la section 4 l'ensemble des RGNM que nous avons identifiées grâce à la méthodologie, règles qui ont ensuite été insérées dans un corpus d'expressions AZee de discours. Enfin, nous exposons nos conclusions et perspectives pour ce travail dans la dernière section.

2 AZee

2.1 Présentation de l'approche

AZee est une approche formelle qui permet de décrire des discours en LS (Filhol, 2021 ; Filhol *et al.*, 2014). En premier lieu, c'est un langage fonctionnel qui permet de décrire précisément les mouvements du signeur afin d'animer des avatars, grâce à des points de l'espace de signation, des

1. Contrairement aux avatars animés par capture de mouvement qui peuvent aujourd'hui être très réalistes (Kim *et al.*, 2023).

postures, des contraintes, des vecteurs, etc.

À un niveau supérieur, linguistique, AZee associe un sens aux formes décrites avec le langage de description mentionné ci-dessus. Cela crée ce que l'on appelle une *règle de production* : une fonction qui représente un sens interprétable et qui produit un ensemble de formes observables. L'ensemble de toutes les règles de production identifiées pour une LS spécifique (dans notre cas, la LSF) forme l'*ensemble de production AZee*. Nous pouvons combiner les différentes règles de l'ensemble de production pour créer des *expressions AZee de discours*, qui représentent formellement des discours de n'importe quelle longueur en LS.

```
1 :info-about
2     'topic
3     :organiser
4     'info
5     :side-info
6         'focus
7         :élection
8         'info
9         :président
```

FIGURE 1 – Expression AZee de discours signifiant « l'organisation des élections présidentielles »

Nous donnons un exemple d'expression AZee de discours dans la figure 1, qui est la combinaison de plusieurs règles de production :

- deux règles avec deux arguments obligatoires :
  - l. 1 `info-about`(*topic, info*) signifiant 'une *info*, qui porte le focus, est donnée à propos de *topic*'
  - l. 5 `side-info`(*focus, info*) signifiant 'une *info* supplémentaire, non essentielle, est donnée à propos de *focus*'
- trois règles sans argument obligatoire, nommées d'après leur signification :
  - l. 3 `organiser`
  - l. 7 `élection`
  - l. 9 `président`

Cette expression AZee signifie « l'organisation des élections présidentielles »en LSF et génère les formes associées. Il s'agit d'un court extrait de la brève 1E-VF, issue du corpus que nous présentons dans la section 3.1.

2.2 Méthodologie

Une méthodologie a été développée pour identifier les règles de production, basée sur l'étude manuelle de corpus de LS (Hadjadj *et al.*, 2018 ; Martinod *et al.*, 2022). Elle consiste à rechercher, alternativement, des critères de forme (e.g. avancement du menton' ; 'déplacement de l'index entre la droite et la gauche') et des critères de sens (e.g. 'notion de multiplicité' ; 'négation de quelque chose') jusqu'à trouver un ensemble de formes unique pour un sens donné.

La méthodologie en question commence par un critère de forme ou de sens que l'on choisit d'étudier. La figure 2 illustre cela, en partant de s_1, un critère de sens. La première étape consiste à rechercher

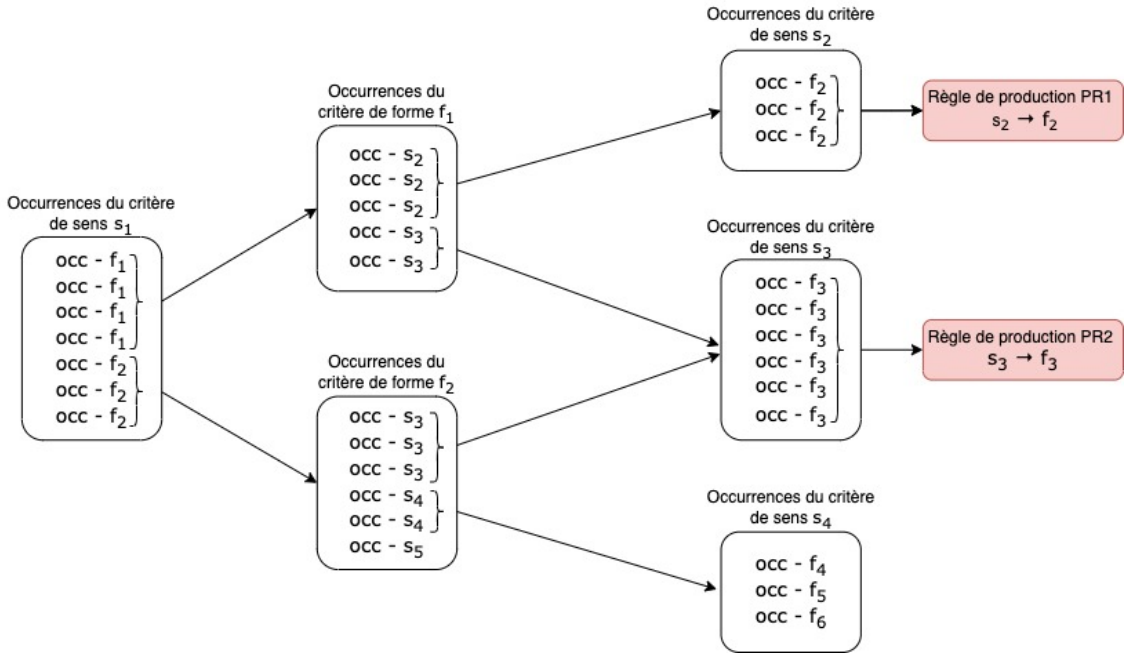


FIGURE 2 – Illustration de la méthodologie AZee, initié avec un critère de sens s_1

toutes les occurrences de s_1 dans le corpus, et à décrire, pour chaque occurrence, un ensemble de formes associées à s_1. Différents groupes de formes identiques sont alors constitués : par exemple f_1, f_2 dans la figure 2.

Chacun de ces groupes devient le critère de départ d'une nouvelle itération. Pour ces nouvelles itérations, l'objectif est maintenant de décrire la signification associée aux occurrences répertoriées. Pour f_1 dans la figure 2, nous trouvons s_2, s_3, et s_4, s_5 pour f_2.

Le processus se poursuit ensuite, jusqu'à ce qu'un seul ensemble de formes soit trouvé pour toutes les occurrences d'un sens. Par exemple, on trouve f_2 pour les occurrences du sens s_2 (figure 2, en haut à droite). Pour ce critère s_2, les formes f_2 correspondent aux formes trouvées lors de la première itération : c'est un cas qui peut se produire. Une nouvelle règle de production PR1 associant s_2 à f_2 peut être créée.

Parfois, au cours du processus, des fusions peuvent apparaître. Par exemple, le même sens s_3 est trouvé en recherchant des occurrences de f_1 et f_2, deux critères de forme différents. Une nouvelle itération est alors réalisée avec s_3 comme critère de sens, et un nouvel ensemble de formes f_3 est identifié : une nouvelle règle de production peut être créée, PR2.

3 Application de la méthodologie sur un corpus

3.1 Sélection du corpus

Nous avons décidé de travailler avec le corpus *40 brèves*, dont la deuxième version a été publiée en 2022 par Challant & Filhol.

Il s'agit d'un corpus parallèle qui aligne :

— 40 brèves journalistiques en français écrit ;
— 120 vidéos en LSF : chacune des 40 brèves a été traduite par trois traducteurs professionnels sourds, ce qui représente une heure de LSF au total ;
— 120 expressions AZee de discours, représentant formellement les 120 vidéos de LSF avec AZee, sans GNM dans la version de 2022.

Le genre journalistique est intéressant car il permet de couvrir un large éventail de sujets, tout en utilisant un style de langue quasi canonique, sans erreurs ou disfluences. Il garantit également une certaine neutralité : les GNM réalisés par les signeurs sont susceptibles de constituer un contenu pertinent au niveau sémantique, et non de donner l'opinion du signeur à propos d'un sujet, ce qui pourrait être le cas avec d'autres genres de discours.

De plus, ce corpus comprend plusieurs signeurs, ce qui est un réel avantage pour la recherche de nouvelles règles de production. Cela nous permet de nous assurer qu'une règle n'est pas spécifique à un signeur en particulier.

3.2 Application de la méthodologie

Nous avons appliqué la méthodologie présentée dans la section 2.2 sur le corpus. Deux personnes, locutrices de la LSF, ont été impliquées dans le processus, pour un total approximatif de 200 heures de travail.

Le critère de départ choisi est un critère de forme : 'GNM, produits simultanément avec l'activité des mains'. Nous nous sommes intéressés aux expressions faciales mais pas seulement : nous avons également pris en compte les mouvements du menton, de la poitrine ou même des épaules.

Nous avons utilisé un logiciel qui permet de visionner les différentes vidéos image par image ou au ralenti pour pouvoir observer les mouvements des signeurs en détail, mais aussi à vitesse réelle pour comprendre le rythme du discours et mieux en saisir le sens.

Lors de l'application de la méthodologie, il a été très difficile de décrire précisément l'ensemble des formes observées avec des mots car les différences de formes sont parfois très subtiles. Si nous prenons l'exemple des sourcils, ils peuvent avoir une multitude de positions, et pas seulement : sourcils froncés - sourcils neutres - sourcils levés. Les différentes positions étant vraiment complexes à décrire, nous avons principalement utilisé des captures d'écran des vidéos pour nous aider dans notre démarche. En terme de sens, le défi est de ne pas se laisser influencer par ce qui est signé avec les mains.

Au cours du processus d'itération, certains groupes ont fusionné, comme l'explique la figure 2 avec le critère de sens s_3 qui se trouve dans deux groupes différents de formes f_1 et f_2 et qui fusionne en un seul groupe, avec l'ensemble de formes f_3. Par exemple, nous avons identifié pour deux sens différents, à savoir 'concentration' et 'interrogation', le même ensemble de formes 'le menton avance, les sourcils sont froncés et les lèvres sont pincées'. Nous avons ensuite trouvé un dénominateur de sens commun pour les deux significations, et une règle de production appelée `closer-look` a été déterminée.

Après avoir appliqué la méthodologie, nous avons fait plusieurs choix concernant l'inclusion des règles de production dans notre ensemble de production AZee.

FIGURE 3 – Illustration des formes de trois RGNM

Premièrement, nous avons décidé de considérer une règle de production comme autonome uniquement si l'on observe que les GNM apportent un sens différent de celui qui est signé avec les mains. Dans le cas contraire, nous avons considéré que la RGNM faisait partie intégrante du signe. En effet, l'association forme-sens est au cœur de l'approche AZee, quels que soient les articulateurs impliqués dans l'ensemble des formes. Il n'y a donc pas de raison théorique de séparer en deux règles de production différentes ce qui est signé par les mains et ce qui est signé par les articulateurs non-manuels, s'ils construisent ensemble le même sens. Par exemple, lors de l'application de la méthodologie, nous avons remarqué une expression faciale qui véhicule le sens 'suspicieusement'. Cette expression faciale apparaît uniquement sur l'activité des mains signifiant 'suspicion', et vice-versa : 'suspicion' n'apparaît jamais sans cette mimique. Nous avons donc considéré que l'expression faciale *suspicieusement* est incluse dans la règle `suspicion`, sans créer une nouvelle règle de production `suspicieusement`. Le même phénomène s'est produit pour `approximativement`, qui a toujours été observé avec la même expression faciale.

Ensuite, alors que la majorité des RGNM que nous avons identifiées a été observée avec les trois signeurs du corpus (ce qui contribue à nous donner confiance dans la robustesse des règles), nous avons réalisé que certaines règles n'apparaissaient que dans les productions de l'un des trois signeurs. Le sens étant par ailleurs assez difficile à interpréter, nous avons décidé de ne pas prendre en compte ce type de GNM. C'était le cas des sourcils levés et des yeux écarquillés produits par le signeur OC. Nous avons associé ce phénomène davantage au style du signeur qu'au contenu linguistique, et nous n'avons donc pas ajouté la règle de production à l'ensemble de production.

Enfin, nous n'avions parfois qu'une ou deux occurrences d'une règle de production dans l'ensemble du corpus, avec des formes et un sens difficiles à saisir. Puisque nous ne pouvions pas les comparer avec un grand nombre d'occurrences afin d'afiner leurs formes et leur sens, nous avons préféré ne pas inclure ces règles dans notre ensemble de production. C'est le cas de certaines règles ayant le sens de 'sans doute', 'difficile à croire' ou 'affecté'.

4 Résultats

Nous avons identifié 23 RGNM dans le corpus. Elles ont toutes un seul argument, appelé *sig*.

Chaque règle de production est nommée en anglais, d'après sa signification. Des captures d'écran

illustrent les formes des trois premières règles dans la figure 3, mais une simple photo ne suffit pas à capturer l'ensemble du mouvement et à obtenir les formes exactes des règles. Pour y remédier, nous donnons entre parenthèses un exemple de chaque règle dans son contexte. Nous indiquons l'identifiant de la vidéo dans laquelle se trouve l'exemple (par exemple : 1A-JP), suivi des bornes temporelles permettant de situer l'exemple dans la vidéo en secondes et de la ligne correspondante dans l'expression AZee.

Voici la liste des 23 RGNM :

- `almost-reaching` (2H-OC, temps : 26.56–27.00, ligne 215)
- `big-threatening` (2R-VF, temps : 01.44–02.12, ligne 18)
- `closer-look` (2J-VF, temps : 29.52–31.00, ligne 222)
- `continuously` (2K-OC, temps : 18.32–18.84, ligne 155)
- `decidedly` (1K-JP, temps : 32.24–33.08, ligne 237)
- `do-you-realise` (2Q-JP, temps : 32.28–33.44, ligne 274)
- `impressive-grandiose` (1J-JP, temps : 10.96–11.92, ligne 93)
- `inter-subjectivity` (1B-OC, temps : 22.24–23.08, ligne 175)
- `it-is-a-shame` (2H-JP, temps : 32.36–33.72, ligne 279)
- `most-probably` (1R-JP, temps : 15.08–16.88, ligne 137)
- `much-almost-too-much` (1A-JP, temps : 06.28–08.52, ligne 37)
- `nothing-sticks-out` (2D-JP, temps : 09.04–10.08, ligne 67)
- `peacefully` (1R-JP, temps : 14.36–15.48, ligne 127)
- `something-sticks-out` (2O-JP, temps : 29.24–29.08, ligne 191)
- `takes-a-while` (2R-OC, temps : 23.92–24.84, ligne 256)
- `too-scared-to-look` (1O-VF, temps : 17.00–17.72, ligne 43)
- `trouble-disturbance` (2Q-JP, temps : 29.12–30.28, ligne 252)
- `uneasy-awkward` (1C-JP, temps : 25.36–26.04, ligne 250)
- `with-chaos` (1F-OC, temps : 09.44–10.28, ligne 97)
- `with-no-precision` (2C-VF, temps : 05.24–06.16, ligne 39)
- `with-surprise` (1E-JP, temps : 29.28–30.28, ligne 240)
- `with-uncertainty` (2R-VF, temps : 25.88–26.56, ligne 243)
- `with-worry` (1B-VF, temps : 20.48–22.02, ligne 175)

Dès lors que cet ensemble de règles a été établi, nous avons pu compléter le corpus *40 brèves* en appliquant ces nouvelles règles de production. La nouvelle version du corpus (v3) est mise à disposition de la communauté scientifique sur la plateforme Ortolang [2]. Au total, 533 occurrences de RGNM figurent désormais dans le corpus. La figure 4 présente la fréquence des différentes règles de production dans le corpus : chaque règle apparaît entre 2 et 89 fois.

Dans la figure 4, nous pouvons voir que les RGNM les plus fréquentes sont `decidedly`, `trouble-disturbance`, `closer-look`, `uneasy-awkard`. Cela s'explique notamment par le fait que les nouvelles journalistiques de ce corpus relatent des évènements assez graves : prise d'otage, catastrophe naturelle, rébellion contre le pouvoir, etc.

2. Le corpus est disponible à cette adresse : https://www.ortolang.fr/market/corpora/40-breves

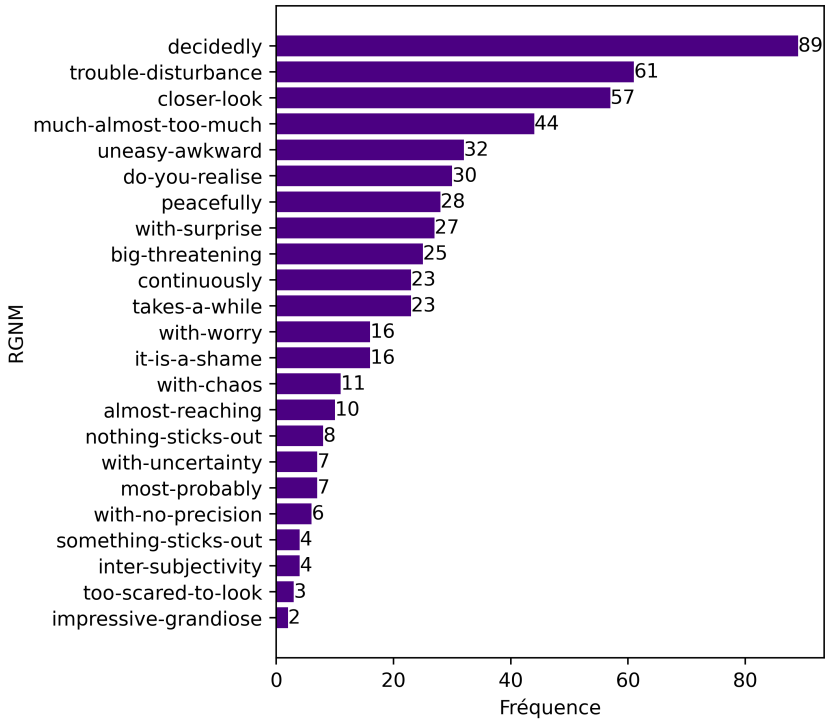


FIGURE 4 – Fréquence des RGNM dans le corpus

Par ailleurs, nous avons remarqué quelques utilisations intéressantes des RGNM, nous détaillons certains cas ci-dessous.

Tout d'abord, nous voulons souligner que les règles présentes dans le corpus ont une réelle contribution sémantique : elles sont nécessaires pour comprendre le message complet du discours. Par exemple, comme l'illustre la figure 5 avec un extrait de la brève 2H-JP, le sens de la règle de production `info-about` avec deux arguments `non tête` et `il n'y a pas` est complètement différent lorsqu'elle est incluse dans la RGNM `it-is-a-shame`. Le signeur parle ici de l'absence d'une nouvelle étoile sur les maillots des joueurs de football français, après une défaite contre l'Italie, ce qui est dommage pour l'équipe de France.

```
:it-is-a-shame
    'sig
    :info-about
        'topic
        :non tête
        'info
        :il n'y a pas
```

FIGURE 5 – Expression AZee de discours avec une RGNM qui contribue au sens du discours

Nous avons également remarqué que les RGNM peuvent parfois être combinées entre elles : le sens et les formes de chaque règle peuvent être additionnés, comme le montre la figure 6. Par exemple, dans la brève 1N-JP, `decidedly` et `with-chaos` sont combinés sur le signe 'attaquer'.

```
:decidedly
    'sig
    :with-chaos
        'sig
        :attaquer
            'patient
            ^Lssp
```

FIGURE 6 – Expressions AZee de discours avec des RGNM qui sont combinées

Les RGNM sont également employées dans des constructions élaborées, comme dans la brève 2L-OC (Figure 7). En effet, dans la règle `not-but` ci-dessous, qui présente déjà un contraste entre deux parties, les RGNM sont utilisées pour renforcer ce contraste, avec des règles qui insistent sur le sens de chaque partie : `much-almost-too-much` sur 'ventes' qui est nié, suivi de la correction avec `trouble-disturbance` sur 'ventes peu rentables', ce qui signifie que les ventes n'ont pas été à la hauteur des espérances.

```
:not-but
    'negated
    :much-almost-too-much
        'sig
        [ventes]
    'correction
    :trouble-disturbance
        'sig
        [ventes peu rentables]
```

FIGURE 7 – Expression AZee de discours avec des RGNM en contraste

Pour finir, la nouvelle version du corpus d'expressions AZee, dans laquelle nous avons ajouté les RGNM, comprend 12 442 règles de production nommées au total, soit 982 de plus que dans la version précédente. Pour calculer la fréquence des règles de production dans le corpus, nous avons utilisé toutes les règles de production ayant au moins un argument obligatoire. Comme le montre la figure 8, les règles les plus fréquentes sont les mêmes que dans la version précédente (voir Challant & Filhol (2022)) : `info-about`, `side-info`, `instance-of`... Néanmoins, cinq nouvelles règles figurent dans le top 20 : `decidedly`, `trouble-disturbance`, `closer-look`, ainsi que `mult-around` et `mult-in-a-row`, deux règles ayant été identifiées par Martinod *et al.* après la publication de la première version du corpus.

5 Conclusion et perspectives

Cet article présente l'étude que nous avons menée sur un phénomène qui n'avait encore jamais été étudié à l'aide de l'approche AZee, c'est-à-dire les GNM. Nous avons présenté la méthodologie qui permet de trouver de nouvelles règles de production AZee, que nous avons ensuite appliquée sur le

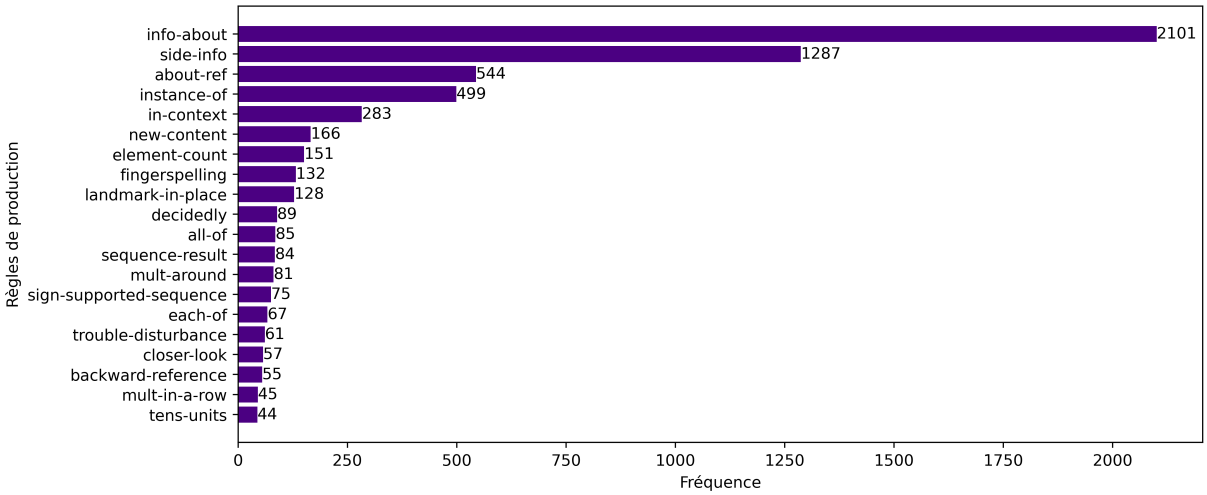


FIGURE 8 – Fréquence des 20 règles de production les plus utilisées dans le corpus

corpus *40 brèves*. 23 règles de production ont été identifiées, ce qui a permis d'augmenter l'ensemble des règles de production AZee actuellement connues pour la LSF. Enfin, nous avons complété le premier corpus d'expressions AZee existant en insérant ces règles là où elles étaient nécessaires, et nous avons rendu cette nouvelle version disponible.

Pour pouvoir proposer une évaluation des RGNM que nous avons identifiées, nous pouvons les animer sur des signeurs virtuels. Puisque chaque règle de production spécifie à la fois les formes à produire et leur synchronisation relative, les expressions AZee peuvent être directement animées sur un avatar, RGNM incluses. Ceci semble maintenant à portée de main grâce aux efforts récents, par exemple le logiciel d'expressions faciales créé par McDonald *et al.* (2022) ou la synthèse à l'aide de Blender (Sharma *et al.*, 2024). Cela constitue une première perspective pour notre travail.

Ensuite, nous avons travaillé sur un petit corpus et avec un genre particulier, il est donc probable que d'autres RGNM existent, qui n'étaient pas présentes dans notre corpus. Nous aimerions étudier d'autres types de corpus, afin de couvrir au mieux la LSF avec des RGNM et d'effectuer des comparaisons entre les genres. Par exemple, nous pourrions étudier des corpus comportant des structures plus iconiques, comme les descriptions de scènes (e.g. *Mocap1* (Benchiheub *et al.*, 2016)) ou la narration d'une histoire (*LS-COLIN* (Cuxac *et al.*, 2014)).

Enfin, ce travail sur les GNM peut être analysé du point de vue de la linguistique formelle. Différents tests pourront être effectués sur la nouvelle version, et des questions peuvent être soulevées telles que : existe-t-il des motifs réguliers pour les RGNM ou leur contexte ? Est-il possible d'avoir de très grandes expressions sous ces règles ? Ou bien ont-elles tendance à contenir des productions plus courtes ?

Références

BENCHIHEUB M.-E.-F., BERRET B. & BRAFFORT A. (2016). Collecting and Analysing a Motion-Capture Corpus of French Sign Language. In *7th International Conference on Language Resources and Evaluation - Workshop on the Representation and Processing of Sign Languages (LREC-WRPSL*

2016), p. 7–12, May 23-28, Portoroz, Slovenia : Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur - UPR 3251 (Limsi), Complexité, Innovation, Activités Motrices et Sportives - EA 4532 (CIAMS) (2020). ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr, v1, https ://hdl.handle.net/11403/mocap1/v1.

BENITEZ-QUIROZ C. F., GÖKGÖZ K., WILBUR R. B. & MARTINEZ A. M. (2014). Discriminant Features and Temporal Structure of Nonmanuals in American Sign Language. *PLoS ONE*, **9**(2), e86268. DOI : 10.1371/journal.pone.0086268.

CHALLANT C. & FILHOL M. (2022). A First Corpus of AZee Discourse Expressions. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, p. 1560–1565, Marseille, France.

CUXAC C., BOUTET D., DUBOIS C. & FIORE S. (2014). Corpus LS-Colin sur plusieurs genres discursifs (Christelle Drecours, Juliette Dalle et Stéphanie Authier). Structures formelles du langage ; Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur ; Centre d'analyses et de mathématiques sociales ; Institut de recherche en informatique de Toulouse. DOI : 10.34847/COCOON.CD3EFAF4-BB76-38DC-AF8C-6A2C65AA4C78.

DE VOS C., VAN DER KOOIJ E. & CRASBORN O. (2009). Mixed Signals : Combining Linguistic and Affective Functions of Eyebrows in Questions in Sign Language of the Netherlands. *Language and Speech*, **52**(2–3), 315—-339. DOI : 10.1177/0023830909103177.

FILHOL M. (2021). *Modélisation, traitement automatique et outillage logiciel des langues des signes*. Habilitation à diriger des recherches, Université Paris-Saclay. HAL : tel-03197108.

FILHOL M., HADJADJ M. & CHOISIER A. (2014). Non-Manual Features : The Right to Indifference. In *International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

HADJADJ M., FILHOL M. & BRAFFORT A. (2018). Modeling French Sign Language : A proposal for a semantically compositional system. In *International Conference on Language Resources and Evaluation*, p. 7, Miyazaki, Japan : ELRA.

HUENERFAUTH M., LU P. & ROSENBERG A. (2011). Evaluating Importance of Facial Expression in American Sign Language and Pidgin Signed English Animations. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, p. 99–106, Dundee Scotland, UK : ACM. DOI : 10.1145/2049536.2049556.

KIM C., CHA H.-S., KIM J., KWAK H., LEE W. & IM C.-H. (2023). Facial Motion Capture System Based on Facial Electromyogram and Electrooculogram for Immersive Social Virtual Reality Applications. *Sensors*, **23**(7), 3580. DOI : 10.3390/s23073580.

KIMMELMAN V., IMASHEV A., MUKUSHEV M. & SANDYGULOVA A. (2020). Eyebrow Position in Grammatical and Emotional Expressions in Kazakh-Russian Sign Language : A quantitative study. *PLOS ONE*, **15**(6). DOI : 10.1371/journal.pone.0233731.

LEWIN D. & SCHEMBRI A. C. (2011). Mouth gestures in British Sign Language : A case study of tongue protrusion in BSL narratives. *Sign Language and Linguistics*, **14**(1), 94–114. DOI : 10.1075/sll.14.1.06lew.

MARTINOD E., DANET C. & FILHOL M. (2022). Two new AZee production rules refining multiplicity in French Sign Language. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages : Multilingual Sign Language Resources*, p. 132–138, Marseille, France : European Language Resources Association.

MCDONALD J., JOHNSON R. & WOLFE R. (2022). A Novel Approach to Managing Lower Face Complexity in Signing Avatars. In *Proceedings of the 7th International Workshop on Sign Language*

Translation and Avatar Technology : The Junction of the Visual and the Textual : Challenges and Perspectives, p. 67–72, Marseille, France : European Language Resources Association.

PFAU R. (2008). The Grammar of Headshake : A Typological Perspective on German Sign Language Negation. *Linguistics in Amsterdam*, **1|1**, 37–74.

PFAU R. & QUER J. (2010). *Nonmanuals : their Grammatical and Prosodic Roles*, In *Sign Languages, Cambridge Language Surveys*, p. 381–402. Cambridge University Press.

SHARMA P., CHALLANT C. & FILHOL M. (2024). Facial Expressions for Sign Language Synthesis using FACSHuman and AZee. In *Proceedings of the 11th Workshop on the Representation and Processing of Sign Languages*, Torino, Italy.

WOLFE R. & MCDONALD J. C. (2021). A Survey of Facial Nonmanual Signals Portrayed by Avatar. *Graz University Library*, **Vol. 93**, 161–223. DOI : 10.25364/04.48 :2021.93.6.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Extraction d'entités nommées décrivant des chaînes de traitement bioinformatiques dans des articles scientifiques en anglais

Clémence Sebe[1] Sarah Cohen-Boulakia[1] Olivier Ferret[2] Aurélie Névéol[1]
(1) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
(2) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[1]prenom.nom@lisn.upsaclay.fr, [2]olivier.ferret@cea.fr

RÉSUMÉ

Les chaînes de traitement d'analyses de données biologiques utilisées en bioinformatique sont une solution pour la portabilité et la reproductibilité des analyses. Ces chaînes figurent à la fois sous forme descriptive dans des articles scientifiques et/ou sous forme de codes dans des dépôts. L'identification de publications scientifiques décrivant de nouvelles chaînes de traitement et l'extraction de leurs informations sont des enjeux importants pour la communauté bioinformatique. Nous proposons ici d'étendre le corpus *BioToFlow* ayant trait aux articles décrivant des chaînes de traitement bioinformatiques et de l'utiliser pour entraîner et évaluer des modèles de reconnaissance d'entités nommées bioinformatiques. Ce travail est accompagné d'une discussion critique portant à la fois sur le processus d'annotation du corpus et sur les résultats de l'extraction d'entités.

ABSTRACT

Extracting named entities describing bioinformatic workflows from the literature in English.

Workflows used in bioinformatic are a solution for analysis portability and reproducibility. These workflows are either described in publications and/or are in source code in repositories. Identifying new workflows in scientific articles and extracting related information is a challenge for the bioinformatics community. Herein, we propose to extend a corpus of articles describing bioinformatic workflows (*BioToFlow*) and to use it to train and evaluate bioinformatics named entity recognition models. We also engage in a critical discussion of both the corpus annotation process and the results of information extraction.

MOTS-CLÉS : Chaînes de traitement bioinformatiques, Annotation, Reconnaissance d'entités nommées.

KEYWORDS: Bioinformatic workflows, Annotation, Named entity recognition.

1 Introduction

La biologie est un domaine dans lequel l'arrivée de nouvelles technologies dites à haut-débit a permis l'acquisition de très grands volumes de données biologiques. Ces données brutes sont nombreuses mais aussi très hétérogènes et l'enjeu de la bioinformatique est de croiser, intégrer et analyser ces données pour faire avancer les connaissances en biologie. Face aux masses de données disponibles, la communauté bioinformatique a développé un grand nombre d'outils bioinformatiques. Une analyse de données bioinformatique consiste en l'enchaînement d'un ensemble d'outils bioinformatiques,

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 422–434.

chaque outil consommant les données brutes en entrée et générant de nouvelles données, consommées à leur tour par l'outil suivant. Ces analyses peuvent être facilement implémentées via un script python ou un notebook pour des analyses rapides et impliquant de petits volumes de données. Mais lorsqu'il convient d'automatiser des analyses complexes pouvant impliquer l'utilisation de plusieurs dizaines d'outils et des volumes importants de données, il est nécessaire d'utiliser des solutions adaptées, plus faciles à déployer sur des grappes de calculs et offrant des solutions pour la portabilité et la reproductibilité de l'analyse. Les *systèmes de workflows scientifiques* ont été conçus pour répondre à ces besoins. Deux systèmes sont en particulier de plus en plus utilisés dans la communauté bioinformatique : Nextflow (Di Tommaso *et al.*, 2017) et Snakemake (Mölder *et al.*, 2021). Dans ces systèmes, une chaîne de traitement (ou *workflow*) est un code structuré où les étapes d'analyse sont bien distinguables. Lorsqu'une chaîne de traitement originale est conçue, le code est mis à disposition de la communauté dans un dépôt github et la description de la chaîne de traitement est publiée sous la forme d'un article scientifique dans une revue bioinformatique. Au 9 février 2024, le nombre d'articles ayant un lien github vers une chaîne de traitement dans le système de gestion Nextflow est ainsi de 89 articles et 91 pour le système Snakemake [1].

Un enjeu important pour la communauté bioinformatique est d'identifier les articles scientifiques décrivant une nouvelle chaîne de traitement et d'extraire les informations qui lui sont associées (par exemple, les outils bioinformatiques ou les données utilisées). À terme, la perspective de ce travail est, pour une même chaîne de traitement, de pouvoir comparer et fusionner des informations extraites des articles scientifiques d'une part et des codes issus de dépôts publics d'autre part.

Dans ce contexte, nous avons récemment proposé une première méthode de modélisation et d'extraction des composants des chaînes de traitement avec un schéma décrivant un ensemble d'entités nommées et de relations (Sebe *et al.*, 2023). Nous avons par ailleurs introduit le corpus *BioToFlow*, composé de 24 articles décrivant des chaînes de traitement (20 Nextflow et 4 Snakemake) et annotés par 3 annotateurs. Des expériences de reconnaissance d'entités ont été réalisées sur ce corpus et ont généré de premiers résultats prometteurs.

Cet article présente la suite de ce travail : nous nous focalisons sur l'extraction d'entités nommées à l'aide d'un ensemble plus important et plus varié d'articles, annotés par un ensemble plus important d'annotateurs. Plus précisément, nos contributions sont les suivantes :

— l'introduction et l'annotation de façon croisée par quatre annotatrices d'un corpus de 52 articles en anglais étendant *BioToFlow* et décrivant des chaînes de traitement bioinformatiques issues à la fois des systèmes Nextflow (26 articles) et Snakemake (26 articles) ;
— l'étude de l'utilisation de ce corpus pour l'entraînement et l'évaluation de modèles pour l'extraction automatique d'entités relatives aux chaînes de traitement ;
— une discussion critique des résultats de cette étude et des performances des modèles obtenus au niveau des différentes classes d'entités nommées.

1. Ce qui correspond à l'extraction d'articles de PubMed Central via la requête *(nextflow[Abstract] OR snakemake[Abstract] OR nextflow[Title] OR snakemake[Title]) AND github[All Fields]*

2 Extension du corpus annoté *BioToFlow*

2.1 Corpus *BioToFlow*

BioToFlow [2] est un corpus contenant 24 articles scientifiques (issus de revues telles Bioinformatics ou F1000Research) annotés manuellement à l'aide de 16 entités modélisant la composition des chaînes de traitement. C'est un corpus de petite taille avec une répartition déséquilibrée des articles relatifs aux chaînes de traitement issues des différents systèmes de gestion (4 Snakemake et 20 Nextflow).

Schéma d'annotation. L'ensemble des entités considérées dans *BioToFlow* est décrit à la figure 1. Les entités sont relatives à la chaîne de traitement elle-même (partie gauche) et à ses composants (partie droite). Chaque chaîne de traitement est désignée par un nom (*WorkflowName*). L'analyse qu'elle effectue fait référence à des méthodes algorithmiques (*Method*) implémentées par des outils bioinformatiques (*Tool*). Une chaîne de traitement met en jeu des données (*Data*) attendues dans un format de fichier particulier (*file*) et/ou issues d'une base de données (*Database*). Sur un plan technique, une chaîne de traitement est implémentée dans un langage de programmation (*ProgrammingLanguage*) et peut faire appel, pour s'exécuter, à un environnement d'exécution (*Environnement*), un système de conteneur (*Container*), un système de gestion de workflow (*ManagementSystem*) et nécessiter l'utilisation de bibliothèques (*LibraryPackage*) ou d'infrastructures particulières (*Hardware*). Chaque composant de la chaîne de traitement peut avoir un numéro de version (*Version*), une description (*Description*), des paramètres de lancement (*Parameter*) et des informations bibliographiques (*Biblio*).

Les annotations issues de ce schéma réalisées sur les articles du corpus *BioToFlow* sont considérées dans ce qui suit comme le *gold standard* pour les processus de reconnaissance d'entités nommées.

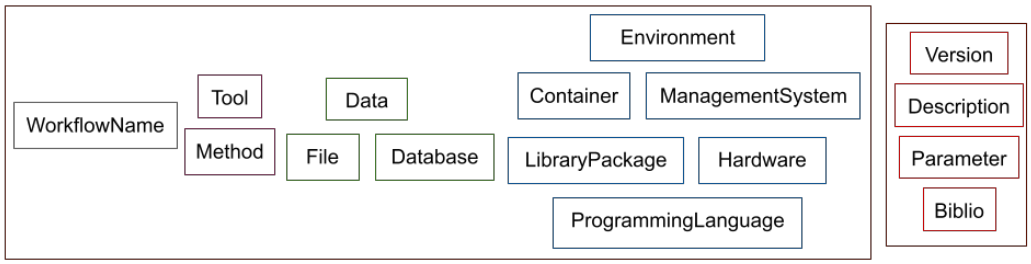


FIGURE 1 – Entités caractéristiques d'un workflow bioinformatique. À gauche : entités constitutives d'un workflow. À droite : entités liées à ses composants.

2.2 Nouveaux articles annotés

Choix des nouveaux articles. Nous avons suivi la méthodologie d'extraction des articles issus de PubMed Central de Sebe *et al.* (2023) pour augmenter la taille du corpus avec l'ajout de 28 nouveaux articles rééquilibrant le nombre d'articles décrivant des chaînes de traitement Nextflow et Snakemake :

2. https://doi.org/10.5281/zenodo.10650467

nous avons choisi de manière aléatoire parmi l'ensemble des publications disponibles 22 nouveaux articles décrivant des chaînes de traitement sous Snakemake et 6 sous Nextflow. La nouvelle version du corpus contient donc un total de 52 articles, avec 26 articles décrivant des chaînes de traitement de chaque système (Nextflow et Snakemake).

Démarche d'annotation des 28 nouveaux articles. Les 28 nouveaux articles ont été annotés en termes d'entités par quatre annotatrices (bio)informaticiennes selon la méthodologie décrite par Fort (2016). Le logiciel BRAT (Stenetorp *et al.*, 2012) et l'outil BRAT-Eval (Verspoor *et al.*, 2013) ont été respectivement utilisés pour l'annotation de ces articles et l'évaluation de la qualité de l'annotation (accord inter-annotatrices). La tâche d'annotation s'est effectuée en deux phases.

Phase 1 d'annotation : formation des annotatrices. L'objectif de la première phase était de former les annotatrices. Cinq articles du corpus initial *BioToFlow* ont été utilisés pour ce faire en enlevant les annotations de référence mais en utilisant comme préannotation les résultats d'un modèle entraîné sur le reste de ce corpus. Les quatre annotatrices ont annoté ces cinq articles à l'aide du guide d'annotation de Sebe *et al.* (2023). La qualité de leur annotation a été évaluée en comparant les annotations obtenues et celles du gold standard (score F1) issu de la version initiale de *BioToFlow*. La durée d'annotation des cinq articles varie entre deux et trois heures selon les annotatrices. Cette première phase a été concluante, avec un score F1 de 73 % à 83 % en mode strict (deux portions de texte doivent être strictement identiques et avoir la même étiquette) et 84 % à 88 % en mode relâché (où le recouvrement entre deux entités avec la même étiquette est accepté).

Phase 2 d'annotation : annotation des nouveaux articles. La seconde phase d'annotation a porté sur les 28 nouveaux articles à annoter. Des exemples de phrases annotées sont présentés dans l'Annexe A. Chaque annotatrice a annoté entre 8 et 17 articles. Il est à noter que ces articles sont plus longs et de nature plus variée que ceux du corpus initial *BioToFlow*. Alors que la version initiale de *BioToFlow* avec 24 articles contient 29 577 tokens et 10 125 tokens annotés (34 %), le nouvel ensemble de 28 articles regroupe 48 842 tokens, dont 17 661 sont annotés (36 %).

Statistiques. La version étendue de *BioToFlow* proposée est composée de 52 articles pour un ensemble de 78 419 tokens (dont 27 786 tokens annotés). Les entités sont réparties selon le tableau 1.

Entités	Occurrences	Entités	Occurrences
Data	2 434	Version	454
Tool	1 482	Hardware	429
Description	1 300	Database	288
Biblio	1 251	ManagementSystem	243
Method	936	Container	108
WorkflowName	851	ProgrammingLanguage	104
File	780	LibraryPackage	101
Parameter	464	Environment	83

TABLE 1 – Nombre d'entités par catégorie dans la version étendue de *BioToFlow*.

Les entités sont distribuées de façon très variable avec un nombre important d'occurrences pour les entités générales de la chaîne de traitement qui décrivent les données (gènes, protéines...), indiquent les noms d'outils utilisés ou encore le nom de cette chaîne de traitement. Au contraire, des informations plus techniques n'apparaissent pas toujours dans les articles scientifiques. Il en résulte un nombre plus réduit d'occurrences pour des entités telles que *Version* ou *Environnement* par exemple.

2.3 Qualité des annotations

Le tableau 2 donne les accords inter-annotateurs obtenus entre les différentes annotatrices deux à deux pour les nouveaux articles annotés. Les scores calculés sont tous supérieurs à 70 % en mode relâché : ceci signifie que le guide d'annotation est suffisamment intelligible, complet et non ambigu et qu'il existe peu de divergences dans la manière d'annoter des quatre annotatrices.

	A2	**A3**	**A4**
A1	*66,7* 72,7	*83,2* 86,3	*79,8* 80,6
A2		*69,7* 75,1	*66,8* 73,0

TABLE 2 – Accord inter-annotateur en *mode strict* et en mode relâché entre annotatrices ayant annoté des articles communs (en pourcentage).

Toutefois, le tableau 3 présente le détail des accords pour chaque type d'entité. L'hétérogénéité observée suggère que certaines entités sont plus simples à annoter que d'autres (Fort *et al.*, 2012).

Entités	**P**	**R**	**F1**	**Entités**	**P**	**R**	**F1**
Biblio	*96,8* 96,8	*99,6* 99,6	*98,2* 98,2	Environment	*57,1* 57,1	*52,6* 52,6	*54,8* 54,8
Container	*80,0* 80,0	*100,0* 100,0	*88,9* 88,9	ManagementSystem	*90,7* 91,5	*95,5* 96,4	*93,0* 93,9
Data	*65,8* 70,4	*62,7* 67,3	*64,2* 68,8	Method	*57,7* 61,9	*61,7* 66,2	*59,6* 64,0
Description	*56,5* 62,8	*68,1* 75,5	*61,7* 68,6	Tool	*76,8* 80,5	*79,8* 84,0	*78,2* 82,2

TABLE 3 – Détail des scores moyens en pourcentage obtenus pour certaines entités en *mode strict* et en mode relâché.

Tandis que des entités telles que *Biblio* ou *Container* ont des scores d'accord élevés, démontrant que la tâche d'annotation pour ces entités est simple, d'autres entités obtiennent au contraire des scores très inférieurs. Nous analysons ci-après trois causes possibles de ce constat.

La première cause identifiée est celle de l'ambiguïté dans le tagset (Fort *et al.*, 2012). C'est le cas des entités *Description* et *Method*. Après échange avec les annotatrices, il ressort que certains articles

scientifiques décrivent les méthodes bioinformatiques sans nécessairement les nommer. Les entités *Description* et *Method* sont alors souvent imbriquées et pas toujours délimitées de façon identique. Dans Zhang & Jonassen (2020) par exemple, la phrase « When the user is satisfied with the quality of the reads, the workflow proceeds to the next step : *quantification of read abundance or expression level* for transcripts or genes », l'entité *quantification of read abundance or expression level* a été annotée soit en tant que *Description*, soit en tant que *Method* selon les annotatrices.

Une deuxième cause identifiée est relative au critère de discrimination de Fort *et al.* (2012). Par exemple, les entités *Data* et *Description* ont une quantité d'annotations variant très fortement d'une annotatrice à l'autre. Certains articles sont rédigés par des bioinformaticiens, décrivant de façon précise les aspects méthodologiques et techniques, tandis que d'autres sont centrés sur le résultat biologique fourni. Dans cette seconde catégorie d'articles, les termes désignant des objets biologiques (gènes, RNA, SNP…) ont été annotés en *Data* par certaines annotatrices puisqu'il s'agit de la désignation de données. D'autres annotatrices les ont annotés en tant que *Description* ou pas annoté du tout, considérant qu'il s'agissait du contexte (biologique) de l'article.

Une troisième cause identifiée est le domaine d'expertise de l'annotatrice, qui influe sur son choix d'annotation. Par exemple, des bibliothèques telles que *Numpy* ou *Scikit-Learn* sont parfois annotées comme des outils bioinformatiques par des annotatrices ayant une formation initiale en biologie tandis que les bioinformaticiennes issues de l'informatique vont les annoter comme des *LibraryPackage*.

3 Expériences d'extraction d'entités nommées

3.1 Cadre expérimental

Choix du modèle et de son implémentation. Pour l'extraction de nos entités cibles, nous avons choisi le modèle neuronal biLSTM-CRF de Wajsbürt (2021), implémenté par l'outil NLStruct [3], dans sa version 0.2.0. Ce modèle est en effet capable de prendre en compte les entités imbriquées, ce qui est nécessaire pour certaines de nos entités, et a par ailleurs montré de bonnes performances dans le domaine biomédical, proche du nôtre. Nous avons plus précisément utilisé ce modèle dans deux configurations : d'une part avec le modèle de langue BERT (Devlin *et al.*, 2019), entraîné en domaine général ; d'autre part avec le modèle de langue SciBERT (Beltagy *et al.*, 2019), plus spécifiquement entraîné à partir d'articles scientifiques et a priori plus adapté à notre cas de figure.

Expériences réalisées. Nous avons classiquement choisi de répartir les articles en deux grands ensembles : un premier ensemble regroupant 75 % du corpus pour constituer le jeu d'entraînement (soit 39 articles) et un second ensemble correspondant aux 25 % restants pour le jeu de test (soit 13 articles), la séparation entre ces deux ensembles se faisant par tirage aléatoire. Au sein du jeu d'entraînement, 2/3 des articles (soit 26 articles) sont utilisés pour l'entraînement proprement dit des modèles et 1/3 pour leur validation (soit 13 articles). La particularité ici est que nous avons construit quatre volets pour ce découpage, par tirage aléatoire, afin de limiter la dépendance à un découpage particulier. Pour chaque découpage, quatre versions du modèle de reconnaissance d'entités sont produites avec des graines aléatoires différentes. Les hyperparamètres sont donnés en Annexe B.

3. https://github.com/percevalw/NLStruct

3.2 Résultats obtenus

Les scores obtenus par les modèles de langue testés BERT et SciBERT en faisant varier les graines aléatoires et les différents jeux d'entraînement et de validation sont présentés dans le tableau 4. L'empreinte carbone de tous les entraînements et évaluations est équivalent à 413 grammes de CO_2, calculée à l'aide de Green Algorithms [4]. Les scores entre chacun de nos volets sont globalement proches, la différence entre les volets extrêmes ne dépassant pas 1,7 point de F1 dans les deux configurations. Nous avons utilisé le test *Almost Stochastic Order* (Dror *et al.*, 2019) avec un niveau de confiance de 0,05 pour mesurer la significativité entre les deux modèles et obtenons que les modèles entraînés sur SciBERT sont stochastiquement dominants par rapport à ceux entraînés sur BERT ($\epsilon_{min} = 0$). Dans chacun des cas, utiliser un modèle de langue pré-entraîné à partir d'un corpus d'articles scientifiques issus des domaines biomédical et informatique (SciBERT) est plus performant.

	BERT			**SciBERT**		
	P	**R**	**F1**	**P**	**R**	**F1**
V1	65,2 ± 0,7	68,5 ± 0,6	66,8 ± 0,3	70,4 ± 0,3	70,2 ± 0,6	70,3 ± 0,4
V2	66,7 ± 0,3	66,8 ± 0,2	66,8 ± 0,1	70,7 ± 0,4	68,8 ± 0,6	69,7 ± 0,1
V3	66,9 ± 0,2	67,6 ± 0,6	67,3 ± 0,3	70,5 ± 0,5	71,0 ± 0,4	70,7 ± 0,4
V4	68,0 ± 0,3	69,0 ± 0,4	68,5 ± 0,2	71,2 ± 0,5	71,6 ± 1,0	71,4 ± 0,6
All	66,7 ± 0,9	68,0 ± 0,9	67,3 ± 0,6	70,7 ± 0,5	70,4 ± 1,1	70,5 ± 0,7

TABLE 4 – Moyenne des scores (et écarts-types) en pourcentage obtenus avec l'outil NLStruct en mode relâché pour chaque volet de découpage entraînement/validation. La moyenne des scores est calculée en fonction des résultats obtenus pour chaque graine aléatoire.

Le détail des scores de certaines entités sont données en *mode strict* et en mode relâché dans le tableau 5. Les scores F1 varient de 30 % à 98 %. Les scores faibles s'accordent avec les difficultés mises en exergue lors de l'annotation manuelle. L'entité *Method* a, en particulier, plus de mal à être extraite. Au contraire, d'autres (*Biblio* ou *Container*) obtiennent des scores F1 supérieurs à 80 %.

Les premiers résultats obtenus sur ce nouveau corpus étendu sont légèrement inférieurs (67,3 % avec le modèle BERT et 70,5 % avec SciBERT) à ceux obtenus par Sebe *et al.* (2023), dont les scores sont de 70,7 % avec un modèle entraîné sur BERT et 72,4 % sur un modèle entraîné avec SciBERT. Par ailleurs, la stabilité des résultats obtenus pour nos différents volets suggère que la taille du corpus annoté est maintenant suffisante pour l'entraînement des modèles de reconnaissance d'entités bioinformatiques. Pour mieux comprendre le fonctionnement des modèles testés, nous avons étudié la capacité des modèles à mémoriser.

3.3 Mémorisation et généralisation

Afin de déterminer l'impact de la mémorisation sur les performances de l'extraction d'entités, nous avons évalué les performances d'une baseline réalisant une simple projection des entités du corpus d'entraînement sur le corpus de test, en fonction de nos différents jeux d'entraînement avec un outil

4. http://calculator.green-algorithms.org/

Entités	P	R	F1	Entités	P	R	F1
Biblio	*94,1* 96,1	*96,4* 98,1	*95,2* 97,1	Environment	*64,1* 65,2	*90,9* 92,5	*75,1* 76,3
Container	*92,9* 92,9	*81,9* 81,9	*86,9* 86,9	ManagementSystem	*65,0* 66,7	*81,5* 83,6	*72,3* 74,1
Data	*51,5* 62,6	*45,7* 56,4	*48,4* 59,3	Method	*26,0* 30,0	*54,5* 63,3	*35,2* 40,7
Description	*37,3* 58,7	*36,5* 58,1	*36,9* 58,4	Tool	*65,7* 72,7	*63,7* 69,1	*64,7* 70,9

TABLE 5 – Détail des scores moyens en pourcentage obtenus pour certaines entités en *mode strict* et en mode relâché pour le premier volet de l'outil NLStruct (moyenne des résultats obtenus en fonction des différentes graines aléatoires).

proposé par Grouin (2016). Le tableau 6 présente les performances obtenues, qui sont globalement très inférieures à celles des modèles d'extraction des entités.

	P	R	F1
Propagation	*23,6* $\pm$ *0,3* 45,8 $\pm$ 0,4	*13,83* $\pm$ *0,7* 30,6 $\pm$ 1,7	*17,4* $\pm$ *0,6* 36,6 $\pm$ 1,1

TABLE 6 – Moyenne des scores (et écarts-types) en pourcentage obtenus avec un outil de propagation d'annotation en *mode strict* et en mode relâché en fonction de chaque jeu d'entraînement.

Entités	P	R	F1	Entités	P	R	F1
Biblio	*67,5* 81,1	*33,5* 43,4	*44,8* 56,5	Environment	*39,1* 75,0	*37,7* 61,0	*33,1* 59,6
Container	*92,9* 92,9	*100,0* 100,0	*96,3* 96,3	ManagementSystem	*89,6* 89,6	*96,1* 96,1	*92,2* 92,2
Data	*17,6* 61,6	*6,2* 26,3	*9,1* 36,5	Method	*8,6* 25,6	*7,5* 23,7	*8,0* 24,6
Description	*3,7* 21,8	*3,6* 22,3	*3,7* 22,0	Tool	*29,4* 34,2	*41,1* 49,4	*34,2* 40,4

TABLE 7 – Moyenne des scores obtenus en pourcentage avec l'outil de propagation d'annotation en *mode strict* et en mode relâché en fonction des différents jeux d'entraînement sur certaines entités.

Dans le tableau 7 relatif aux résultats de certaines entités, on observe une grande hétérogénéité dans la distribution des scores. Ainsi en *mode strict*, trois catégories d'entités semblent se distinguer.

La première catégorie correspond aux entités pour lesquelles la mémorisation fonctionne très bien. Il s'agit des entités prenant un nombre restreint de valeurs, telles les entités *Container* ou *Management-*

System. Les performances obtenues dans ce cas par la simple mémorisation sont meilleures que celles des modèles neuronaux.

La deuxième catégorie regroupe les entités pour lesquelles la mémorisation présente des performances moyennes. C'est le cas de l'entité *Tool* : certains outils bioinformatiques génériques sont communs à de nombreuses chaînes de traitement alors que d'autres correspondent à des outils spécifiques à chaque domaine bioinformatique. Le modèle sait reconnaître les outils génériques souvent cités car ils ont été annotés précédemment dans le corpus *BioToFlow* mais ne peut distinguer les outils spécifiques si ces derniers n'y figuraient pas. Ainsi, comme pour les entités où la mémorisation est efficace, on peut penser que l'injection de connaissances spécifiques au domaine dans les modèles neuronaux pourrait être bénéfique pour la reconnaissance d'entités. La contribution potentielle d'une telle injection est d'autant plus intéressante dans le domaine bioinformatique que des bases de connaissances recensant des informations sur les outils comme BioTools [5] (Ison *et al.*, 2016), les conteneurs et systèmes de gestion des chaînes de traitement existent déjà.

La dernière catégorie correspond aux entités que le modèle ne peut mémoriser, par exemple *Description* et *Method*. Ces entités sont complexes car pouvant être composées d'un ou plusieurs tokens. Les modèles devront posséder la capacité de généraliser pour extraire ce type d'entités.

3.4 Comparaison des chaînes de traitement Nextflow et Snakemake

Une autre perspective à explorer pour l'amélioration des scores d'extraction consiste à traiter les articles relatifs aux deux systèmes de gestion Nextflow et Snakemake de manière séparée. De fait, même s'il s'agit de deux systèmes très utilisés dans la communauté bioinformatique, il semblerait qu'ils soient utilisés par des communautés d'utilisateurs distinctes, ce qui pourrait impliquer des différences de style et de structure des articles.

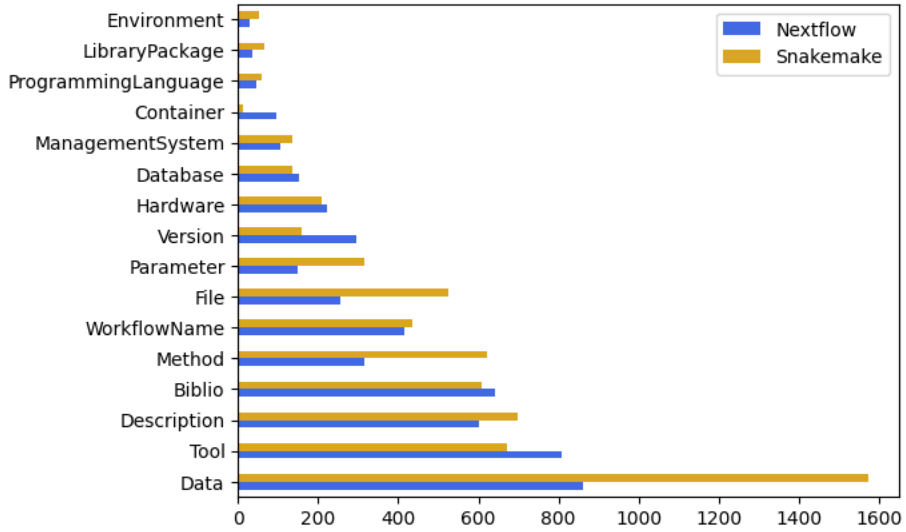


FIGURE 2 – Occurrences des entités figurant dans le corpus enrichi selon les systèmes de gestion Nextflow et Snakemake.

5. https://bio.tools/

La figure 2 représente la répartition de chacun des types d'entités dans les articles relatifs au système Nextflow et dans ceux relatifs au système Snakemake. On constate que les articles décrivant des chaînes de traitement sous Snakemake comportent presque le double de mentions d'entités *Data* et *Method* que ceux se rapportant à Nextflow. A contrario dans les articles relatifs à Nextflow, figurent plus souvent les noms des outils bioinformatiques utilisés. Tester deux modèles de langue différents est donc une piste à explorer pour l'amélioration des scores d'extraction.

4 Conclusion et perspectives

L'intersection entre le domaine du traitement automatique des langues et celui de la bioinformatique ouvre de nouvelles perspectives pour la recherche et l'analyse des chaînes de traitement bioinformatiques contenues dans la littérature. Ainsi, nous présentons ici deux contributions portant sur l'introduction d'un corpus de 52 articles étendant *BioToFlow* et sur l'utilisation de ce corpus pour l'entraînement et l'évaluation de modèles d'extraction d'entités.

Constitution du corpus *BioToFlow* étendu. Nous proposons une nouvelle version du corpus *BioToFlow* [6] composée de 52 articles variés annotés en termes d'entités nommées. Ce corpus est riche en entités, tant en nombre qu'en variété.

Lors de l'annotation manuelle par de nouvelles annotatrices, de nouveaux questionnements ont émergé sur la définition de certaines entités, notamment la distinction entre les entités *Method* et *Description*. Ces points devront être rediscutés entre annotatrices afin de diminuer les désaccords et le guide d'annotation devra être mis à jour en conséquence. Une fois ce travail réalisé, nous pouvons espérer un accroissement des performances de nos modèles d'extraction d'entités nommées.

Méthodes d'extraction d'entités nommées. Les résultats d'extraction des entités nommées obtenus sur notre corpus d'articles sont hétérogènes. Certaines entités obtiennent de très bons scores d'extraction, supérieurs à 80 %, alors que d'autres présentent des scores très inférieurs. Ceci s'explique par la combinaison des facteurs suivants :

— l'ambiguïté citée dans le paragraphe précédent concernant la distinction entre les entités *Method* et *Description* ;
— les inégalités élevées en matière de fréquence d'occurrence des entités ;
— les limites des modèles de langue utilisés en matière de connaissance du vocabulaire spécifique à la bioinformatique et aux chaînes de traitement, en particulier en ce qui concerne les noms des outils bioinformatiques.

Ce dernier point ne pourra être résolu que par l'injection de connaissances spécifiques au domaine des chaînes de traitement bioinformatiques dans les modèles de langue.

Futurs travaux L'extraction des chaînes de traitement dans les articles scientifiques nécessite non seulement l'extraction de leurs constituants, ce qui est l'objet du travail présenté, mais également des relations qu'ils entretiennent. La prochaine étape de ce travail est donc la définition de modèles pour

6. https://doi.org/10.5281/zenodo.11204427

l'extraction de ces relations. À plus long terme, la liaison référentielle entre les composants extraits des chaînes de traitement dans la littérature scientifique en anglais et ceux se trouvant dans les codes issus de dépôts publics sera nécessaire pour fusionner les informations venant des deux sources.

Remerciements

Nous remercions Noémie Bossut et Marie Schmit pour leur précieuse contribution à l'annotation des publications scientifiques. Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre de France 2030 portant la référence ANR-22-PESN-0007.

Références

BELTAGY I., LO K. & COHAN A. (2019). SciBERT : A Pretrained Language Model for Scientific Text. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3615–3620, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1371.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DI TOMMASO P., CHATZOU M., FLODEN E. W., BARJA P. P., PALUMBO E. & NOTREDAME C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, **35**(4), 316–319. Number : 4 Publisher : Nature Publishing Group, DOI : 10.1038/nbt.3820.

DROR R., SHLOMOV S. & REICHART R. (2019). Deep Dominance - How to Properly Compare Deep Neural Models. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 2773–2785, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1266.

FORT K. (2016). *Collaborative annotation for reliable natural language processing : Technical and sociological aspects*. John Wiley & Sons.

FORT K., NAZARENKO A. & ROSSET S. (2012). Modeling the complexity of manual annotation tasks : a grid of analysis. In M. KAY & C. BOITET, Éds., *Proceedings of COLING 2012*, p. 895–910, Mumbai, India : The COLING 2012 Organizing Committee.

GROUIN C. (2016). Controlled Propagation of Concept Annotations in Textual Corpora. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 4075–4079, Portorož, Slovenia : European Language Resources Association (ELRA).

ISON J., RAPACKI K., MÉNAGER H. & AL (2016). Tools and data services registry : a community effort to document bioinformatics resources. *Nucleic Acids Research*, **44**(D1), D38–D47. DOI : 10.1093/nar/gkv1116.

MÖLDER F., JABLONSKI K. P., LETCHER B., HALL M. B., TOMKINS-TINCH C. H., SOCHAT V., FORSTER J., LEE S., TWARDZIOK S. O., KANITZ A., WILM A., HOLTGREWE M., RAHMANN S., NAHNSEN S. & KÖSTER J. (2021). *Sustainable data analysis with Snakemake*. Rapport interne 10 :33, F1000Research. Type : article, DOI : 10.12688/f1000research.29032.1.

SEBE C., NÉVÉOL A., COHEN-BOULAKIA S. & GAIGNARD A. (2023). Extraction d'informations sur les workflows scientifiques à partir de la littérature. volume Extraction et Gestion des Connaissances, RNTI-E-39, p. 313.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). brat : a Web-based Tool for NLP-Assisted Text Annotation. In F. SEGOND, Éd., *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 102–107, Avignon, France : Association for Computational Linguistics.

VERSPOOR K., JIMENO YEPES A., CAVEDON L., MCINTOSH T., HERTEN-CRABB A., THOMAS Z. & PLAZZER J.-P. (2013). Annotating the biomedical literature for the human variome. *Database*, **2013**, bat019. DOI : 10.1093/database/bat019.

WAJSBÜRT P. (2021). *Extraction et normalisation d'entités simples et structurées dans les documents médicaux*. These de doctorat, Sorbonne université.

ZHANG X. & JONASSEN I. (2020). RASflow : an RNA-Seq analysis workflow with Snakemake. *BMC bioinformatics*, **21**(1), 110. DOI : 10.1186/s12859-020-3433-x.

A Exemple d'annotations

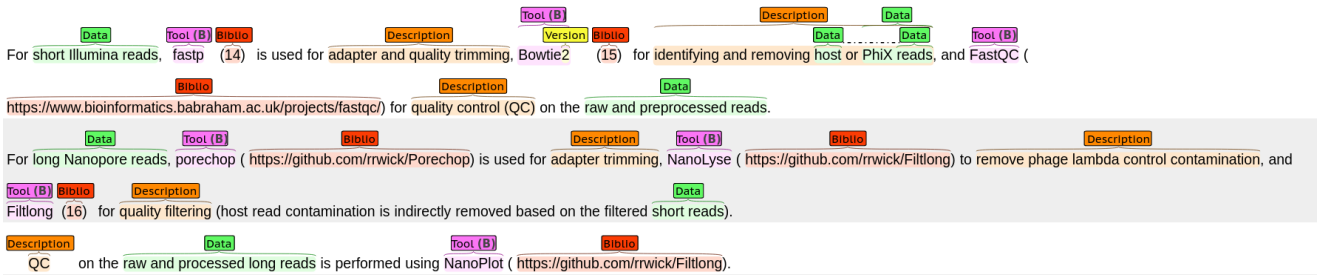


FIGURE 3 – Exemple d'annotation provenant de l'article PMID35118380.

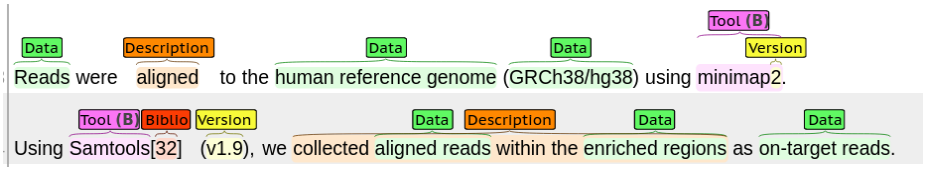


FIGURE 4 – Exemple d'annotation provenant de l'article PMID34103501.

B Hyperparamètres utilisés

Paramètre	Valeur
Encodeur de base	BERT-base-uncased SciBERT_scivocab_uncased
Longueur des séquences	256
Nombres d'itérations max.	5000
Optimiseur	AdamW
Taux d'apprentissage	1e-3
Dropout	0,1
Warmup ratio	0,1
Graines aléatoires	1 - 8 - 22 - 42

TABLE 8 – Hyperparamètres utilisés pour Nlstruct.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Génération contrôlée de cas cliniques en français à partir de données médicales structurées

Hugo Boulanger*[1], Nicolas Hiebel*[2], Olivier Ferret[1], Karën Fort[3], Aurélie Névéol[2]
(1) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
(2) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
(3) Sorbonne Université / Université de Lorraine, CNRS, Inria, LORIA, France
[1]prenom.nom@cea.fr, [2]prenom.nom@lisn.upsaclay.fr, [3]karen.fort@loria.fr

RÉSUMÉ

La génération de texte ouvre des perspectives pour pallier l'absence de corpus librement partageables dans des domaines contraints par la confidentialité, comme le domaine médical. Dans cette étude, nous comparons les performances de modèles encodeurs-décodeurs et décodeurs seuls pour la génération conditionnée de cas cliniques en français. Nous affinons plusieurs modèles pré-entraînés pour chaque architecture sur des cas cliniques en français conditionnés par les informations démographiques des patient·es (sexe et âge) et des éléments cliniques. Nous observons que les modèles encodeurs-décodeurs sont plus facilement contrôlables que les modèles décodeurs seuls, mais plus coûteux à entraîner.

ABSTRACT

Using structured health information for controlled generation of clinical cases in French.

Text generation opens up new prospects for overcoming the lack of open corpora in fields such as healthcare, where data sharing is bound by confidentiality. In this study, we compare the performance of encoder-decoder and decoder-only language models for the conditioned generation of clinical cases in French. We fine-tune several pre-trained models for each architecture on French clinical cases conditioned by patient demographic information (gender and age) and clinical features. We observe that encoder-decoder models are easier to control than decoder-only models, but more costly to train.

MOTS-CLÉS : Génération contrôlée, textes cliniques, textes synthétiques, français.

KEYWORDS: Controlled Generation, Clinical Texts, Synthetic Texts, French.

1 Introduction

L'explosion actuelle de l'intelligence artificielle générative (Cusumano, 2023) ne doit pas faire oublier que les modèles de langue génératifs textuels ont pour compétence première de générer du texte. Les modèles actuels ont permis de pousser cette compétence à un niveau tel qu'il devient difficile de distinguer un texte produit par un humain d'un texte produit par une machine (Casal & Kessler, 2023), ouvrant ainsi la voie à de multiples applications. Dans cet article, nous considérons le cas de documents de référence ne pouvant pas être diffusés, en particulier du fait des informations à caractère personnel qu'ils contiennent, mais suffisamment génériques pour mutualiser des moyens de traitement

*. Les deux auteurs ont contribué de manière égale à ce travail. L'ordre est alphabétique.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 435–448.

à l'échelle d'une communauté. Une façon de développer de tels traitements est de travailler à partir de documents comparables dans leur nature aux documents de référence mais générés automatiquement à partir de ces derniers. Le cas des documents constitutifs des dossiers électroniques patient est à cet égard emblématique, même s'il est loin d'être unique. C'est celui que nous considérons ici.

Dans cette optique, la capacité à contrôler finement le processus de génération est central et multidimensionnel. Pour ne retenir que les principales de ces dimensions, les documents générés doivent être comparables aux documents de référence en termes de style, de structuration, de contenu tout en préservant les informations personnelles qu'ils recèlent. Si les informations directement identifiantes peuvent faire l'objet d'une désidentification robuste en amont, celle-ci ne rend pas les documents anonymes au sens du règlement général sur la protection des données (RGPD). En effet, la désidentification, qu'elle soit automatique ou manuelle, ne protège pas des possibilités de recoupements d'informations médicales, en particulier pour les pathologies rares. Si la possibilité de contrôler la génération en termes de contenu est importante du point de vue de la cohérence médicale des textes générés, elle l'est donc également sur le plan de la préservation des informations personnelles.

Dans cet article, nous proposons ainsi une méthodologie permettant d'exercer un contrôle sur la génération de texte en termes de contenu. Plus précisément, l'idée est de pouvoir conditionner la génération de comptes rendus médicaux par des profils patients. À l'instar de travaux réalisés sur la génération de profils patients synthétiques en termes de données structurées (Walonoski *et al.*, 2017), ces profils prennent la forme d'un ensemble de concepts médicaux. Cette approche, qui relève d'une problématique de génération données-vers-texte, présente l'avantage, par rapport à une approche par amorce textuelle (*prompt*), de pouvoir contrôler finement l'information servant au conditionnement. Ce dernier est mis en œuvre par l'entraînement d'un modèle de langue neuronal à l'aide d'un ensemble de couples composés chacun d'un profil patient sous forme de concepts et d'un compte rendu de référence correspondant à ce profil. Dans ce cadre, les contributions de notre article sont les suivantes :

— une méthodologie de contrôle du contenu de la génération de comptes rendus médicaux ;
— une méthode de constitution d'un ensemble d'entraînement pour la réalisation de ce contrôle ;
— deux formes de mise en œuvre de la stratégie de contrôle proposée ;
— une évaluation multidimensionnelle automatique des résultats de cette stratégie.

2 Travaux connexes

2.1 Génération contrainte de texte

Depuis l'avènement des premiers grands modèles de langues tels que ceux de la famille des GPT (Radford *et al.*, 2018), générer du texte ressemblant à une production humaine semble facile et le problème de la génération a évolué pour changer de cible : le but n'est plus de simplement générer du texte vraisemblable, mais de pouvoir contrôler plus finement ce qui est généré. Les textes produits par les modèles génératifs peuvent ne pas être pertinents ou présenter un contenu offensant voire dangereux (Bender *et al.*, 2021). C'est pourquoi de nombreux travaux portent sur le contrôle de la génération. Le contrôle peut concerner plusieurs aspects de la génération comme le lexique ou le style du texte (Zhang *et al.*, 2023). Plusieurs méthodes de contrôle ont été explorées, dont l'entraînement d'un modèle avec des exemples conditionnés selon des critères choisis (Keskar *et al.*, 2019) ou la modification des probabilités des tokens de sortie lors de l'inférence (Kruszewski *et al.*, 2023).

Les approches « données vers texte » (*data-to-text*) (Lin *et al.*, 2023) contraignent la génération à partir de données structurées (graphes, tableaux et, dans notre cas, les *slots*). Les architectures privilégiées sont des modèles encodeurs-décodeurs pouvant avoir des architectures internes variées, combinant des modèles pré-entraînés en encodeurs et/ou décodeurs. Il est aussi possible d'affiner directement des modèles encodeurs-décodeurs, tels que le modèle T5 (Raffel *et al.*, 2020). Les modèles de langue causaux, comme par exemple les modèles utilisant une architecture de décodeur de transformeur (Vaswani *et al.*, 2017), utilisent le contexte en début de séquence pour générer la suite de la séquence.

2.2 Génération dans le domaine biomédical

Dans le domaine biomédical, la génération de texte est notamment explorée pour produire des comptes-rendus de discussions entre médecins et patients (Eremeev *et al.*, 2023; Ben Abacha *et al.*, 2023; Asada & Miwa, 2023). L'automatisation de cette tâche pourrait en effet grandement soulager une partie de la charge de travail des médecins.

Pour répondre à la difficulté d'accès aux textes médicaux, Ive *et al.* (2020) proposent une méthodologie de génération de cas cliniques synthétiques en anglais à partir de cas cliniques réels. La génération est conditionnée par des entités extraites automatiquement des documents réels. Cependant, peu de corpus, et donc de travaux, portent sur d'autres langues que l'anglais (Névéol *et al.*, 2018). Dans cette étude, nous nous intéressons au cas du français.

3 Méthodologie générale

Comme nous l'avons esquissé dans l'introduction, l'idée directrice de ce travail est de conditionner le processus de génération par les données structurées dont le texte généré devra faire état. Bien entendu, retrouver les données de conditionnement au sein des textes générés ne peut être le seul critère d'évaluation des modèles : il suffirait en effet à ces derniers de reproduire leur entrée pour être jugés comme parfaits. Ce conditionnement doit donc intégrer une proximité de nature par rapport aux documents de référence que l'on souhaite émuler.

Comme évoqué à la section 2.1, ce double conditionnement peut se faire par un affinage a priori du modèle de langue servant à la génération ou bien par son contrôle lors de la génération. Nous avons opté pour la première solution dans la mesure où la seconde suppose d'appliquer des processus d'analyse textuelle élaborés lors de la génération pour vérifier le respect du conditionnement, ce qui est coûteux. La première solution suppose néanmoins de disposer de données d'entraînement associant données de conditionnement et textes exemples conformes à ce conditionnement. Pour ce faire, nous avons adopté une stratégie comparable à Peng *et al.* (2018) pour la génération d'histoire, reprise par Ive *et al.* (2020) pour les comptes rendus médicaux, et consistant à extraire automatiquement les données de conditionnement des textes exemples. Cette stratégie suppose bien évidemment de disposer de processus d'analyse textuelle capables d'extraire ces données de conditionnement des textes exemples avec un niveau de performance suffisamment élevé. Elle induit par conséquent un couplage étroit entre les capacités de génération et celles d'analyse, mais permet de se passer d'une annotation manuelle coûteuse. Dans le cas présent, nous nous focalisons sur les concepts médicaux et sommes donc dépendants de modèles permettant d'extraire ces concepts de comptes rendus médicaux,

mais la généricité de cette stratégie permet de prendre en compte facilement de nouveaux éléments de conditionnement, dès lors qu'ils peuvent être extraits automatiquement de textes exemples.

4 Corpus et modèles génératifs

4.1 Corpus de cas cliniques en français

Les données utilisées pour nos expériences proviennent de deux corpus de cas cliniques librement disponibles. Le premier est le corpus CAS (Grabar *et al.*, 2018), un corpus de cas cliniques désidentifiés en français [1]. Le second est le corpus E3C (Magnini *et al.*, 2020), un corpus multilingue de cas cliniques désidentifiés. Nous nous intéressons uniquement aux cas cliniques en français de ce dernier.

4.2 Construction des contraintes selon un profil patient

Nous souhaitons pouvoir contraindre la génération par des éléments cliniques afin de créer des cas cliniques cohérents. Nous avons échangé avec des cliniciens afin de définir les éléments saillants dans des cas cliniques. Ces éléments sont ensuite sélectionnés comme conditionnement de la génération. Le tableau 1 présente un exemple des différentes catégories d'éléments importants qui ont été retenus pour un cas clinique du corpus E3C. On y retrouve les informations démographiques du patient (âge et sexe), la localisation de la pathologie, des informations histologiques, différents signes ou symptômes, des traitements et procédures effectués, des résultats biologiques et des scores (mesures ou codes). En accord avec les recommandations des cliniciens, nous identifions une vingtaine de contraintes par cas, en sélectionnant si possible des éléments de chaque catégorie avec une majorité de symptômes, traitements et procédures. Cette façon de faire permet de sélectionner les informations importantes des cas cliniques selon les médecins.

Types d'éléments cliniques	**Exemple de valeurs**
Âge	54
Sexe	Masculin
Localisation	Vessie
Histologie	adénocarcinome de l'ouraque peu différencié
Signe	hématurie
Procédure	scanner CT
Traitement	chimiothérapie par Méthotrexate-Vinblastine-Endoxan-Cisplatine
Score	T III A (selon la classification de Sheldon)
Bio	une négativité pour les cytokératines (ck) 7 et 20

TABLE 1 – Exemples d'éléments de contrôle manuellement définis pour un cas clinique. Le cas clinique correspondant est présenté dans le tableau 2.

1. Contacter les auteurs pour accéder au corpus https://deft.lisn.upsaclay.fr/2020

4.3 Extraction des contraintes des documents

Concernant les données démographiques, nous nous sommes appuyés pour le corpus CAS sur les annotations présentes relatives à l'âge et au sexe des patients. Le corpus E3C ne disposant pas de ces informations, nous avons annoté les 1 009 cas cliniques en français du corpus pour obtenir l'âge et le sexe des patients. Pour les autres entités cliniques, nous annotons automatiquement les deux corpus pour que les annotations des deux corpus soient homogènes et facilitent ainsi l'apprentissage des contraintes par les modèles de génération. Pour réaliser cette annotation automatique, nous utilisons des modèles de reconnaissance d'entités cliniques entraînés sur le corpus privé MERLOT (Campillos *et al.*, 2018), qui contient des annotations manuelles pour ces entités.

Nous avons construit nos contraintes en partant des annotations démographiques manuelles et des annotations automatiques en entités cliniques. Pour chaque document, nous sélectionnons l'âge et le sexe lorsqu'ils sont renseignés. Lorsque l'âge exact n'est pas renseigné, nous utilisons les catégories d'âge issues des descripteurs obligatoires du thésaurus MeSH (Medical Subject Headings) [2].

Pour les entités cliniques, nous avons sélectionné les catégories d'annotation de MERLOT pour correspondre aux catégories discutées avec les médecins. Ainsi, nous retenons pour chaque cas clinique les dix procédures (*PROC*) et dix symptômes (*DISO*) ayant le tf.idf le plus élevé. Nous sélectionnons aussi les substances (*CHEM*) et les mesures (*MEAS*). Ces dernières sont filtrées pour ne garder que des mesures informatives (les chiffres seuls comme *6* sont par exemple annotés comme *MEAS* mais sans information supplémentaire). De cette manière, nous obtenons en moyenne 26 contraintes ($\pm$9,5) par cas clinique.

4.4 Modèles génératifs

Nous comparons les performances de deux architectures différentes pour la génération contrainte de textes cliniques : l'architecture encodeur-décodeur et l'architecture décodeur seul. Nous nous appuyons pour cela sur des modèles transformeurs pré-entraînés.

Encodeur-décodeur Cette architecture est spécialisée dans la génération de texte à partir de données structurées. Notamment, l'affinage du modèle T5 (Raffel *et al.*, 2020) s'est imposé comme une méthode standard pour ce genre de tâches. Nous avons choisi d'utiliser la version multilingue de T5, appelée mT5 (Xue *et al.*, 2021), d'un milliard de paramètres comme modèle pré-entraîné, et les modèles Flan-T5-Large (780 millions de paramètres) et Flan-T5-XL (3 milliards de paramètres) (Chung *et al.*, 2022) comme modèles affinés avec instructions.

Décodeur seul Cette architecture est spécialisée dans la génération de texte à partir d'amorces textuelles (*prompt*). Nous avons choisi plusieurs modèles pour cette architecture : Bloom (Scao *et al.*, 2022), un modèle génératif entraîné sur plusieurs langues, et Bloomz, une variante entraînée spécialement pour réaliser différentes tâches (traduction, résumé automatique etc.). Nous prenons pour chacun de ces deux modèles deux versions en termes de taille : un et sept milliards de paramètres.

2. https://www.nlm.nih.gov/bsd/indexing/training/CHK_030.html

5 Expérimentations

5.1 Représentation des données structurées

L'utilisation de ces modèles génératifs nécessite de transformer les données structurées en format textuel. Nous avons choisi de linéariser les entrées de manière différente pour les modèles encodeurs-décodeurs et les modèles décodeurs seuls. Pour les modèles encodeurs-décodeurs, nous ajoutons devant chaque entité un token spécial lié à la classe de l'entité. Nous séparons les informations démographiques (âge, sexe) des contraintes médicales (symptôme, procédure etc.) par un token spécial *contraintes*. Pour les modèles décodeurs seuls, nous avons choisi de ne pas ajouter de tokens spéciaux. La figure 1 présente un exemple de représentation des données pour les encodeurs-décodeurs.

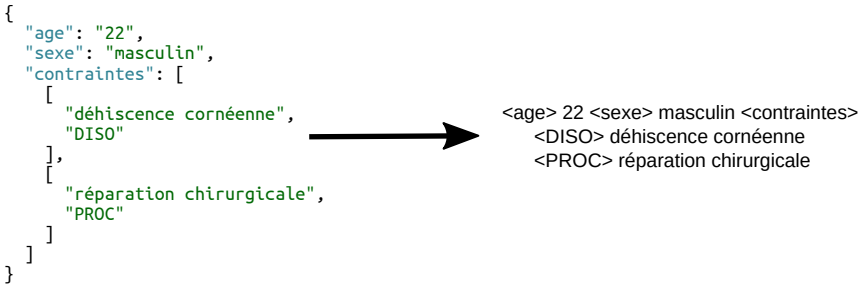


FIGURE 1 – Exemple de représentation des données pour l'architecture encodeur-décodeur.

5.2 Affinage

Le jeu d'entraînement sur lequel nous affinons nos modèles est constitué de 1 424 cas cliniques, pour un peu plus de 500 000 tokens hors contraintes. Pour l'affinage des modèles, nous choisissons de geler les poids des modèles pré-entraînés et ajoutons des matrices entraînables LoRA (Hu *et al.*, 2022). L'emplacement des matrices entraînables dépend du type de modèle. Pour les modèles encodeurs-décodeurs, nous ajoutons des matrices LoRA sur les *queries* et *values* des couches transformeurs et sur la tête de modélisation du modèle. Pour les modèles décodeurs seuls, les matrices LoRA sont ajoutées sur les couches linéaires des modèles. Nous ajoutons les tokens spéciaux aux plongements via des vecteurs initialisés aléatoirement. Le traitement des plongements lexicaux varie selon trois configurations définies comme suit :

Configuration « gelé » : les plongements sont gelés mais nous ajoutons des matrices LoRA pour leur permettre une adaptation à la tâche à un faible coût mémoire.

Configuration « dégelé » : les plongements sont dégelés, pour permettre l'adaptation à la tâche, mais à un coût plus élevé.

Configuration « partiel » : seuls les plongements des tokens spéciaux sont dégelés.

5.3 Génération des cas cliniques

Notre jeu de test est constitué de 156 cas cliniques et de leurs contraintes. Les contraintes sont données en entrée aux modèles génératifs et les cas cliniques réels servent de référence. Le décodage est

réalisé en utilisant une recherche par faisceaux (*beams*) avec 5 faisceaux, et de l'échantillonnage (*sampling*) avec un top-p de 0,90, un top-k de 100, une température de 1 et une pénalité de répétition de 3. Faire de l'échantillonnage lors de la génération signifie qu'un même modèle peut générer des textes différents avec la même entrée. Nous effectuons cinq générations pour chaque exemple de test afin de prendre en compte cette variabilité.

5.4 Métriques d'évaluation

L'évaluation automatique de la génération de texte est notoirement difficile (Novikova *et al.*, 2017). De nombreuses métriques existent, qui permettent de mesurer différents aspects de la génération de texte (Frisoni *et al.*, 2022). Nous avons sélectionné certaines d'entre elles afin de couvrir plusieurs dimensions de l'évaluation.

Adéquation aux contraintes - *Exactitude* Cette mesure sert à évaluer la capacité du modèle à respecter les contraintes qui lui sont imposées. Nous calculons la proportion de contraintes respectées dans les textes générés par rapport au nombre total de contraintes imposées.

Qualité de la langue - *Perplexité* La perplexité évalue l'adéquation des données textuelles avec la distribution de probabilité d'un modèle de langue. Nous utilisons un modèle spécifique au français, GPTFR (Simoulin & Crabbé, 2021). Pour cette métrique, nous souhaitons que la perplexité obtenue sur les données générées se rapproche de la perplexité obtenue sur les données réelles (égale à 19 ici pour le corpus d'entraînement).

Diversité des textes générés - *Self-BLEU* Le score Self-BLEU (Zhu *et al.*, 2018) est la moyenne des scores BLEU de toutes les phrases d'un corpus entre elles. Ainsi, un corpus redondant aura un score Self-BLEU élevé tandis qu'un corpus varié aura un score plus faible.

Proximité avec le corpus naturel - *Corpus-BLEU* Corpus-BLEU (Yu *et al.*, 2017) est une mesure de proximité entre deux corpus et correspond à la moyenne des scores BLEU entre chaque phrase du corpus généré et toutes les phrases du corpus naturel. Nous calculons Corpus-BLEU en comparant les cas cliniques du corpus de test avec les textes générés.

Proximité avec le cas clinique correspondant aux contraintes - *BLEU* Le score BLEU (Papineni *et al.*, 2002) est calculé entre le texte généré et le cas clinique réel d'où proviennent les contraintes. Il mesure la proximité avec les données réelles de façon plus spécifique que le score Corpus-BLEU.

6 Résultats

6.1 Génération des cas cliniques

Le tableau 2 présente des exemples de textes générés à partir d'un ensemble de contraintes par un modèle encodeur-décodeur (Flan-T5-XL gelé) et un modèle décodeur seul (Bloomz 1b1 dégelé). Le tableau 3 présente quant à lui l'évaluation automatique des cas cliniques générés avec les différentes architectures étudiées. Parmi nos baselines, la simple copie des entités de conditionnement (« Copie ») obtient comme attendu une exactitude de 100 %, mais aussi une perplexité très grande. La baseline « Corpus » correspond à une copie du corpus de test dans laquelle nous avons enlevé les retours à la ligne. Cette modification explique pourquoi les scores BLEU et corpus-BLEU ne sont pas

Contraintes extraites automatiquement (hors balises)	âge : 54 ; sexe : masculin ; contraintes : hématurie isolée, examen tomodensitométrique, masse, 4 cm, adénocarcinome peu différencié, de type III, bilan d' extension, cystoprostatectomie radicale totale, lymphadénectomie iliaque, obturatrice, omphalectomie, entérocystoplastie de substitution, adénocarcinome de l'ouraque peu différencié, très localement mucosécrétant, ulcéré, carcinome transitionnel, grade III, Antigène Carcino-Embryonnaire, Leu-M1, CD 15, cytokératines, épithélium vésical, classification de Sheldon, Méthotrexate, Vinblastine, Endoxan, Cisplatine
Cas clinique réel	Un homme de 54 ans a consulté pour hématurie isolée. Une échographie, puis un examen tomodensitométrique, démontraient une masse de 4 cm de diamètre, au centre nécrotique, antérieure au dôme vésical, envahissant uniquement la graisse adjacente (Figure 1A.). Les biopsies, faites à la cystoscopie, révèlaient un adénocarcinome peu différencié, avec des zones d'épithélioma transitionnel de type III (Figure 1B). Le bilan d' extension était négatif. Le patient bénéficiait d'une cystoprostatectomie radicale totale avec lymphadénectomie iliaque et obturatrice, associée à une omphalectomie et une entérocystoplastie de substitution selon Studer. L'examen anatomo-pathologique complet de la pièce opératoire, révèlait un adénocarcinome de l'ouraque peu différencié, invasif, très localement mucosécrétant (Figure 1C). Il était ulcéré dans la lumière vésicale, et contenait des zones de carcinome transitionnel de grade III, ainsi que des cellules argentaffines. Deux poches de lymphonoeuds et la graisse environnante, étaient envahis. L'immunohistochimie montrait un positivité des cellules néoplasiques pour l'Antigène Carcino-Embryonnaire (ACE), Leu-M1 (CD 15), et une négativité pour les cytokératines (ck) 7 et 20. L'épithélium vésical était positif pour les ck 7. Il s'agissait d'un stade T III A (selon la classification de Sheldon). Le patient subissait une chimiothérapie par Méthotrexate - Vinblastine - Endoxan -Cisplatine par voie systémique. Avec 35 mois de recul, il va bien.
Génération Bloomz 1b1 dégelé	Il s'agit d'un patient âgé de 54 ans sans antécédents pathologiques particuliers qui a consulté pour une hématurie isolée. L'examen clinique trouvait un patient en bon état général, apyrétique, avec des signes d'hypertension artérielle et une sensibilité abdominale à la palpation. La biologie était normale. L'échographie abdominale objectivait une masse tissulaire hétérogène du flanc gauche mesurant 4 cm de grand axe. L'IRM montrait une masse tissulaire homogène au niveau du flanc gauche (Figure 1). L'étude histologique de la pièce opératoire confirmait le diagnostic d'adénocarcinome peu différencié de type III selon la classification de Sheldon [1]. Le bilan d'extension par résonance magnétique (RMN) ne retrouvait pas de métastase cérébrale ou pulmonaire. Une cystoprostatectomie radicale totale associée à une lymphadénectomie iliaque et obturatrice ainsi qu'une omphalectomie et entérocystoplastie de substitution ont été réalisées. L'étude anatomopathologique de la pièce opératoire avait confirmé le diagnostic d'adénocarcinome de l'ouraque peu différencié évoluant vers un carcinome transitionnel stade III selon la classification de Sheldon [1].
Génération Flan-T5-XL gelé	Il s'agit d'un patient âgé de 54 ans, sans antécédents pathologiques particuliers, qui a présenté une hématurie isolée évoluant depuis 2 mois. L'examen tomodensitométrique a objectivé une masse de 4 cm au niveau de la paroi postérieure de l'ouraque (Figure 1). Le diagnostic d'adénocarcinome peu différencié de type III a été retenu. Le bilan d' extension était négatif. Une cystoprostatectomie radicale totale associée à une lymphadénectomie iliaque non obturatrice et une omphalectomie avec entérocystoplastie de substitution a été réalisée. L'étude anatomopathologique de la pièce opératoire a confirmé le diagnostic d'adénocarcinome de l'ouraque peu différencié comportant un aspect très localement mucosécrétant et ulcéré. Cet aspect est compatible avec un carcinome transitionnel de grade III (Antigène Carcino-Embryonnaire : CD34, Leu-M1 et CD 15). Les cellules tumorales exprimaient les cytokératines spécifiques de l'épithélium vésical (classification de Sheldon). Un traitement associant Méthotrexate, Vinblastine, Endoxan et Cisplatine a été débuté.

TABLE 2 – Exemples de textes générés par deux modèles à partir de contraintes automatiquement extraites d'un cas clinique réel.

parfaits et, de façon plus surprenante, fait baisser la perplexité de 30,5 à 19,5. Le score d'exactitude révèle quant à lui des limites de nos données et du calcul d'exactitude. La majorité de ces erreurs concerne le sexe du patient, lorsqu'il n'est pas marqué via l'accord du terme « patient » en genre ou l'utilisation du qualificatif « masculin » ou « féminin ». Les autres erreurs proviennent majoritairement de reformulations ou d'erreurs dans les contraintes.

Les résultats des modèles montrent plusieurs tendances. Les modèles encodeurs-décodeurs ayant bénéficié d'une période d'affinage avec instructions, les modèles Flan, obtiennent globalement de meilleurs résultats que les modèles pré-entraînés sans instructions. Les modèles Flan ont en outre

	Modèle	Exactitude↑	Perplexité	Self-BLEU-4↓	Corpus-BLEU-4↑	BLEU-4↑
Baselines	Copie	100	194,3	14,4	25,5	1,1
	Corpus	98,8	19,5	33,4	97,4	97,5
	Bloom 1b1 gelé∗	s/o	11,5±1,5	86,1±0,4	64,8±0,4	s/o
	Bloom 1b1 dégelé∗	s/o	10,2±0,9	82,9±0,4	60,6±0,5	s/o
	Bloom 7b1 gelé∗	s/o	8,4±2,8	79,3±1,2	57,1±0,5	s/o
Encodeurs-décodeurs	mT5-large gelé	78,0±0,6	13,6±0,2	53,5±0,5	55,8±0,5	12,0±0,1
	mT5-large dégelé	73,6±0,8	13,4±0,2	53,8±0,4	56,4±0,3	10,9±0,2
	mT5-large partiel	75.7±0.3	13,2±0,4	55.1±0.5	56.5±0.2	11.6±0.3
	Flan-T5-large gelé	81,5±1,1	14,8±0,4	52,8±0,4	55,3±0,4	12,0±0,1
	Flan-T5-large dégelé	80,3±1,0	15,6±0,5	51,9±0,2	55,0±0,4	11,7±0,2
	Flan-T5-large partiel	80,9±0,9	16,1±0,2	50,9±0,3	54,3±0,4	11,6±0,1
	Flan-T5-XL gelé	84,2±0,8	14,9±0,2	50,2±0,2	54,5±0,2	12,8±0,1
	Flan-T5-XL dégelé	85,3±0,8	14,9±0,2	49,0±0,1	53,8±0,4	12,9±0,2
	Flan-T5-XL partiel	82,2±1,3	15,4±0,2	50,3±0,2	54,6±0,3	12,0±0,2
Décodeurs	Bloom 1b1 gelé	40,5±3,9	8,8±0,2	62,5±5,8	42,3±11,1	4,7±1,0
	Bloom 1b1 dégelé	29,6±0,9	9,3±0,4	63,6±4,7	50,4±9,7	4,0±0,5
	Bloom 7b1 gelé	43,5±2,5	9,9±0,6	54,0±2,1	47,5±2,0	5,8±1,0
	Bloomz 1b1 gelé	45,4±4,2	9,2±0,2	61,9±7,6	41,8±11,0	5,2±1,3
	Bloomz 1b1 dégelé	32,1±1,7	9,6±0,2	65,7±6,0	47,0±13,2	4,3±0,7
	Bloomz 7b1 gelé	39,8±3,0	9,9±0,2	55,0±1,9	49,8±1,5	5,4±0,4

TABLE 3 – Évaluation des données générées à partir des contraintes provenant du jeu de test. Modèles baselines marqués par « ∗ » : entraînement et génération sans contrainte.

l'avantage d'être plus rapidement affinés à taille égale, avec une période d'entraînement de 16 h pour Flan-T5-large contre 60 h pour mT5-large. Nous observons, comme attendu, que les modèles Flan-T5-XL sont les plus performants des modèles encodeurs-décodeurs testés. Ces derniers génèrent des textes plus variés (Self-BLEU), et présentent la meilleure exactitude. Les textes générés ressemblent le plus aux références (BLEU) et la perplexité et le Corpus-BLEU sont meilleurs que ceux de la version plus petite du modèle. Le Corpus-BLEU reste cependant assez stable quel que soit le modèle initial et le traitement des plongements lexicaux. Il est cependant à noter que les modèles mT5 obtiennent une perplexité inférieure, probablement due au fait que le modèle initial soit multilingue tandis que les modèles Flan-T5 n'ont vu de français que sur des tâches de traduction.

Nous observons que les décodeurs seuls obtiennent de moins bons résultats que les encodeurs-décodeurs. Les modèles décodeurs sont également nettement plus instables d'une génération à une autre, avec des écarts-types importants au niveau de l'exactitude, du Self-BLEU et du Corpus-BLEU. Au niveau de la perplexité, ces modèles obtiennent des scores plus faibles et s'éloignent donc du corpus d'entraînement. Le modèle permettant de calculer la perplexité étant un décodeur seul, l'architecture commune biaise potentiellement les décodeurs pour cette métrique. En revanche, le temps d'entraînement des décodeurs est beaucoup plus court : 10 à 15 minutes pour les modèles à un milliard de paramètres et 30 minutes pour les modèles à sept milliards de paramètres.

Nous pouvons également identifier quelques bonnes pratiques concernant le pré-entraînement des modèles et la configuration des plongements lexicaux. Les modèles ayant bénéficié d'un pré-entraînement avec instructions sont globalement plus performants que les modèles avec un pré-entraînement simple

sur une tâche de modélisation de la langue. Cela s'observe principalement pour l'exactitude et le score BLEU. Nous pouvons aussi observer que les modèles dégelés ont de moins bonnes performances que les modèles gelés. Cependant, nous avons remarqué que les modèles dégelés convergent plus rapidement, en temps et en époques d'entraînement. Les résultats plus faibles de Flan-T5-large gelé sont peut être le résultat d'un sous-entraînement.

6.2 Impact environnemental

Modèle	Entraînement	Génération	Perplexité	Total
mT5-large	7,26	0,75	0,01	8,02
flan-T5-large	1,95	0,75	0,01	2,71
flan-T5-XL	7,26	0,75	0,01	8,02
Bloom(z) 1b1	0,03	0,78	0,01	0,82
Bloom(z) 7b1	0,05	0,64	0,01	0,70

TABLE 4 – Impact environnemental en $kgCO_2$éq des expériences finales pour chaque modèle. Chaque ligne somme les émissions des différentes configurations associées. Les émissions totales sont de 20,27 $kgCO_2$éq.

La compilation des émissions en $kgCO_2$éq peut être retrouvée dans le tableau 4. L'impact environnemental est essentiellement lié à l'entraînement des modèles encodeurs-décodeurs, qui est plus long et requiert plus de GPU pour les modèles plus grands. Ces évaluations ont été réalisées avec le MachineLearning Impact calculator présenté dans (Lacoste *et al.*, 2019) avec les valeurs d'émission de la France (0,101 $kgCO_2$éq/kWh) se trouvant dans (Moro & Lonza, 2018).

6.3 Limites

L'ensemble de mesures que nous avons mis en place nous permet d'avoir une vision assez bonne sur ce que nos modèles génèrent. Il y a néanmoins des limites à n'utiliser que l'exactitude, en particulier telle qu'elle est calculée, pour décrire la fidélité de la retranscription des informations. L'exactitude recherche ici une correspondance exacte entre les contraintes et le texte. Toute reformulation du modèle est donc écartée bien qu'elle puisse être correcte. De plus, utiliser cette mesure seule ne nous donne pas d'information sur de potentiels ajouts d'informations ou d'entités par les modèles. Dans cette étude, nous avons exclusivement utilisé des métriques automatiques pour l'évaluation des textes générés. Il est difficile d'évaluer manuellement la qualité des textes générés sans connaissances cliniques. Une évaluation manuelle par des experts cliniques permettrait d'estimer la cohérence médicale des textes générés de manière plus fiable. Nous avons enfin constaté que les générations par un même modèle peuvent être instables. Un filtrage des textes pour garder le meilleur candidat pourrait améliorer les résultats (Hiebel *et al.*, 2023).

7 Conclusion

Dans cette étude, nous générons des comptes rendus médicaux en français conditionnés par des données cliniques structurées. Nous comparons des modèles d'architectures différentes, des encodeurs-décodeurs et des décodeurs seuls, que nous affinons sur un corpus de cas cliniques à l'aide de matrices LoRA. Nous proposons une méthodologie d'évaluation fondée sur un ensemble de mesures automatiques : exactitude, perplexité, Self-BLEU, Corpus-BLEU et BLEU. Nous observons que les modèles à architecture encodeurs-décodeurs obtiennent de meilleurs résultats sur la tâche de génération à partir de données structurées, mais avec un entraînement plus coûteux. Concernant les différentes stratégies d'affinage au niveau des plongements lexicaux, la meilleure stratégie consiste à ajouter des matrices LoRA sur les plongements lexicaux et non de les dégeler, bien que cela allonge l'apprentissage. La puissance de calcul disponible dans un cadre hospitalier limite la possibilité d'utiliser des modèles plus gros et/ou plus lourds. D'après nos résultats, les architectures décodeurs sont plus légères et donc plus adaptées. Il faudrait cependant générer plusieurs candidats et les filtrer pour compenser l'irrégularité de ces modèles. Il serait aussi intéressant d'explorer les performances de modèles à architecture encodeur-décodeur plus petits que ceux testés dans cette étude (le modèle Flan-T5-small ne contient par exemple que 80 millions de paramètres). La quantification (*quantization*) des modèles pourrait aussi être une solution pour réduire la charge de calcul, à condition que les modèles quantifiés donnent des résultats comparables à leurs homologues standards.

Remerciements

Ce travail a été réalisé dans le cadre d'un projet de l'Agence Nationale de la Recherche, CODEINE (artificial text COrpus DEsIgNed Ethically), ANR-20-CE23-0026-01. Il a été réalisé grâce au super-calculateur Factory-IA financé par le Conseil Régional d'Île-de-France.

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2023-AD011014538 attribuée par GENCI.

Références

ASADA M. & MIWA M. (2023). BioNART : A biomedical non-AutoRegressive transformer for natural language generation. In D. DEMNER-FUSHMAN, S. ANANIADOU & K. COHEN, Éds., *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, p. 369–376, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.bionlp-1.34.

BEN ABACHA A., YIM W.-W., FAN Y. & LIN T. (2023). An empirical study of clinical note generation from doctor-patient encounters. In A. VLACHOS & I. AUGENSTEIN, Éds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 2291–2302, Dubrovnik, Croatie : Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.168.

BENDER E. M., GEBRU T., MCMILLAN-MAJOR A. & SHMITCHELL S. (2021). On the dangers of stochastic parrots : Can language models be too big ? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, p. 610–623, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3442188.3445922.

CAMPILLOS L., DELÉGER L., GROUIN C., HAMON T., LIGOZAT A.-L. & NÉVÉOL A. (2018). A French clinical corpus with comprehensive semantic annotations : development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). *Language Resources and Evaluation*, **52**(2), 571–601. DOI : 10.1007/s10579-017-9382-y.

CASAL J. E. & KESSLER M. (2023). Can linguists distinguish between chatgpt/ai and human writing ? : A study of research ethics and academic publishing. *Research Methods in Applied Linguistics*, **2**(3), 100068. DOI : https ://doi.org/10.1016/j.rmal.2023.100068.

CHUNG H. W., HOU L., LONGPRE S., ZOPH B., TAY Y., FEDUS W., LI Y., WANG X., DEHGHANI M., BRAHMA S. *et al.* (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv :2210.11416*.

CUSUMANO M. A. (2023). Generative ai as a new innovation platform. *Communications of the ACM*, **66**(10), 18—-21. DOI : 10.1145/3615859.

EREMEEV M., VALMIANSKI I., AMATRIAIN X. & KANNAN A. (2023). Injecting knowledge into language generation : a case study in auto-charting after-visit care instructions from medical dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2373–2390, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.133.

FRISONI G., CARBONARO A., MORO G., ZAMMARCHI A. & AVAGNANO M. (2022). NLG-metricverse : An end-to-end library for evaluating natural language generation. In N. CALZOLARI, C.-R. HUANG, H. KIM, J. PUSTEJOVSKY, L. WANNER, K.-S. CHOI, P.-M. RYU, H.-H. CHEN, L. DONATELLI, H. JI, S. KUROHASHI, P. PAGGIO, N. XUE, S. KIM, Y. HAHM, Z. HE, T. K. LEE, E. SANTUS, F. BOND & S.-H. NA, Éds., *Proceedings of the 29th International Conference on Computational Linguistics*, p. 3465–3479, Gyeongju, Corée du Sud : International Committee on Computational Linguistics.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, p. 122–128, Bruxelles, Belgique : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5614.

HIEBEL N., FERRET O., FORT K. & NÉVÉOL A. (2023). Can synthetic text help clinical named entity recognition ? a study of electronic health records in French. In A. VLACHOS & I. AUGENSTEIN, Éds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 2320–2338, Dubrovnik, Croatie : Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.170.

HU E. J., YELONG SHEN, WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2022). LoRA : Low-rank adaptation of large language models. In *International Conference on Learning Representations*, En ligne.

IVE J., VIANI N., KAM J., YIN L., VERMA S., PUNTIS S., CARDINAL R., ROBERTS A., STEWART R. & VELUPILLAI S. (2020). Generation and evaluation of artificial mental health records for natural language processing. *npj Digital Medicine*, **3**. DOI : 10.1038/s41746-020-0267-x.

KESKAR N. S., MCCANN B., VARSHNEY L. R., XIONG C. & SOCHER R. (2019). CTRL : A conditional transformer language model for controllable generation. *CoRR*, **abs/1909.05858**.

KRUSZEWSKI G., ROZEN J. & DYMETMAN M. (2023). disco : a toolkit for distributional control of generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3 : System Demonstrations)*, p. 144–160, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-demo.14.

LACOSTE A., LUCCIONI A., SCHMIDT V. & DANDRES T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv :1910.09700*.

LIN Y., RUAN T., LIU J. & WANG H. (2023). A survey on neural data-to-text generation. *IEEE Transactions on Knowledge and Data Engineering*, p. 1–20. DOI : 10.1109/TKDE.2023.3304385.

MAGNINI B., ALTUNA B., LAVELLI A., SPERANZA M. & ZANOLI R. (2020). The E3C Project : Collection and Annotation of a Multilingual Corpus of Clinical Cases. In J. MONTI, F. DELL'ORLETTA & F. TAMBURINI, Éds., *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020*, volume 2769 de *CEUR Workshop Proceedings*, Bologne, Italie : CEUR-WS.org.

MORO A. & LONZA L. (2018). Electricity carbon intensity in european member states : Impacts on ghg emissions of electric vehicles. *Transportation Research Part D : Transport and Environment*, **64**, 5–14. The contribution of electric vehicles to environmental challenges in transport. WCTRS conference in summer, DOI : https ://doi.org/10.1016/j.trd.2017.07.012.

NÉVÉOL A., DALIANIS H., VELUPILLAI S., SAVOVA G. & ZWEIGENBAUM P. (2018). Clinical natural language processing in languages other than english : opportunities and challenges. *Journal of Biomedical Semantics*, **9**(1), 12. DOI : 10.1186/s13326-018-0179-8.

NOVIKOVA J., DUŠEK O., CERCAS CURRY A. & RIESER V. (2017). Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 2241–2252, Copenhague, Danemark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1238.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In P. ISABELLE, E. CHARNIAK & D. LIN, Éds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphie, Pennsylvanie, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

PENG N., GHAZVININEJAD M., MAY J. & KNIGHT K. (2018). Towards controllable story generation. In M. MITCHELL, T.-H. K. HUANG, F. FERRARO & I. MISRA, Éds., *Proceedings of the First Workshop on Storytelling*, p. 43–49, Nouvelle-Orléans, Louisiane : Association for Computational Linguistics. DOI : 10.18653/v1/W18-1505.

RADFORD A., NARASIMHAN K., SALIMANS T. & SUTSKEVER I. (2018). Improving language understanding by generative pre-training.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, **21**(1).

SCAO T. L. *et al.* (2022). Bloom : A 176b-parameter open-access multilingual language model. DOI : 10.48550/ARXIV.2211.05100.

SIMOULIN A. & CRABBÉ B. (2021). Un modèle Transformer Génératif Pré-entrainé pour le _____ français. In P. DENIS, N. GRABAR, A. FRAISSE, R. CARDON, B. JACQUEMIN, E. KERGOSIEN & A. BALVET, Éds., *Traitement Automatique des Langues Naturelles*, p. 246–255, Lille, France : ATALA. HAL : hal-03265900.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. VON LUXBURG, S. BENGIO, H. M. WALLACH, R. FERGUS, S. V. N. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, p. 5998–6008.

WALONOSKI J., KRAMER M., NICHOLS J., QUINA A., MOESEL C., HALL D., DUFFETT C., DUBE K., GALLAGHER T. & MCLACHLAN S. (2017). Synthea : An approach, method, and

software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, **25**(3), 230–238. DOI : 10.1093/jamia/ocx079.

XUE L., CONSTANT N., ROBERTS A., KALE M., AL-RFOU R., SIDDHANT A., BARUA A. & RAFFEL C. (2021). mT5 : A massively multilingual pre-trained text-to-text transformer. In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 483–498, En ligne : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.41.

YU L., ZHANG W., WANG J. & YU Y. (2017). Seqgan : sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, p. 2852–2858 : AAAI Press.

ZHANG H., SONG H., LI S., ZHOU M. & SONG D. (2023). A survey of controllable text generation using transformer-based pre-trained language models. volume 56, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3617680.

ZHU Y., LU S., ZHENG L., GUO J., ZHANG W., WANG J. & YU Y. (2018). Texygen : A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, p. 1097–1100, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3209978.3210080.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

L'impact de genre sur la prédiction de la lisibilité du texte en FLE

GAO Lingyun[1] Rodrigo Wilkens[1] Thomas François[1]
(1) CENTAL, IL&C, Université catholique de Louvain, Belgique
lingyun.gao.nom@uclouvain.be

RÉSUMÉ

Cet article étudie l'impact du genre discursif sur la prédiction de la lisibilité des textes en français langue étrangère (FLE) à travers l'intégration de méta-informations du genre discursif dans les modèles de prédiction de la lisibilité. En utilisant des architectures neuronales basées sur CamemBERT, nous avons comparé les performances de modèles intégrant l'information de genre à celles d'un modèle de base ne considérant que le texte. Nos résultats révèlent une amélioration modeste de l'exactitude (*accuracy*) globale lors de l'intégration du genre, avec cependant des variations notables selon les genres spécifiques de textes. Cette observation semble confirmer l'importance de prendre en compte les méta-informations textuelles tel que le genre lors de la conception de modèles de lisibilité et de traiter le genre comme une information riche à laquelle le modèle doit accorder une position préférentielle.

ABSTRACT

The impact of genre on the prediction of text readability in FFL (French as a Foreign Language)

This article examines the influence of discourse genre on readability prediction in texts for French as a foreign language (FLE), focusing on the integration of genre-related meta-information into readability models. Utilizing neural architectures based on CamemBERT, we assessed the performance of models that incorporate genre information against a baseline model that considers only the text. Our findings indicate a modest enhancement in overall accuracy with the inclusion of genre, though little significant variations were observed across specific text genres. These results seem to confirm the importance of taking into account textual meta-information such as genre when designing readability models, and of treating genre as rich information to which the model should give a preferential position.

MOTS-CLÉS : lisibilité, français langue étrangère, genre discursif.

KEYWORDS: readability, French as a foreign language, genre.

1 Introduction

Dans le contexte de la didactique des langues étrangères, la sélection de matériaux pédagogiques adaptés aux étudiants constitue un problème fondamental. Dès lors, certains chercheurs ont entrepris de développer des algorithmes capables de déterminer automatiquement le niveau de compétence nécessaire pour comprendre un texte, en vue de faciliter les tâches de préparation pour les formateurs et les examinateurs. Ces modèles, appelés formules de lisibilité, prennent en compte différentes caractéristiques des textes pour prédire automatiquement le degré de lisibilité d'un texte pour une population ciblée (François, 2011). La notion de lisibilité a été définie par Dale & Chall (1949, 19)

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 449–471.

comme : la somme totale de tous les éléments d'un document imprimé qui influencent le succès d'un groupe de lecteurs, mesuré au moyen de la compréhension, de la vitesse de lecture et de l'intérêt.

On connaît généralement les formules de lisibilités, dites classiques, telles que celles proposées par Flesch (Flesch, 1948), Flesch-Kincaid (Kincaid *et al.*, 1975) ou Henry (Ters, 1976). Depuis les années 2000, l'évolution des solutions en traitement automatique du langage (TAL) a transformé la manière d'évaluer la lisibilité des textes écrits. Les approches se sont alors appuyées reposaient sur des modèles statistiques plus complexes et des routines de TAL capables de capturer des caractéristiques textuelles plus complexes, tels que la syntaxe (Schwarm & Ostendorf, 2005), les relations de discours (Pitler & Nenkova, 2008), ou des caractéristiques acquisitionnelles (Vajjala & Meurers, 2012). Enfin, suite à la révolution neuronales, les formules de lisibilité se sont majoritairement appuyées sur des réseaux de neurones profond (Azpiazu & Pera, 2019), l'architecture transformer (Yancey *et al.*, 2021) et les plongements de mots (Filighera *et al.*, 2019).

La plupart des travaux récents se concentrent soit sur la nature des informations textuelles à capturer (variables ou plongements de mots) ou sur les algorithmes, mais très peu sur les caractéristiques de la situation de lecture. Les chercheurs comme Hiebert & Pearson (2010) ont constaté que les méta-informations textuelles liés à la situation de communication (ou de lecture ici) telle que genre discursif peuvent jouer un rôle essentiel dans la compréhension des textes. Pourtant, Beier *et al.* (2022) indiquent que cette méta-information est souvent négligée dans les études sur la lisibilité des textes. Ignorer le genre dans l'évaluation de la lisibilité risque de conduire à une compréhension inexactes des structures, une incapacité à prendre en compte les spécificités lexicales (termes de spécialité, éléments brachygraphiques, etc.) et à une normalisation des conventions stylistiques.

L'objectif de cet article est donc explorer l'impact du genre discursif sur la qualité des prédictions de la lisibilité des textes en français langue étrangère (FLE). Nous proposons d'explorer différentes approches pour intégrer la notion de genre dans des modèles de lisibilité neuronaux et d'évaluer son effet sur la performance de ces modèles. Avec cette étude, notre contribution constitue un premier pas en vue de démonter l'importance de prendre en compte le genre dans des modèles neuronaux pour la lisibilité du FLE. Plus généralement, cela ouvre des perspectives quant à la prise en plus des caractéristiques de la situation de lecture dans la lisibilité computationnelle.

Dans la suite de l'article, nous brossons un état de l'art centré sur les études réalisées selon le paradigme de la lisibilité computationnelle en mettant l'accent sur l'importance du "genre" (section 2.1). Nous rappelons également quelques notions fondamentales concernant les genres discursifs (section 2.2). Dans la partie suivante (section 3), nous décrivons la méthodologie de cette étude, qui comprend la construction des corpus et la modélisation. La section des résultats (section 4) met en lumière l'impact du genre sur la performance des modèles. Enfin, nous discutons nos principaux résultats avant de conclure (section 5).

2 État de l'art

2.1 Lisibilité computationnelle et genre

Durant la période classique, les formules de Flesch (1948) et de Dale & Chall (1948) avaient déjà été critiquées pour leur manque de robustesse sur des textes de nature différente que ceux du corpus d'entraînement. Par exemple, Brown (1965) montre que la formule de Dale and Chall surestime

la complexité des textes scientifiques. En réaction, Jacobson (1965) entraîne sa propre formule spécialisée pour les textes scientifiques.

Avec l'avènement de l'apprentissage automatique basé sur l'ingénierie de caractéristiques, les approches en lisibilité sont devenues plus aptes à capturer automatiquement diverses caractéristiques textuelles comme la distribution des mots, (Collins-Thompson & Callan, 2005), les relations syntaxiques (Schwarm & Ostendorf, 2005) ou des relations de discours (Pitler & Nenkova, 2008). Ceci a permis de développer des modèles plus robustes, supposément moins sensibles aux particularités des textes analysés.

Par la suite, le développement de l'apprentissage profond a ouvert de nouvelles perspectives en permettant d'encoder directement les propriétés linguistiques pertinentes sans l'intermédiaire d'une ingénierie de caractéristiques complexe. Les travaux récents réalisés dans cette veine privilégient des architectures neurales avancées. Azpiazu & Pera (2019) ont proposé une architecture de Réseau Neuronal Récurrent multi-attentive pour évaluer la lisibilité selon une approche multilingue. Filighera *et al.* (2019) ont employé les réseaux neuronaux et les plongements lexicaux tels que word2vec, GloVe et ELMo, qui offrent des performances comparables aux approches de pointe basées sur l'ingénierie de caractéristiques, tout en étant plus faciles et moins coûteux à adapter à de nouveaux types de textes. Jian *et al.* (2022) ont introduit un modèle hybride basé sur les réseaux de neurones convolutionnels pour évaluer automatiquement la lisibilité des textes anglais, améliorant ainsi leur efficacité et leur précision. Dans sa thèse, Ma (2022) a employé des modèles Transformer pré-entraînés et a obtenu des améliorations par rapport aux approches basées sur les n-grammes. Enfin, l'exploration de l'apprentissage par renforcement profond, notamment par Mohammadi & Khasteh (2019), a également montré des résultats prometteurs.

Depuis quelque temps, certains chercheurs comme Beier *et al.* (2022) ont cependant mentionné du point de vue de la psychologie cognitive et de l'éducation, la nécessité de prendre en considération, non seulement des éléments linguistiques, mais aussi les informations métalinguistiques liées à la situation de lecture telles que la langue première des lecteurs, le sujet/domaine traité du texte, et le genre textuel. Il n'est toutefois pas le premier à se pencher sur cette question. Certains chercheurs ont déjà rapporté une corrélation entre les performance de modèles de prédiction de la lisibilité et le genre. Ainsi, Nelson *et al.* (2012) ont noté une différence de performance de plusieurs formules sur les genres informatif versus narratif. Dell'Orletta *et al.* (2012, 2014) ont montré, sur un corpus en italien, que la prédiction de la lisibilité était fortement influencée par le genre de textes et que, pour cette raison, une notion de lisibilité orientée par genre est nécessaire. Sheehan *et al.* (2013) se distinguent par leur utilisation explicite du genre, via une méthodologie en deux phases qui intègre une classification initiale des textes par genre (informatif, littéraire, ou mixte) avant l'étape d'évaluation de la lisibilité par un modèle de régression.

Plutôt que d'entraîner des modèles différents par genre, Kate *et al.* (2010) exploitent des modèles n-grams entraînés sur des genres différents au sein d'un modèle unique et observent une amélioration de performance. À l'inverse, il est intéressant de noter que Falkenjack *et al.* (2016) ont cherché à prédire l'appartenance à un genre sur la base de variables textuelles de lisibilité, prouvant ainsi l'existence d'un lien fort entre le genre et la lisibilité. Enfin, Li (2022) a évalué la lisibilité des textes de manuels selon leur genre en appliquant des métriques de la linguistique mathématique et a montré l'influence des genres sur la lisibilité et également sur la compréhensibilité des étudiants non natifs de la langue cible.

Dans le domaine du FLE, qui nous intéresse plus particulièrement, Yancey *et al.* (2021) ont évalué la performance d'un modèle de lisibilité affiné avec CamemBERT en fonction de huit genres différents

couramment utilisés en FLE. Ils ont observé un écart de performance de 20 % pour leur meilleur modèle entre les dialogues et des genres atypiques (tels que les chansons, prospectus, recettes, etc.). Toutefois, le genre n'est pas inclus en tant qu'une variable dans leur modèle, ce qui nous a poussé à prolonger cette expérience, avec une perspective centrée sur le genre.

De ces différents travaux, on peut retenir que l'effet du genre sur les performances des modèles de lisibilité a déjà été démontré. Une approche consiste à entraîner des modèles différents par genre (personnalisation), mais elle s'avère coûteuse. Par conséquent, nous privilégierons la stratégie de Sheehan *et al.* (2013), qui consiste à informer un modèle unique du genre de textes qu'il doit analyser. Nous nous démarquons toutefois en adoptant une architecture neuronale, en l'appliquant, pour la première fois, à la lisibilité du FLE et en explorant la meilleure stratégie pour informer le modèle.

2.2 Genre

La question du genre s'ancre dans la pluralité des disciplines et des points de vue : diverses perspectives coexistent pour aborder cette notion (Ablali, 2010). En linguistique, la définition du genre varie selon les écoles. Ainsi, dans le monde anglophone, trois courants principaux dominent les études de genre : les études de genre rhétoriques (Bazerman *et al.*, 1988), la linguistique systémique fonctionnelle et l'anglais sur objectifs spécifiques (Swales, 1990; Bhatia, 1993). Pour le premier, le genre est décrit comme une action sociale, dans de différents contextes de communication (Miller, 1984). Quant à la linguistique systémique fonctionnelle, le genre y est défini comme un processus social comportant plusieurs étapes (*stage*, en anglais) et axé sur des objectifs spécifiques dans différents contextes sociaux (Martin *et al.*, 2010). Enfin, en anglais sur objectifs spécifiques, le genre est considéré comme « des événements communicatifs reconnaissables caractérisés par des objectifs communicationnels et par différents motifs au niveau de la structure, du style, du contenu, et du public visé » (Swales, 1990, 58).

Dans le contexte francophone, la définition de cette notion reflète également une multitude de points de vue et de disciplines. Plusieurs théoriciens tels que Jean-Michel Adam, Dominique Maingueneau et Patrick Charaudeau y ont apporté leur contribution. Adam (1997) met l'accent sur les structures textuelles, analysant comment les textes se constituent en genres selon leurs caractéristiques formelles et fonctionnelles. Maingueneau (2007) considère le genre comme une notion discursive, se concentrant sur les contextes d'énonciation et les conventions influençant la production et la réception des textes. Enfin, Charaudeau (2011) s'intéresse à la dimension communicative et pragmatique des genres, étudiant comment ils structurent la communication dans divers contextes sociaux. Dans le domaine de la didactique du FLE, Beacco (2013) a approfondi la notion de genre de discours. Selon ces auteurs, les genres de discours représentent une forme métalinguistique de communication spécifique à une communauté de discours donnée, et guident les locuteurs dans leur communication verbale.

Il est à noter également que dans le monde francophone, il existe des hésitations terminologiques et un certain flottement des cadres théoriques de référence sur le genre (Adam, 2011). Les notions de "type de texte" et de "genre de texte" sont parfois considérés comme synonyme. La définition du genre/type varie, selon le point de vue du théoricien ou de la traduction, comme démontré par Adam (2011) dans son livre sur la base des titres des numéros de revues de linguistique et de didactique consacrés à cette question.

Nous considérons que le genre discursif est défini par la situation de production et de réception du texte (Adam, 2011). Pour cette étude, qui se concentre sur la lisibilité des textes en FLE, il est

essentiel de prendre en compte la réalité sur le terrain. Dans l'enseignement du FLE, les "types de texte" sont souvent inclus dans la notion de genre. Selon Adam (2011), les textes sont catégorisés en types et prototypes, selon leurs caractéristiques structurelles et fonctions communicatives. Cette notion se distingue de la notion de genre, mais est souvent confondue avec le genre à cause des confusions terminologiques historique détaillées dans (Adam, 2011). Ainsi sur le terrain, ces deux notions sont souvent mélangées. Ce flottement terminologique se retrouve dans le corpus utilisé dans cette étude (cf. section 3.2), mais, dans cet article, la notion de "genre" est utilisée comme un terme parapluie pour les méta-informations textuelles incluant le genre textuel et le type des textes.

2.3 Modélisation de genre en TAL

Cette étude cherchant également à prédire le genre de textes automatiquement, il convient de souligner que des recherches ont observé l'impact de genre textuel sur la performance des modèles de TAL dans diverses circonstances. Dans sa thèse, Frérot (2005) a montré que, selon le genre de corpus, les performances de différentes stratégies de rattachement prépositionnel pour une tâche d'analyse syntaxique varient. Dans le domaine de la recherche d'informations, Mothe & Tanguy (2005) ont découvert que la difficulté des requêtes peut être calculée et prédite à partir de l'analyse d'un certain nombre de traits linguistiques.

L'objectif principal de la modélisation de genre consiste à regrouper les documents par genre. Ainsi, la classification des genres est appliquée dans des tâches différentes. De l'extraction de terminologie pour des textes spécialisés (Todirascu & Guillaume, 2011; Todriascu *et al.*, 2012), à la recherche d'information sur les genres journalistiques (Petrenz & Webber, 2011), genres du Web (Mehler *et al.*, 2010) et genres littéraires (Ollagnier *et al.*, 2015).

Dans les travaux de l'enseignement du FLE assisté par le TAL, le genre est également pris en considération, comme dans l'étude d'analyse linguistique des productions écrites dans apprenants (Audras & Ganascia, 2005).

3 Méthodologie

L' objectif de cette étude est déterminer dans quelle mesure l'intégration de l'information de genre influe sur la performance des modèles de prédiction de la lisibilité. Nous cherchons à savoir si l'ajout de cette information peut enrichir les représentations des textes et quelle méthode d'intégration s'avère être la plus efficace. Pour ce faire, nous comparons les performances d'une architecture informée du genre de texte à celles d'une architecture de référence qui ne connaît pas le genre. Nous avons limité notre étude à l'architecture transformers, car elle correspond à l'état de l'art en lisibilité et peut apprendre le genre indirectement, offrant ainsi un modèle de référence difficile à atteindre.

3.1 Corpus

3.1.1 Présentation des corpus sources

Cette étude se concentre sur l'effet de l'intégration du genre discursif des textes en tant que méta-information sur la performance des modèles de lisibilité en FLE, visant à prédire les niveaux de

compétence linguistique selon le CECR (Conseil de l'Europe, 2001). Le corpus utilisé est issu du corpus élaboré par François (2011) dans sa thèse de doctorat et comprend des extraits des manuels de FLE couvrant les six niveaux du CECR ainsi qu'une diversité de genres textuels tels que textes, dialogues, lettres/emails, publicités, poèmes/chansons, et recettes. En 2021, Yancey *et al.* (2021) ont constitué FLE-CORP, en suivant la même méthodologie, mais en intégrant des genres supplémentaires et en regroupant les textes des niveaux C1 et C2 pour mieux équilibrer les classes. Une portion de ce corpus a déjà servi à une recherche antérieure par Yancey *et al.* (2021), axée sur l'apprentissage profond et l'emploi de variables cognitives et pédagogiques. La section restante du corpus servira à l'entraînement de notre modèle de prédiction de genre, soulignant la continuité et l'évolution dans l'utilisation des ressources pour affiner les outils de mesure de la lisibilité.

3.1.2 Préparation du corpus

À l'origine, nous avions prévu de réutiliser le corpus FLE-CORP (Yancey *et al.*, 2021), qui comprend 8 genres différents. Cependant, les observations des auteurs, qui ont noté une performance élevée de leurs modèles sur les dialogues (niveaux A1, A2) et faible sur des genres diversifiés comme "varias", nous ont conduit à réévaluer la pertinence des données pour notre étude. Une exploration fine du corpus a révélé que certains textes classés comme "Texte à trous" ou "Varias" devaient être éliminés. Le genre "Texte à trous" correspond à des exercices à trous dans les manuels, tandis que "Varias" regroupe plusieurs genres peu organisés. La catégorie "Texte" mélange des textes narratifs et informatifs. Idéalement, il aurait été préférable de réannoter ces textes, mais cette entreprise dépasse le périmètre de cette étude. Nous avons finalement décidé de garder ces genres par souci de comparaison avec les travaux de Yancey *et al.* (2021).

Pour favoriser l'équilibre des classes, les niveaux C1 et C2 ont été regroupés pour former une classe "C" plus peuplée. Ce corpus comprend ainsi 5 classes au total (A1, A2, B1, B2, C). Une fois les textes du genre "Varias" éliminés, nous avons observé un certain déséquilibre des classes, surtout pour le niveau B2. Pour limiter ce déséquilibre, nous avons récupéré des textes du niveau B2 dans le corpus de François (2011). Le corpus final, nommé Corpus-FLE-GENRES, est ainsi construit à partir des données provenant des deux corpus.

Comme l'information du genre des textes n'est pas toujours disponible sur le terrain réel, nous avons décidé de créer un corpus pour entraîner notre modèle de prédiction des genres, à partir des données qui n'ont pas été intégrées dans le Corpus-FLE-GENRES. Ce corpus, nommé Corpus-Genre, est constitué des textes et de leur genre. L'information de genre semble être partiellement associée à la lisibilité des textes dans nos données (cf. table 1), par exemple, les dialogues apparaissent davantage aux niveaux A1 et A2. Nous avons donc préparé un jeu de données dans lequel nous avons éliminé le texte de chaque donnée, ne conservant que la variable de genre en vue d'entraîner un modèle de lisibilité. Ce corpus sera nommé Genre-Text.

3.1.3 Analyse du corpus

Corpus-FLE-GENRE et Genre-Text contiennent 2 101 données au total, ce qui fait 650 données de moins par rapport au corpus utilisé dans l'étude de Yancey *et al.* (2021). Quant à corpus-GENRE contient, il inclut 1 220 textes. La table 1 présente la distribution des données en ce qui concerne les niveaux CECR et les genres dans Corpus-FLE-GENRE. Ce corpus est par la suite réparti en corpus d'entraînement, de validation et de test ayant une proportion de 70/10/20 pour les expériences.

Niveau/Genre	Dialogue	Informative	Mail	Narrative	Phrase	Texte	Total
A1	105	68	43	27	109	98	450
A2	53	131	31	42	65	128	450
B1	28	86	28	28	26	254	450
B2	17	67	24	28	65	100	301
C	0	104	8	55	39	244	450
Total	203	456	134	180	304	824	2 101

TABLE 1 – Distribution des données par genre et par niveau dans Corpus-FLE-GENRE

Nous pouvons constater que Corpus-FLE-GENRE offre une belle variété au niveau du nombre de textes associés à chaque genre et à chaque niveau. Le niveau B2 a toutefois un nombre de textes inférieur de 149 par rapport aux autres niveaux (qui en comportent 450 chacun). Le niveau A1 comporte un nombre relativement élevé de dialogues et de phrases, ce qui est cohérent avec les compétences de communication basique et la construction de phrases simples à ce stade de l'apprentissage. À mesure que le niveau augmente, on note une augmentation du nombre de textes informatifs et narratifs, ce qui va de pair avec la progression des compétences linguistiques vers un niveau plus élaboré. Enfin, le niveau C met particulièrement l'accent sur les textes informatifs et narratifs, reflétant la maîtrise avancée de la langue nécessaire pour comprendre et produire des textes détaillés et nuancés. Cependant, quand on prête attention aux genres majoritaires dans chaque niveau, nous pouvons observer des similitudes entre chaque niveau à savoir que le genre "Text" est majoritaire dans les niveaux A2, B1, B2 et C. La situation est la même pour le genre informatif. Cette situation est très prononcée dans la classe B1 et surtout B2. Quant à Genre-Text, il ne comprend pas de données issues du niveau B2 et très peu du niveau C, ce qui ne semble néanmoins pas constituer une limite. Par contre, le corpus a un grand nombre de textes du genre texte et Phrase, mais très peu de textes du genre narratif et mail.

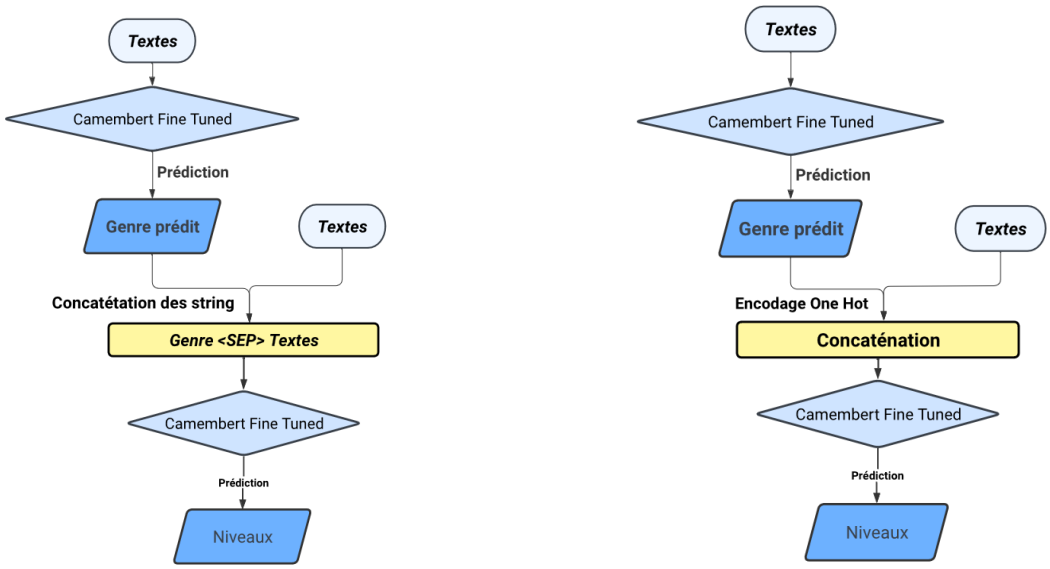
3.1.4 Architectures

Pour notre modèle de référence, nous avons opté pour un CamemBERT affiné (voir figure en annexe 3) disposant uniquement des informations textuelles (sans inclure le genre de texte), car cela correspond à l'approche couramment appliquée en lisibilité.

Ensuite, pour évaluer l'apport de l'information de genre, nous avons entraîné un modèle CamemBERT affiné similaire au modèle de référence, mais avec l'information sur le genre du texte en tant que paramètre d'entrée donné en même temps que le texte. Nous avons exploré deux stratégies pour informer les modèles de lisibilité avec le genre. Premièrement, une chaîne de caractères indiquant le genre (modèle $\text{Concat}_{\text{théorique}}$, voir figure en annexe 5b) est concaténé avec le texte à analyser, et cet input est utilisé pour affiner CamemBERT. Cette stratégie vise à tirer parti de l'information déjà encodée dans CamemBERT concernant les tokens associés au genre. Deuxièmement, nous concaténons le vecteur du texte d'entrée après encodage par CamemBERT avec un vecteur "one-hot" représentant le genre (modèle $\text{One-Hot}_{\text{théorique}}$, voir figure en annexe 4b). Le résultat de cette concaténation est ensuite transmis à la tête de classification de CamemBERT. Ces deux stratégies sont illustrées respectivement dans la Figure 4b en annexe.

La comparaison de ces modèles permet d'évaluer l'apport de l'information de genre à l'identification automatique de la lisibilité d'un texte. Néanmoins, ces modèles constituent un idéal théorique, partant

du principe que la méta-information de genre est connue avec certitude, ce qui n'est pas réalisable avec les limites de la technologie actuelle. Nous avons donc envisagé deux autres modèles afin d'évaluer notre question de recherche dans un environnement plus proche de la réalité des modèles de lisibilité. Ces modèles partagent les mêmes structures que les modèles de concaténation présentés ci-dessus, mais les informations sur le genre proviennent d'un modèle de prédiction automatique du genre. Ces modèles alternatifs sont appelés $\text{Concat}_{\text{réaliste}}$ (voir figure 6b) et $\text{One-Hot}_{\text{réaliste}}$ (voir figure 6c).



(a) Modèle basé sur le nom du genre ($\text{Concat}_{\text{réaliste}}$) (b) Modèle basé sur le vecteur de genre ($\text{One-hot}_{\text{réaliste}}$)

FIGURE 1 – Modèle de concaténation de genre réaliste

Dans cette optique, un modèle automatique est nécessaire pour prédire le genre. Toutefois, un modèle pour prédire les genres en FLE, et plus précisément les genres fournis dans notre contexte d'étude, n'est pas disponible. Nous avons donc développé notre propre modèle d'identification du genre. Comme ce travail ne vise pas à apporter une contribution directe au domaine de la classification automatique du genre, nous avons opté pour un modèle relativement standard, reposant également sur un CamemBERT affiné pour prédire les genres à partir du texte. Il est entraîné sur un échantillon de textes de FLE de notre corpus non utilisés pour l'entraînement des modèles de lisibilité, afin d'éviter d'introduire un biais (pour plus de détails, voir la section 3.1.2).

Tous nos modèles ont été évalués en utilisant une approche de validation croisée à 5 plis pour garantir la robustesse et la généralisabilité des résultats. De plus, l'arrêt anticipé (*early stopping*) basé sur la métrique F-mesure a été employé pour optimiser chaque modèle sans risque de sur-apprentissage. Les performances ont été mesurées à l'aide de métriques standards de classification : l'exactitude , la précision, le rappel, et le F-mesure. Ces mesures sont rapportées sur l'ensemble du corpus, mais également pour chaque niveau de lisibilité et pour chaque genre. Des analyses statistiques ont été menées pour déterminer si les améliorations observées étaient significatives, permettant ainsi d'évaluer de manière rigoureuse l'impact de l'intégration de l'information de genre sur la performance des modèles.

Un modèle de prédiction de la lisibilité basé exclusivement sur les genres sera également développé, mais nous ne discutons pas sa performance, car nous sommes intéressés par son évaluation extrinsèque,

au sein de la tâche de lisibilité.

4 Résultats

Notre étude sur l'intégration de l'information de genre dans la prédiction de la lisibilité des textes en français langue étrangère (FLE) a évalué cinq modèles : CamemBERT, $\text{Concat}_{\text{théorique}}$, $\text{One-Hot}_{\text{théorique}}$, $\text{One-Hot}_{\text{réaliste}}$, et $\text{Concat}_{\text{réaliste}}$. La table 2 présente les scores d'exactitude obtenus par ces modèles sur l'ensemble des données (Global) et sur chacun des genres considérés, ainsi leurs écart-types et leur intervalle de confiance.

Genre	**CamemBERT**	**$\text{Concat}_{\text{théorique}}$**	**$\text{One-Hot}_{\text{théorique}}$**	**$\text{One-Hot}_{\text{réaliste}}$**	**$\text{Concat}_{\text{réaliste}}$**	**Écart-type**
Texte	0,53	0,51	0,53	0,51	0,48	0,04
Informative	0,64	0,65	0,65	0,59	0,56	0,03
Mail	0,43	0,54	0,67	0,52	0,43	0,09
Phrase	0,42	0,46	0,53	0,44	0,53	0,04
Dialogue	0,69	0,74	0,69	0,66	0,68	0,02
Narrative	0,56	0,61	0,57	0,52	0,53	0,06
Global	0,55	0,57	0,59	0,54	0,55	0,02

TABLE 2 – Scores d'exactitude moyens par genre pour chaque modèle avec écart-type et intervalle de confiance à 95 %.

L'intégration de l'information de genre montre des gains d'exactitude de 0,02 ($\text{Concat}_{\text{théorique}}$) et 0,05 ($\text{One-Hot}_{\text{théorique}}$) par rapport au modèle de référence (CamemBERT), bien que ces gains ne soient pas statistiquement significatifs. Certains genres bénéficient davantage de l'intégration de l'information de genre. Par exemple, pour le genre "Dialogue", $\text{Concat}_{\text{théorique}}$ améliore l'exactitude de 0,05 par rapport à CamemBERT. Le genre "Mail" voit l'exactitude du modèle $\text{One-Hot}_{\text{théorique}}$ atteindre 0,67 contre 0,43 pour CamemBERT. L'utilisation de modèles réalistes, qui prédisent le genre des textes, diminue logiquement les performances. Ainsi, l'exactitude du modèle $\text{One-Hot}_{\text{réaliste}}$ diminue de 0,05 par rapport à $\text{One-Hot}_{\text{théorique}}$, et $\text{Concat}_{\text{réaliste}}$ perd 0,02.

Les intervalles de confiance et les écarts-types fournissent des informations cruciales sur la stabilité et la précision des performances des modèles. L'analyse des intervalles de confiance (IC) (voir Figure 5)pour l'accuracy des modèles sur différents genres de textes révèle que les modèles '$\text{Concat}_{\text{réaliste}}$' et 'Baseline' présentent généralement des IC plus étroits, indiquant une plus grande stabilité dans leurs prédictions. Par exemple, '$\text{Concat}_{\text{réaliste}}$' affiche l'IC le plus étroit pour les genres 'texte' (0.039668) et 'mail' (0.104243), ce qui suggère une performance cohérente. En revanche, les modèles '$\text{One-hot}_{\text{réaliste}}$' et '$\text{Concat}_{\text{théorique}}$' montrent des IC plus larges dans certains genres, tels que 'sentence' et 'dialogue', indiquant une variabilité plus importante dans leurs performances de prédiction. Les genres 'narrative' et 'mail' montrent une plus grande variabilité dans les performances de certains modèles, suggérant que ces types de textes posent plus de défis pour des prédictions stables. Par exemple, le modèle 'Baseline' a un IC très large pour le genre 'narrative' (0.277162), tandis que '$\text{Concat}_{\text{théorique}}$' affiche un IC large pour 'mail' (0.324186). Le tableau 5 montre que la stabilité des performances des modèles varie non seulement entre les modèles, mais aussi en fonction des genres de texte.

T-test sur les intervalles de confiance par genres de chaque modèles et les écarts-types sur l'exactitude fournissent des informations cruciales sur la stabilité et la précision des performances des modèles(voir Figure 6 & 7). Les modèles $\text{Concat}_{\text{théorique}}$ et $\text{One-hot}_{\text{théorique}}$ semblent offrir une stabilité comparable

avec une variabilité de performance relativement faible. Les modèles hybrides montrent plus de variabilité, suggérant une performance moins stable. La plupart des comparaisons montrent des P-Values non significatives, indiquant que les améliorations de performance par rapport au modèle Baseline ne sont pas statistiquement significatives pour la plupart des types de texte, à quelques exceptions près comme $\text{Concat}_{\text{réaliste}}$ pour certains genres.

Quant à l'analyse des erreurs, nous allons la traiter selon deux perspectives : l'évaluation des matrices de confusion au niveau des genres et la corrélation entre les caractéristiques linguistiques et les erreurs de prédiction. Les matrices de confusion montrent que les modèles proposés améliorent leurs performances par rapport au modèle de référence (voir Figures 5 et 6). Pour le genre "Texte", les modèles $\text{Concat}_{\text{théorique}}$ et $\text{Concat}_{\text{réaliste}}$ réduisent légèrement les confusions, notamment aux niveaux A1 et C. Cependant, des confusions persistent entre les niveaux intermédiaires tels que A2, B1 et B2, souvent dues à la similarité des caractéristiques entre ces niveaux et genres, indiquant un besoin d'amélioration dans la différenciation des caractéristiques.

Avec l'outil FABRA (Wilkens *et al.*, 2022), nous avons analysé la corrélation entre certaines caractéristiques linguistiques (longueur des textes et longueur moyenne des mots) et les erreurs de prédiction (voir Figure 7). Les corrélations observées sont faibles, indiquant que ces caractéristiques n'ont pas une influence significative sur les erreurs. Cela suggère que d'autres facteurs, comme la complexité syntaxique, la fréquence des mots, et les relations sémantiques, pourraient être plus pertinents pour expliquer les erreurs.

Pour améliorer les modèles, il serait bénéfique d'explorer ces autres variables linguistiques plus en profondeur. Une analyse détaillée de la complexité syntaxique, de la fréquence des mots et d'autres aspects linguistiques pourrait fournir des diagnostics précieux pour affiner les modèles et améliorer la précision des prédictions de lisibilité.

5 Conclusion

Notre étude approfondit la compréhension de l'impact du genre sur la prédiction de la lisibilité en FLE, confirmant l'importance cruciale de ces méta-informations. Bien que l'intégration du genre n'ait pas conduit à des améliorations significatives globalement, nos analyses par genre révèlent des nuances spécifiques, confirmant l'impact soulevé dans certaines études en lisibilité mentionnées dans la section 2.1. Même avec ce signal faible, nous concluons que les informations sur le genre aident à identifier le niveau de lisibilité d'un texte. Toutefois, cette tâche est loin d'être achevée. En effet, notre travail montre que plusieurs éléments doivent être mis en place afin d'obtenir un véritable modèle de lisibilité intégrant le genre. Pour les futures recherches, nous estimons que la collecte de données variées est fondamentale, y compris la nécessité d'un corpus doté d'informations bien calibrées sur le genre. D'autre part, il est important de développer des architectures qui prennent efficacement en compte le genre. Notre étude préliminaire montre que le genre doit être traité comme une information riche et le modèle doit lui accorder une position préférentielle, afin d'affiner et d'améliorer la performance des modèles de lisibilité.

Références

ABLALI D. (2010). Linguistique des genres. exploration sur corpus. *Linguistique & Littérature : Cluny*, **40**, 251.

ADAM J.-M. (1997). Genres, textes, discours : pour une reconception linguistique du concept de genre. *Revue belge de philologie et d'histoire*, **75**(3), 665–681.

ADAM J.-M. (2011). La linguistique textuelle. introduction à l'analyse textuelle des discours. *Semen*, **32**, 182–185. DOI : https ://doi.org/10.4000/semen.9411.

AUDRAS I. & GANASCIA J.-G. (2005). Des outils informatiques au service du passage à l'écrit d'apprenants.

AZPIAZU I. & PERA M. (2019). Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, **7**, 421–436. DOI : 10.1162/tacl_a_00278.

BAZERMAN C. *et al.* (1988). *Shaping written knowledge : The genre and activity of the experimental article in science*, volume 356. University of Wisconsin Press Madison.

BEACCO J.-C. (2013). L'approche par genres discursifs dans l'enseignement du français langue étrangère et langue de scolarisation. *Pratiques. Linguistique, littérature, didactique*, (157-158), 189–200. Number : 157-158 Publisher : Association CRESEF, DOI : 10.4000/pratiques.3838.

BEIER S., BERLOW S., BOUCAUD E., BYLINSKII Z., CAI T., COHN J., CROWLEY K., DAY S. L., DINGLER T., DOBRES J., HEALEY J., JAIN R., JORDAN M., KERR B., LI Q., MILLER D. B., NOBLES S., PAPOUTSAKI A., QIAN J., REZVANIAN T., RODRIGO S., SAWYER B. D., SHEPPARD S. M., STEIN B., TREITMAN R., VANEK J., WALLACE S. & WOLFE B. (2022). Readability Research : An Interdisciplinary Approach. *Foundations and Trends® in Human–Computer Interaction*, **16**(4), 214–324. DOI : 10.1561/1100000089.

BHATIA V. (1993). *Analysing Genre : Language Use in Professional Settings*. Applied linguistics and language study. Longman.

BROWN W. (1965). Science textbook selection and the Dale-Chall formula. *School Science and Mathematics*, **65**(2), 164–167.

CHARAUDEAU P. (2011). Chapitre 1. l'information comme acte de communication. *Medias-Recherches*, **2**, 21–28.

COLLINS-THOMPSON K. & CALLAN J. (2005). Predicting Reading difficulty with Statistical Language Models. *JASIST*, **56**, 1448–1462. DOI : 10.1002/asi.20243.

CONSEIL DE L'EUROPE (2001). *Cadre Européen Commun de Référence pour les Langues : Apprendre, enseigner, évaluer*.

DALE E. & CHALL J. (1948). A formula for predicting readability. *Educational research bulletin*, **27**(1), 11–28.

DALE E. & CHALL J. (1949). The concept of readability. *Elementary English*, **26**(1), 19–26.

DELL'ORLETTA F., VENTURI G. & MONTEMAGNI S. (2012). Genre-oriented Readability Assessment : a Case Study. p. 91–98.

DELL'ORLETTA F., WIELING M., VENTURI G., CIMINO A. & MONTEMAGNI S. (2014). Assessing the Readability of Sentences : Which Corpora and Features ? In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 163–173, Baltimore, Maryland : Association for Computational Linguistics. DOI : 10.3115/v1/W14-1820.

FALKENJACK J., SANTINI M. & JONSSON A. (2016). An Exploratory Study on Genre Classification using Readability Features.

FILIGHERA A., STEUER T. & RENSING C. (2019). *Automatic Text Difficulty Estimation Using Embeddings and Neural Networks*, p. 335–348. DOI : 10.1007/978-3-030-29736-7_25.

FLESCH R. (1948). A new readability yardstick. *Journal of Applied Psychology*, **32**(3), 221–233. Place : US Publisher : American Psychological Association, DOI : 10.1037/h0057532.

FRANÇOIS T. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère*. Thèse de doctorat, Université Catholique de Louvain.

FRÉROT C. (2005). *Construction et évaluation en corpus variés de lexiques syntaxiques pour la résolution des ambiguïtés de rattachement prépositionnel*. Thèse de doctorat, Université de soutenance.

HIEBERT E. H. & PEARSON P. D. (2010). An examination of current text difficulty indices with early reading texts. reading research report #10-01.

JACOBSON M. (1965). Reading difficulty of physics and chemistry textbooks. *Educational and Psychological Measurement*, **25**(2), 449–457.

JIAN L., XIANG H. & LE G. (2022). English text readability measurement based on convolutional neural network : A hybrid network model. *Computational Intelligence and Neuroscience*, **2022**, 1–9. DOI : 10.1155/2022/6984586.

KATE R., LUO X., PATWARDHAN S., FRANZ M., FLORIAN R., MOONEY R., ROUKOS S. & WELTY C. (2010). Learning to Predict Readability using Diverse Linguistic Features. p. 546–554.

KINCAID J. P., FISHBURNE J., ROBERT P. R., RICHARD L. C. & BRAD S. (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel :*. Rapport interne, Defense Technical Information Center, Fort Belvoir, VA. DOI : 10.21236/ADA006655.

LI W. (2022). Text genres, readability and readers' comprehensibility. *European Journal of Computer Science and Information Technology*, **10**, 52–62. DOI : 10.37745/ejcsit.2013/vol10n45262.

MA C. (2022). *Readability Assessment with Pre-trained Transformer Models*. Thèse de doctorat.

MAINGUENEAU D. (2007). Genres de discours et modes de généricité. *Le français aujourd'hui*, (4), 29–35.

MARTIN J., CHRISTIE F. & ROTHERY J. (2010). Social processes in education : a reply to sawyer and watson. *Metaphor*, (4), 51–52.

MEHLER A., SHAROFF S. & SANTINI M. (2010). *Genres on the Web : Computational Models and Empirical Studies*, volume 42. DOI : 10.1007/978-90-481-9178-9.

MILLER C. R. (1984). Genre as social action. *Quarterly journal of speech*, **70**(2), 151–167.

MOHAMMADI H. & KHASTEH S. H. (2019). Text as environment : A deep reinforcement learning text readability assessment model.

MOTHE J. & TANGUY L. (2005). Linguistic features to predict query difficulty.

NELSON J., PERFETTI C., LIBEN D. & LIBEN M. (2012). Measures of Text Difficulty :. *Council of Chief State School Officers, Washington, DC*.

OLLAGNIER A., FOURNIER S. & BELLOT P. (2015). Analyse en dépendance et classification de requêtes en langue naturelle, application à la recommandation de livres [dependency parsing and classification of natural language queries : application to book recommendation]. *Traitement Automatique des Langues*, **56**(3), 23–47.

PETRENZ P. & WEBBER B. (2011). Stable classification of text genres. *Computational Linguistics*, **37**, 385–393. DOI : 10.1162/COLI_a_00052.

PITLER E. & NENKOVA A. (2008). Revisiting readability : a unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Proces-*

sing - EMNLP '08, p. 186, Honolulu, Hawaii : Association for Computational Linguistics. DOI : 10.3115/1613715.1613742.

SCHWARM S. & OSTENDORF M. (2005). Reading level assessment using support vector machines and statistical language models. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, p. 523–530.

SHEEHAN K. M., FLOR M. & NAPOLITANO D. (2013). A two-stage approach for generating unbiased estimates of text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, p. 49–58.

SWALES J. (1990). *Genre Analysis : English in Academic and Research Settings*. Cambridge Applied Linguistics. Cambridge University Press.

TERS F. (1976). Henry (Georges). — Comment mesurer la lisibilité. *Revue française de pédagogie*, **36**(1), 71–74.

TODIRASCU A. & GUILLAUME B. (2011). Classyn : classer les documents selon le genre textuel.

TODRIASCU A., PADÓ S., KISSELEW M., KRISCH J. & HEID U. (2012). French and german corpora for audience-based text type classification.

VAJJALA S. & MEURERS D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, p. 163–173.

WILKENS R., ALFTER D., WANG X., PINTARD A., TACK A., YANCEY K. P. & FRANÇOIS T. (2022). FABRA : French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 1217–1233, Marseille, France : European Language Resources Association.

YANCEY K., PINTARD A. & FRANCOIS T. (2021). Investigating readability of french as a foreign language with deep learning and cognitive and pedagogical features. *Lingue e linguaggio*, **20**(2), 229–258.

Annexes

Annexe 1 : Distribution de longueur de texte par genre et par niveau

Ce tableau présente le nombre de mots maximum, minimum et moyen par genre dans le corpus FLE-GENRE.

Genre	Dialogue	Informative	Mail	Narrative	Phrase	Texte
MaxNbmot	522	1682	646	1427	340	2485
MinNbmot	12	15	23	3	18	15
MoyenNbmot	146,67	228,22	138,52	225,21	73,15	251,48

TABLE 3 – Nombre de mots par genre dans Corpus-FLE-GENRE

Ce tableau montre la distribution de la longueur des textes par niveau de compétence. Elle permet de visualiser comment la complexité textuelle varie en fonction des niveaux de compétence des apprenants.

Niveau	A1	A2	B1	B2	C
MaxNbmot	346	644	719	2485	1682
MinNbmot	12	15	24	3	15
MoyenNbmot	94,01	134,58	191,60	254,48	348,26

TABLE 4 – Nombre de mots par niveau dans Corpus-FLE-GENRE

Annexe 2 : Intervalles de Confiance

Ce tableau présente les intervalles de confiance de précision pour chaque genre et chaque modèle. Ces données sont essentielles pour évaluer la variabilité et la stabilité des performances des modèles.

Modèle	**Texte**	**Phrase**	**Narratif**	**Informatif**	**Dialogue**	**Mail**
Baseline	0.0693	0.0909	0.2772	0.0436	0.2749	0.1286
Concat_théorique	0.0827	0.1515	0.1922	0.1184	0.2749	0.3242
One_hot_théorique	0.0834	0.1773	0.1676	0.1303	0.1632	0.1749
One_hot_réaliste	0.1071	0.0611	0.0850	0.1639	0.2290	0.2145
Concat_réaliste	0.0397	0.0645	0.0862	0.0702	0.1524	0.1042

TABLE 5 – Intervalles de Confiance de Précision par Genre et par Modèle

Annexe 3 : Comparaisons Statistiques

Ces tableaux présentent les résultats des tests T et les valeurs P pour les comparaisons des performances des modèles par genre. Ils aident à déterminer si les différences observées entre les modèles sont statistiquement significatives.

Type de Texte	Modèle de Base	Modèle de Comparaison	Statistique T	Valeur P
text	Baseline	Concat	-0.2733	0.7930
sentence	Baseline	Concat	-1.0075	0.3434
narrative	Baseline	Concat	-0.4031	0.6993
informative	Baseline	Concat	0.3593	0.7287
dialogue	Baseline	Concat	0.0798	0.9384
mail	Baseline	Concat	0.8824	0.4033
text	Baseline	One_hot	-0.5426	0.6081
sentence	Baseline	One_hot	-1.1712	0.2758
narrative	Baseline	One_hot	-0.5081	0.6278
informative	Baseline	One_hot	0.2621	0.8002
dialogue	Baseline	One_hot	0.1984	0.8487
mail	Baseline	One_hot	-0.3696	0.7214
text	Baseline	One_hot_réaliste	1.5300	0.1848
sentence	Baseline	One_hot_réaliste	-0.6657	0.5271
narrative	Baseline	One_hot_réaliste	-0.3196	0.7615
informative	Baseline	One_hot_réaliste	0.2390	0.8174
dialogue	Baseline	One_hot_réaliste	0.2809	0.7876
mail	Baseline	One_hot_réaliste	-1.8058	0.1088
text	Baseline	Concat_réaliste	-6.1665	0.0016
sentence	Baseline	Concat_réaliste	-2.7598	0.0250
narrative	Baseline	Concat_réaliste	-0.8314	0.4445
informative	Baseline	Concat_réaliste	-2.8235	0.0239
dialogue	Baseline	Concat_réaliste	-0.8194	0.4421
mail	Baseline	Concat_réaliste	-2.7423	0.0295

TABLE 6 – Statistiques T et Valeurs P pour les Comparaisons de Modèles par Genre-1

Type de Texte	Modèle de Base	Modèle de Comparaison	Statistique T	Valeur P
text	One_hot	One_hot_réaliste	1.6450	0.1399
sentence	One_hot	One_hot_réaliste	0.7413	0.4848
narrative	One_hot	One_hot_réaliste	0.3397	0.7441
informative	One_hot	One_hot_réaliste	-0.0228	0.9823
dialogue	One_hot	One_hot_réaliste	0.1095	0.9155
mail	One_hot	One_hot_réaliste	-1.5003	0.1719
text	One_hot	Concat_réaliste	-2.4254	0.0688
sentence	One_hot	Concat_réaliste	-1.3219	0.2249
narrative	One_hot	Concat_réaliste	-0.4000	0.7026
informative	One_hot	Concat_réaliste	-2.6811	0.0335
dialogue	One_hot	Concat_réaliste	-1.4040	0.1980
mail	One_hot	Concat_réaliste	-2.4572	0.0429
text	One_hot_réaliste	Concat_réaliste	-4.1242	0.0134
sentence	One_hot_réaliste	Concat_réaliste	-2.6081	0.0334
narrative	One_hot_réaliste	Concat_réaliste	-1.0042	0.3458
informative	One_hot_réaliste	Concat_réaliste	-2.6820	0.0332
dialogue	One_hot_réaliste	Concat_réaliste	-1.5581	0.1578
mail	One_hot_réaliste	Concat_réaliste	-0.7142	0.4977

TABLE 7 – Statistiques T et Valeurs P pour les Comparaisons de Modèles par Genre-2

Annexe 4 : Distribution de longueur de texte par genre et par niveau

Cette figure montre la distribution de la longueur des textes par genre et par niveau de compétence.

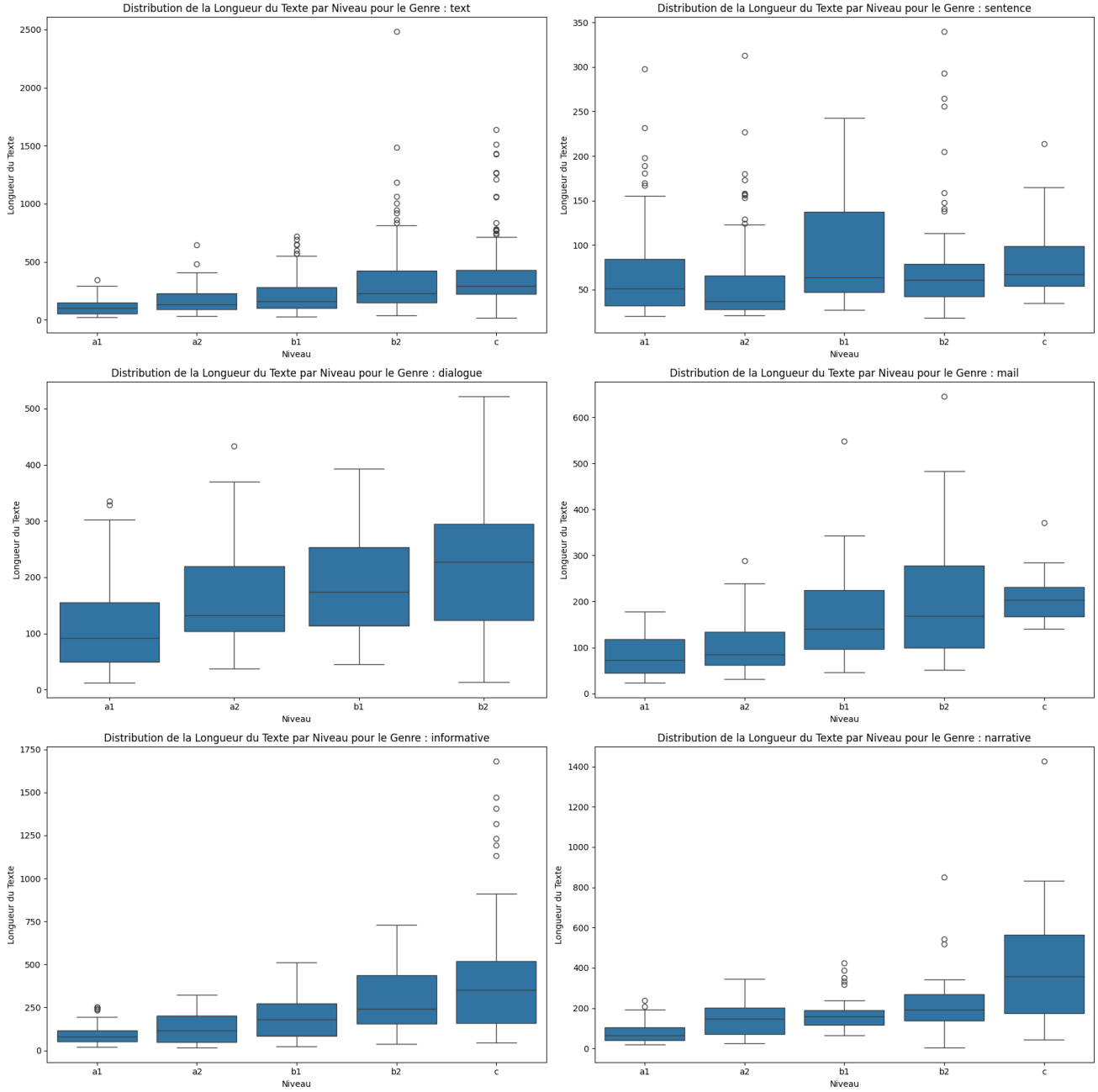


FIGURE 2 – Distribution de longueur de texte par genre et niveau

Annexe 5 : Modèles et Architectures

Ces figure illustre le modèle de base (Baseline) utilisé dans l'étude pour prédire la lisibilité des textes qui ne prend en compte que les informations textuelles sans intégrer les méta-informations de genre.

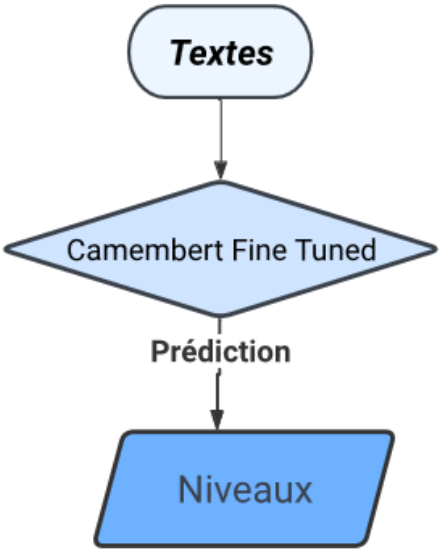
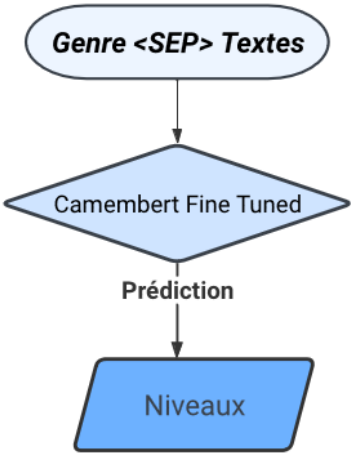
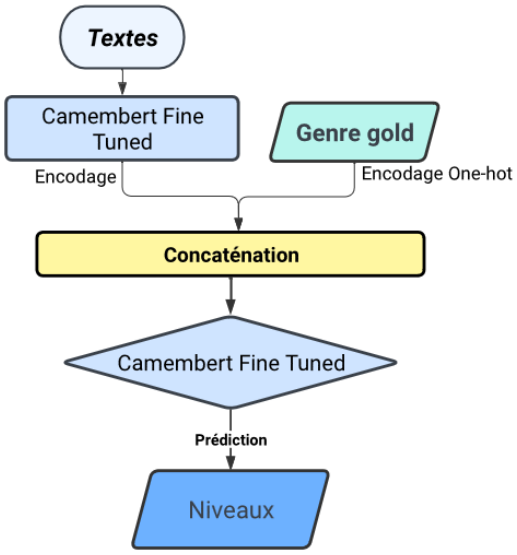


FIGURE 3 – Modèle Baseline

Cette figure présente deux variantes de modèles théoriques utilisant la concaténation de l'information de genre avec les textes. Le modèle (a) utilise le nom du genre, tandis que le modèle (b) utilise un vecteur "one-hot" pour représenter le genre.



(a) Modèle basé sur le nom du genre ($\text{Concat}_{\text{théorique}}$)

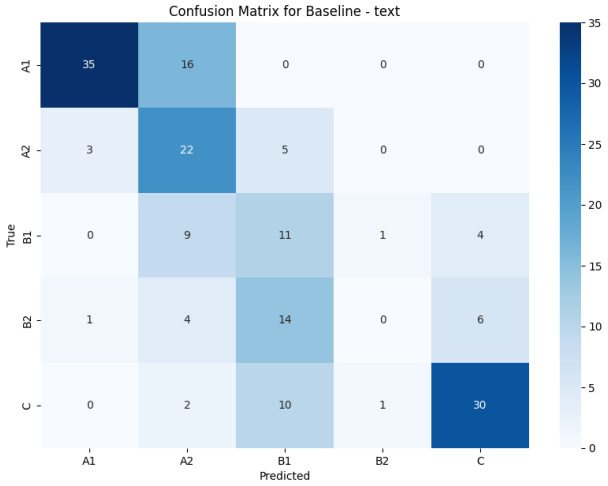


(b) Modèle basé sur le vecteur de genre ($\text{One-hot}_{\text{théorique}}$)

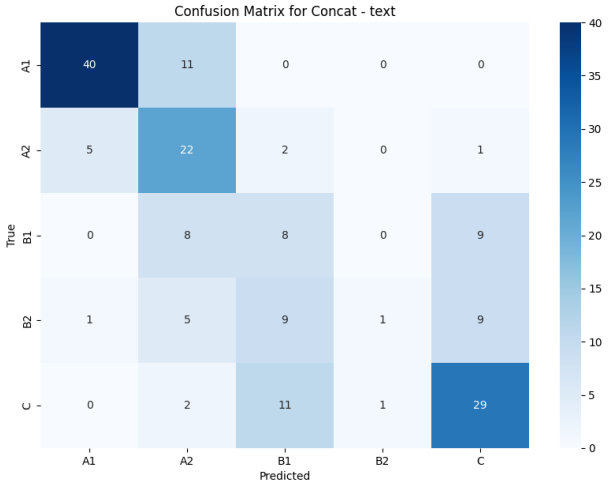
FIGURE 4 – Modèle de concaténation de genre Théorique

Annexe 6 : Matrices de Confusion

Ces matrices de confusion montrent les performances des différents modèles dans la prédiction de la lisibilité pour le genre 'Texte'. Elles illustrent les confusions entre les niveaux de compétence, ce qui aide à identifier les domaines où les modèles peuvent être améliorés.

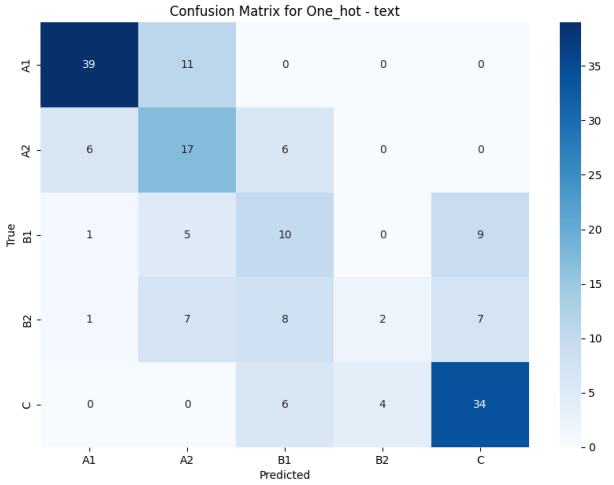


(a) Matrice de confusion - Texte (Baseline)

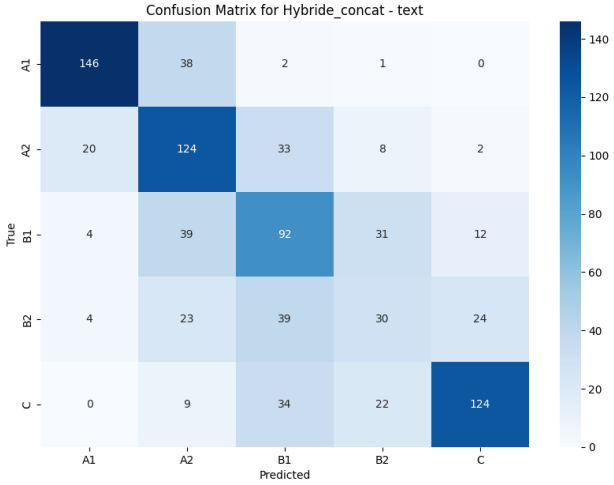


(b) Matrice de confusion - Texte ($\text{Concat}_{\text{théorique}}$)

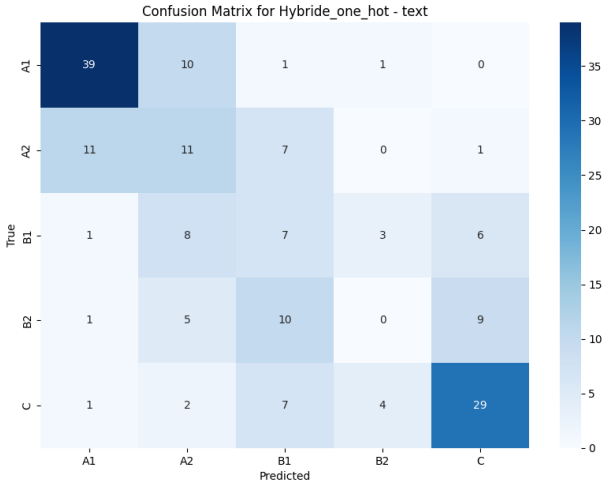
FIGURE 5 – Matrices de confusion de la prédiction de la lisibilité pour le genre 'Texte'



(a) Matrice de confusion - Texte ($\text{One-hot}_{\text{théorique}}$)



(b) Matrice de confusion - Texte ($\text{Concat}_{\text{réaliste}}$)



(c) Matrice de confusion - Texte ($\text{One-hot}_{\text{réaliste}}$)

FIGURE 6 – Matrices de confusion de la prédiction de la lisibilité pour le genre ’Texte’ bis

Annexe 7 : Corrélations et Analyses d'Erreur

Cette figure montre les corrélations entre certaines caractéristiques linguistiques (comme la longueur des textes et la longueur moyenne des mots) et les erreurs de prédiction. Les corrélations observées sont faibles, suggérant que d'autres facteurs linguistiques pourraient être plus pertinents pour expliquer les erreurs.

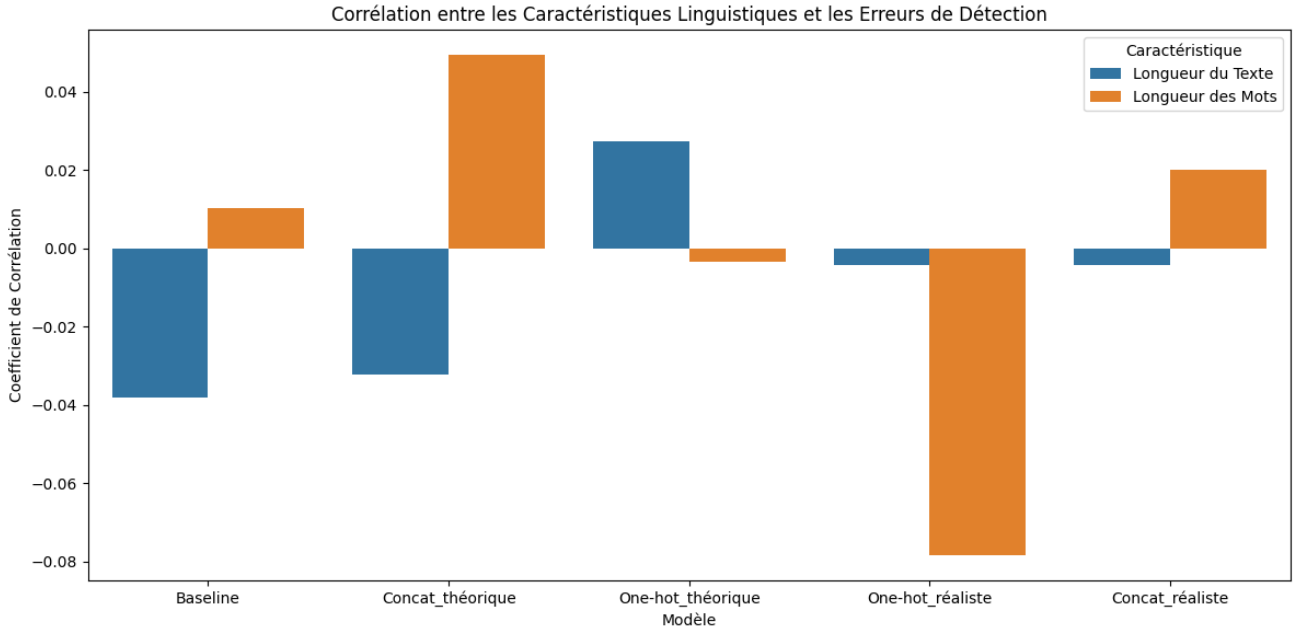


FIGURE 7 – Visualisation Corrélation Caractéristiques linguistiques et Erreurs

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

LLM-Generated Contexts to Practice ESP Vocabulary: Corpus Presentation and Comparison

Iglika Nikolova-Stoupak[1] Serge Bibauw[3] Amandine Dumont[2] Françoise Stas[2]
Patrick Watrin[1] Thomas François[1]
(1) CENTAL, (2) ILV, (3) IACCHOS, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgique
{iglika.nikolova, serge.bibauw, amandine.dumont, francoise.stas,
patrick.watrin, thomas.francois}@uclouvain.be

RÉSUMÉ

Contextes générés par LLM pour pratiquer le vocabulaire anglais de spécialité : présentation de corpus et comparaison

Ce projet analyse la capacité des LLM (grands modèles de langue) et de corpus web à fournir des contextes visant la pratique et l'apprentissage de vocabulaire anglais spécialisé dans un contexte universitaire. Le but sous-jacent est d'augmenter le volume d'exemples utilisables et leur facilité de mise au point tout en conservant la qualité actuelle où ils sont conçus par des spécialistes. Sur la base d'un jeu de contextes de référence — utilisés en classe — visant à l'apprentissage d'une liste de vocabulaire spécialisé, nous comparons les caractéristiques linguistiques de contextes générés par trois LLM récents de différentes tailles (Mistral-7B-Instruct, Vicuna-13B et Gemini 1.0 Pro) et un corpus de contextes extraits automatiquement d'articles de sites web spécialisés. Les caractéristiques textuelles évaluées incluent la longueur, la morphosyntaxe, la sémantique et le niveau discursif. En fin de compte, nous identifions le corpus généré par un LLM (Gemini) dans un scénario one-shot comme étant celui qui se rapproche le plus du corpus de référence.

ABSTRACT

This project analyses the ability of LLMs (large language models) and web-based corpora to provide contexts for the practice and acquisition of specialised English vocabulary in a university context. The underlying purpose is to increase the volume of usable examples and their ease of generation while retaining the currently established quality of learning materials as crafted by specialists. We present a reference corpus of contexts — handpicked by expert teachers — for a specialised vocabulary list, as well as related corpora generated by three recent LLMs of different sizes (Mistral-7B-Instruct, Vicuna-13B, and Gemini 1.0 Pro) and a corpus extracted from articles crawled from specialised websites. We evaluate and compare the corpora based on a representative set of textual characteristics (length-based, morphosyntactic, lexico-semantic, and discourse-related). Ultimately, we identify a corpus generated by an LLM (Gemini) in a one-shot setting as coming closest to the reference one.

MOTS-CLÉS : grands modèles de langue, vocabulaire anglais spécialisé, traits de lisibilité.

KEYWORDS: large language models, specialised English vocabulary, readability features.

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 472–498.

1 Introduction

The present study is part of a broader project conducted at Université catholique de Louvain that aims at leveraging natural language processing (NLP) tools to facilitate the acquisition of English for specific purposes (ESP) vocabulary by providing multiple examples of its natural use in context. The project involves several university courses mapped to proficiency levels B1 to C1 and designed to teach ESP to STEM (science, technology, engineering, and mathematics) students. In these courses, students have to master predefined vocabulary lists comprised of both strictly scientific vocabulary (e.g. "a chemical") and other vocabulary that commonly appears in scientific contexts (e.g. "to assume"). A preliminary survey revealed that the target students currently study the lists as is, at best using flash cards. Research has, however, shown the importance of regularly exposing learners to words in authentic and informative contexts (Huckin & Coady, 1999; Ramos & Dario, 2015; Godwin-Jones, 2018). Collecting authentic contexts is, unfortunately, very time-consuming for ESP teachers, which generally impedes offering enough authentic contexts to students as support for vocabulary learning. This is why we intend to automatically retrieve and generate contexts of use for any target ESP word, thereby significantly decreasing the burden in terms of both time and effort that currently results from the manual collection of context sentences.

In this study, we explore two ways of collecting such contexts : firstly, their extraction from a large corpus made up of websites routinely exploited by ESP teachers ; and secondly, their generation using three large language models (LLMs) : Mistral-7B-Instruct, Vicuna-13B, and Gemini 1.0 Pro. More specifically, this paper analyses the linguistic characteristics – using standard readability- and stylometry-related variables – of sentences in our web-crawled and LLM-generated corpora and their similarities and differences with a reference corpus. The last is made up of examples handpicked by ESP teachers among vocabulary test materials used in university ESP courses.

2 Background

2.1 Automatic Text Retrieval/Generation in Language Learning

The advancement of the Internet and Big Data has long been viewed as an opportunity for language teachers and learners to get hold of a large quantity of learning materials that are characterised with authenticity and timeliness. One of the established roles of NLP in EFL studies is the retrieval of relevant materials from the web, often followed by their evaluation, annotation and/or adaptation and the generation of related exercises (Litman, 2016; Meurers, 2021). Via a survey, Wilson (2004) evaluates students' practices and satisfaction in relation to the use of web resources for independent ESL study as well as assembles a list of recommended websites for learners. Heilman *et al.* (2008) first create a corpus of web-crawled texts to be used for vocabulary and reading practice and then develop a system called REAP Search that allows the selection of particular texts from the corpus based on defined constraints (e.g. the presence of specific words). Similarly, Yoon *et al.* (2017) retrieve a number of YouTube videos based on criteria such as the existence of manual transcriptions and go on to use them in the generation of listening exercises for the TOEIC certificate exam. Other studies (Meurers *et al.*, 2010; Hussin *et al.*, 2010; Jin & Lu, 2018) focus less on the specificities of web-based textual sources than on their later enrichment and annotation for use in language learning.

A recent and revolutionary technology capable of producing humanlike language, LLMs have already

been exploited in a variety of scenarios, including the creation of EFL learning materials. Young & Shishido (2023a) had ChatGPT produce text of different proficiency levels based on articles from an online newspaper. The levels' correctness was confirmed through a readability analysis. In another experiment, Young & Shishido (2023b) used ChatGPT for the generation of dialogues. The general topic and participants were indicated in the prompts, and multiple readability metrics were used to analyse the suitability of the derived dialogues and to determine their best target audience. They concluded that the dialogues are most suitable for the A2 level (followed by B1), while students of higher proficiency levels may miss out on elements like colloquial expressions and phrasal verbs. Shaikh *et al.* (2023) made use of a questionnaire to evaluate users' views on the effectiveness of ChatGPT in specialised EFL studies. Students of different nationalities, proficiency levels and fields of study were asked to converse with ChatGPT on different topics, and engage in vocabulary practice and have the virtual assistant edit text they produce. Students' opinions were generally favourable, in particular with regard to ChatGPT's assistance in vocabulary acquisition.

2.2 Readability Features

Readability, often considered to date back to Sherman's experiments in 1893, is a primary measure used for quantitative description of text (DuBay, 2007). Its main purpose is the estimation of a textual unit's reading difficulty and, thereby, appropriateness for a given audience (typically, children of a certain age). To this aim, numerous readability formulas have been developed throughout the years [1], relying on a variety of shallow textual characteristics (such as sentence or word length), more advanced ones (e.g. syntax or discourse properties), or comparison against vocabulary lists. Although state-of-the-art readability estimations are now mostly provided by deep neural models (Vajjala, 2022), readability formulas based on engineered features have dominated the field for almost 90 years, and some of the best-performing current systems rely on both deep learning and such features within a hybrid architecture (Deutsch *et al.*, 2020; Wilkens *et al.*, 2024).

Features are central to readability because they link the theory of the reading processes to the pragmatic approach typical to predictive modelling. The reading process is made up of three main steps : visual perception, decoding, and comprehension, and each of them can be impacted by a given text's characteristics (François, 2011). For instance, more frequent words are generally decoded faster than rare ones, some syntactic structures seem harder to parse for the brain than others, and lexemes representing abstract concepts are generally activated more slowly in the brain than more concrete ones. There is a large amount of psycholinguistic studies that have stressed a specific aspect of language that is likely to impact the reading process (Ferrand, 2007). In this work, we exploit the long tradition of readability variables, hundreds of which have been investigated, parametrised and tested on different corpora since the 1920s.

Many of the mentioned features have also been utilised in textual descriptions that are not strictly related to complexity. The field of second language acquisition, in particular, aims to describe the language produced by language learners and also resorts to various features, some of which overlap with readability ones. For instance, "lexical richness" (close to "lexical diversity"), strongly interconnected with type-to-token ratio, is a concept defined by Yule (1944) and used to estimate a particular author's (or text's) distinctive linguistic characteristics. A related term used in stylometry is "lexical sophistication", typically associated with word frequency and concreteness (Kyle, 2019). Other metrics aim to measure "lexical density", which refers to the proportion of content words in the

1. For surveys of the field, see François (2011); Collins-Thompson (2014); Vajjala (2022).

text (Ure, 1971). Cech & Kubat (2018) specifically refer to the "morphological richness" of a text, defined as the difference between the vocabulary richness (i.e. type-to-token ratio) of lemmas and words, as an important characteristic in authorship attribution.

3 Methods

In the current study, we collect and analyse two general types of contexts for ESP vocabulary learning : generated by LLMs (Mistral-Instruct, Vicuna, and Gemini Pro) and obtained through web-crawling. Figure 1 illustrates the different steps of this procedure. The resulting corpora, described in Section 3.1, are compared to a reference ESP corpus associated with the same vocabulary items, handpicked by teachers on the basis of a set of stylistic features introduced in Section 3.2. We hypothesise that the most adequate generation method produces contexts closest to the reference corpus in terms of stylistic characteristics.

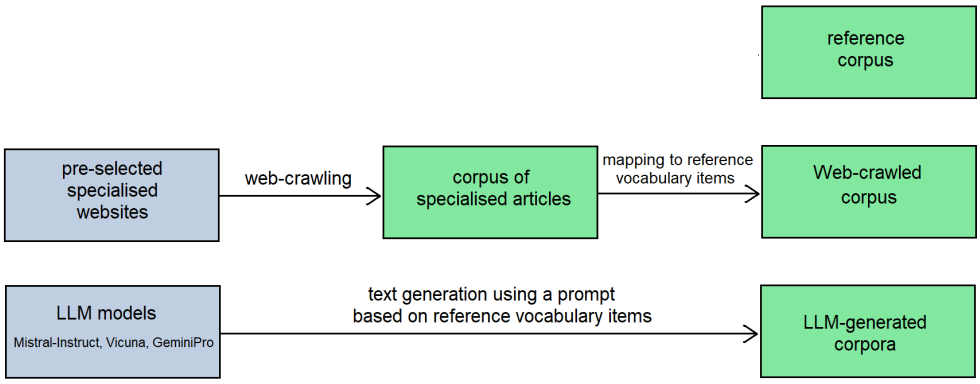


FIGURE 1 – Collection procedure for the examined corpora

3.1 Derivation of the Corpora

3.1.1 Reference

The reference corpus, consisting of 244 contexts, is crafted by ESP teachers from Université catholique de Louvain and consists of a sample of study and examination materials used in the acquisition of vocabulary knowledge in ESP courses. The provided contexts are typically one to three sentences long and reflect key pedagogical qualities as defined by the ESP teachers, including appropriate length, topic relevance, authenticity, timeliness and level-appropriate vocabulary and grammar. The corpus's items are further classified as belonging to CEFR levels B1 or B2 and to the fields of "agronomy" or "general science".

3.1.2 Web-Crawled

The web-crawled context corpus is extracted from a large corpus of articles found in 36 websites (centred around the domains of agronomy, civil engineering and general science), which are commonly

used by ESP teachers when they manually craft context examples [2]. The articles and associated metadata [3] were extracted using the Python tools *beautifulsoup4* [4] and *newspaper* [5] and then preprocessed for noise removal. For the current experiment, sentences in the articles were mapped to the vocabulary items associated with the reference corpus. The recorded metadata was used to ensure compatibility of domains. A set of heuristic rules defined the search for joint matches in terms of words and their POS tags, the latter allowing for lemma rather than word correspondence and resolving cases of polysemy that imply words' parts of speech (e.g. "yield" - both noun and verb). The quality of the issuing corpus was verified manually, and issues that can be fixed automatically (mostly format-related) were identified and resolved. Quality problems that are not readily fixable included text-conversion errors (4 contexts featured mistakes of this type), sentence fragments (a total of 3), and a run-on sentence. Levenstein distance was applied to ensure that no sentences are (closely) identical.

3.1.3 LLM-Generated

To generate contexts based on a prompt, three recent, easily accessible LLMs of different sizes were selected : Mistral-Instruct (7B parameters), Vicuna (13B) and Gemini 1.0 Pro (600B).

Mistral, developed by Mistral AI, makes use of grouped-query and sliding window attention mechanisms, thus substantially increasing inference speed and reducing memory constraints at decoding. Its finetuned "Instruct" version demonstrates superior performance than LLaMA on human as well as automated benchmarks (Jiang *et al.*, 2023). Following a process of trial and error, the prompt's role was set as "system" rather than "user", thus placing focus on the text generation guidelines.

The Vicuna model is based on LLaMA as enhanced via instruction-tuning on data provided by ShareGPT, a platform that features full conversations of users with ChatGPT and facilitates their sharing and reuse (Mehta, 2022). It thus makes use of ChatGPT's established linguistic abilities. Vicuna is associated with better privacy as compared with ChatGPT and can reach as much as 90% of the latter's performance despite its compact size (Lam *et al.*, 2023). In this study, both the Mistral and Vicuna models were used through the "LM studio" interface.

Gemini is a state-of-the-art multimodal model released by Google DeepMind in three versions : Ultra, Pro and Nano. It uses advanced attention mechanisms, such as multi-query attention, and supports a context length of 32k. Its Ultra version achieves higher performance than GPT-4 in 30 out of 32 language benchmarks and uniquely surpasses human performance on the exam benchmark MMLU (Anil *et al.*, 2023). Trained for increased deployability, the Pro version almost matches GPT-3.5 in performance (Akter *et al.*, 2023). Even though the model is proprietary, at the time of writing, it can be accessed freely via the Google AI Studio developer tool within a given quota.

A prompt format was defined and tested that includes the vocabulary item's associated domain and CEFR proficiency level, the part of speech in the case of nouns, verbs, adjectives and adverbs and differentiation between word and expression [6]. Mistral output demonstrated high variance based on its temperature setting [7]. A value of 0.8 was opted for as the threshold below which output was perceivably too homogeneous and consisted of definitions of the target vocabulary rather than

2. See Appendix 1 : List of Crawled Websites for the list of utilised websites
3. title, date, scientific domain, format (html vs pdf)
4. Version 4.12.3 ; https ://pypi.org/project/beautifulsoup4/
5. Version 0.2.8 ; https ://pypi.org/project/newspaper3k/
6. See Appendix 2 : Prompts used for LLM Generation for the utilised prompts
7. a model's temperature defines its level of unexpectedness or creativity

examples of use.

An additional experiment was carried out using one of the models (Gemini was opted for due to its fastest performance), namely a one-shot setting in which we offer the corresponding example from the reference corpus. The purpose was to test the LLM's efficiency in adapting its output to the given example and the underlying potential for multiple contexts per target word to be derived from a single professionally-crafted one.

The models were asked to provide output until it was automatically verified that output was present and that it contained the respective vocabulary items (verbatim or, in the case of verbs and nouns, in any possible form). Readily fixable issues (such as an additional "Explanation" part) were addressed manually, and no more substantial issue was found. The Gemini model exhibited by far the fastest performance (650 seconds, compared with 2624 for Mistral-Instruct and 5447 for Vicuna). Once again, it was ensured that no sentences were (closely) identical [8].

3.2 Stylistic Comparison

For the stylistic analysis of the various contexts, we selected various atomic features related to textual readability belonging to four general categories : length-based, morphosyntactic, lexico-semantic, and discourse-related [9].

The selected length-based features are the numbers of words and syllables per sentence and numbers of letters and syllables per word. Morphosyntactic features include the number of noun phrases per sentence, the number of non-stem words per sentence, the number of punctuation signs per sentence (excluding end-of-sentence punctuation), the percentage of sentences ending in question and exclamation marks, and the overall morphological richness as defined by Cech & Kubat (2018). The selected lexico-semantic features are the number of verbs, first-person pronouns, proper nouns and the joint number of adjectives and adverbs per sentence. We also considered the word-based and lemma-based type-to-token ratios, the percentage of hapax legomena, the percentage of words not present in the Dale-Chall list, the average concreteness (based on Brysbaert *et al.* (2014)'s rated concreteness list of 40k English lemmas), and the 10 most frequent words including and excluding stop words. Finally, the discourse-related features consist of the number of pronouns per sentence, the percentage of anaphora-denoting words per sentence [10], and the cosine distance between all sentences in the respective corpus [11]. The assignment of features to a particular category is occasionally highly subjective ; for instance, the number of verbs in a sentence could be interpreted as being more strongly related to a sentence's syntactic structure than its semantics.

In their work on readability classification, Wilkens *et al.* (2022) concluded that the use of a set of aggregators provides a better estimation of textual qualities than a single selected value. Where relevant, the average, minimal and maximal values for a feature, as well as the standard deviation (SD), were examined. This allowed for comparisons of both the texts' general qualities (as often relevant to a measure of complexity) and the span of values contained (which can serve as an estimation of

8. Our experimental setup featured an 11th Gen Intel Core i7 CPU with 8 cores and TigerLake-LP GT2 integrated GPU.
9. Measures that strongly imply a larger textual unit (e.g. textual cohesion) were naturally excluded.
10. The words considered are the following : definite article (*the*) ; personal pronouns (*he*, *she*, *it*, *they*) ; demonstrative pronouns (*this*, *that*, *these*, *those*) ; relative pronouns (*who*, *which*, *whose*, *whom*, *where*) ; indefinite pronouns (*all*, *some*, *none*, *any*, *each*, *every*) ; adverbs (*here*, *there*, *now*, *then*)
11. calculated using Python's *transformers* library ; sentence transformer model *paraphrase-MiniLM-L6-v2*

textual variety). [12]

The following steps were taken to evaluate continuous features. Firstly, a Shapiro-Wilk test (Shapiro & Wilk, 1965) was used to determine whether the features demonstrate normal distribution. As the only feature that was normally distributed was the percentage of non-stem words per sentence, we used Mann-Whitney U, a non-parametric test, to determine whether differences between the reference corpus and the rest of the corpora were significant. Statistical significance, when present, was assigned one of three levels corresponding to p-values of 0.001, 0.01 and 0.05.

4 Results

The results of our stylistic comparison are reported in Table 1. Only the most relevant features have been listed ; please refer to Appendix 4 : Detailed Results for a comparison of all features. This section provides an overview of the main results, organised according to the four families of features.

Feature	Ref.	Web	Mistral	Vicuna	Gemini	Gemini : one-shot
words in sample	9823	7269	4615	4852	4160	**10366**
words / sentence	*13.59*	*14.93****	*9.34***	*9.6***	*8.51****	***12.26******
letters / word	*5.29*	*5.38*	*5.3*	*5.43*	*5.6**	***5.53***
noun phrases / sentence	*5.87*	*8.16****	*5.56*	*5.53*	*4.92****	***5.39****
non-stem words / s-ce	*33.56*	*31.56****	*35.14****	*35.59****	*36.36****	***36******
punctuation signs / s-ce	*1.56*	*2.7*	*0.77****	*0.99****	*0.75****	***1.25****
verbs / sentence	*2.45*	*3.83*	*2.54****	*2.44****	*2.27****	***2.53***
adj. and adv. / sentence	*2.96*	*4.13****	*2.21****	*2.31****	*2.29****	***2.52*****
1st-person pron. / s-ce	*0.1*	*0.12*	*0.39****	*0.11*	*0.06***	***0.07****
proper nouns / sentence	*0.9*	*1.46*	*0.06****	*0.32****	*0.1****	***0.27******
hapax legomena	16.13	22.56	16.25	18.18	19.75	**14.14**
concreteness	2.46	2.36	2.44	2.44	2.42	**2.41**
pronouns / sentence	*0.88*	*1.27*	*1.06*	*0.8*	*0.5****	***0.73****
anaphora words / s-ce	*20.49*	*10.78*	*10.47*	*10.43*	*13.42****	***24.59***
cos. distance btwn s-ces	*0.14*	*0.1****	*0.18****	*0.17****	*0.18****	***0.15*****

TABLE 1 – Comparison of the corpora based on a sample of textual features. The average values of continuous characteristics are indicated in *italics*, and the statistical significance of their divergence from the reference corpus is marked with * (lowest), ** and *** (highest). The one-shot Gemini corpus is represented in **bold** to denote its highest global closeness to the reference as per Section 4.5.

4.1 Length-Based Features

The one-shot corpus contains the largest number of words, closely followed by the reference and web-crawled ones and then by the three zero-shot LLM-generated corpora, which exhibit similar values. The number of words per sentence, tightly associated with textual complexity, is highest in the web-crawled corpus, followed by the reference one and then by all LLM corpora. Differences at the

12. Refer to Appendix 3 : Features Used in Corpus Comparison for an overview of all investigated features.

"word" level (e.g. the number of letters per word) are minimal. The ranges of length-based features [13] are lowest with the LLM corpora (implying a lack of variety) and highest with the web-crawled one, followed closely by the reference corpus. The web-crawled corpus tends to demonstrate the largest SD. In this category, the Vicuna and Mistral corpora demonstrate the lowest significance in difference with the reference.

4.2 Morphosyntactic Features

The number of noun phrases per sentence is similar between the reference corpus and the LLM ones and significantly higher within the web-crawled corpus. The number of non-stem words, which is the most statistically different feature in the category, is lowest in the web-crawled corpus (interestingly suggesting lower complexity) and highest in the LLM-generated ones, the reference corpus standing in the middle. In relation to the number of punctuation signs, the reference corpus is once again in the middle, this time the web-crawled corpus exhibiting the highest value. The LLM corpora do not demonstrate variety in end-of-sentence punctuation. Morphological richness is stable at 0.02 for all corpora. Once again, the web-crawled corpus has the highest SD and value ranges are typically narrower for LLM corpora.

4.3 Lexico-Semantic Features

The number of verbs per sentence (associated with the presence of complex sentences) is closely stable, with the exception of the web-crawled corpus, where it is significantly higher. The number of adjectives and adverbs (which demonstrates the highest statistical difference in the category), as well as the number of proper nouns per sentence, are highest within the web-crawled corpus and lowest within the LLM ones, the reference corpus standing in the middle. The most hapax legomena are found in the web-crawled corpus and the fewest in the LLM ones (in particular, the one-shot corpus), implying re-use of vocabulary. First-person pronouns are generally rare, the highest value of 0.39 per sentence being associated with the Mistral corpus. The percentage of words outside of the Dale-Chall frequency list is highly stable, and so are the average concreteness of words and type-to-token ratios. Once again, the web-crawled corpus demonstrates the highest value ranges and SD. With the exception of the web-crawled corpus, the most frequent words excluding stop words are narrowly related to the texts's specialisation (e.g. "water", "climate"). When stop words are retained, the words are highly identical, the Mistral corpus uniquely featuring the pronoun "I".

4.4 Discourse-Related Features

Cosine distance (i.e. the estimated semantic difference between examples) diverges the most from the reference corpus. While average cosine distance demonstrates similar values, the maximal one is highest with the LLM corpora. The number of pronouns per sentence is highest in the web-crawled corpus and varies among the LLM ones. Vicuna's average value is closest to the reference but at the expense of a significant difference in distribution. Anaphora-denoting words are most prominent in the one-shot followed by the reference corpus, values being significantly lower in the other corpora. Uniquely for this category, the web-crawled corpus does not exhibit high SD values.

13. i.e. the differences between their maximal and minimal values

4.5 Additional experiments

In this subsection, we report three additional experiments that we carried out for the purpose of gaining deeper insight about our corpora.

First, we divided each corpus into two according to the proficiency level of the target vocabulary items (B1 or B2). In such a scenario, the web-crawled corpus, which consists of authentic texts not specifically conceived for language learners, increases in significance of difference with the reference one ; in particular, in relation to level B2, for which it exhibits 12 significantly different features (as opposed to 5 when the entire corpus is considered).

In contrast, sensitivity in relation to proficiency levels is noticeable among the LLM corpora. Features associated with textual complexity, such as the total number of words, the number of letters per word and the number of first-person pronouns per sentence, demonstrate significantly differing values compared to when the corpus is taken in its entirety. In particular, the Gemini corpus shows the highest modification of values based on proficiency level. Interestingly, there are even cases where LLM corpora show higher sensitivity to the level at hand than the reference corpus [14]. As the different CEFR levels are also associated with different domains, the most frequent words in the LLM corpora now reflect the domain at hand (derivatives of "science" being common for the scientific domain and words such as "crop" and "soil" for agronomy).

As a second experiment, the complete set of textual characteristics was used within a global distance metric in order to determine which corpus is globally closest to the reference one. For this purpose, min-max normalisation was applied, and the Euclidean distance between corpora was calculated. The one-shot Gemini corpus ranked first (2.96), followed by the web-crawled one (3.8) and the three zero-shot LLM corpora, which in turn demonstrated relatively similar values [15].

Finally, as regards the one-shot scenario, the features of the corresponding corpus were compared to those of its zero-shot counterpart, revealing that their majority [16] come closer to the reference, reducing the significance in divergence in 10 out of all 13 cases. The one-shot generation method also leads to significantly increased global closeness to the reference.

5 Discussion

Given the high value ranges and SD pertaining to the web-crawled corpus, as well as the tendency between its and LLM-generated texts to diverge from the baseline in opposite directions, the two types of corpora can work together effectively within an educational framework to provide a variety of contexts for ESP vocabulary. The web-crawling method is computationally efficient and tends to provide texts that are naturally close to the reference examples. In turn, the benefits of LLM generation include sensitivity to the CEFR level at hand and high malleability, including the potential to derive a large number of examples from a single one within a one-shot setting.

Table 2 shows the example sentences for the verb "to avoid" in all discussed corpora. In accordance

14. For instance, when solely B1 examples are considered, the reference corpus unintuitively has higher values for the percentages of non-stem words and words outside the Dale-Chall frequency list compared to when the entire corpus is considered.

15. Mistral : 4.03 ; Vicuna : 4.06 ; Gemini : 4.57

16. excluding "percentage of hapax legomena", "average concreteness" and several standard deviation values

with the results stated in Section 4, the web-crawled example exhibits high complexity : it is longest and contains three proper nouns and a direct quotation. The Vicuna and Gemini examples show similarity with the reference if one considers the number of verbs and the joint number of adjectives and adverbs [17]. Mistral adds variety with its use of first-person language, which however results in the example's reduced formality and in-domain quality. In turn, Gemini's example is narrowly associated with the scientific domain. Much akin to the reference, the "Gemini : one-shot" example features additional vocabulary items that are relevant for the same learner audience ("essential", "biases", "skew"), yet it does not demonstrate any perceivable copying of the reference content.

Reference	There is still time to reverse the warming trend and **avoid** global environmental and economic catastrophe.
Web-Crawled	"There are few institutional structures to achieve co-operation globally on the sort of scales now essential to **avoid** very serious consequences," warns lead author Dr Brian Walker of Australia's CSIRO.
Mistral	I try to **avoid** using my phone while I study because it can be a great distraction.
Vicuna	To **avoid** overfishing, it is essential to manage fisheries sustainably and establish marine protected areas.
Gemini	To **avoid** contamination, the scientist carefully wore gloves and a lab coat while conducting the experiment.
Gemini : one-shot	In scientific experiments, it is essential to **avoid** biases that could skew the results.

TABLE 2 – Examples for the vocabulary item "avoid" (domain "science" ; level "B1")

6 Conclusion and Future Directions

This work discussed context corpora derived via two discrete NLP-based methods (web-crawling and generation by LLMs), which can be used to aid university students in the acquisition of ESP vocabulary. The stylistic characteristics of the generated contexts were compared to a professionally crafted baseline of context examples through quantitative evaluation based on readability features. The comparison revealed higher similarity among different LLM-generated corpora than between them and corpora of a different nature. The "Gemini : one-shot" corpus was discovered to be globally closest to the reference, thereby showing that generation can be refined through prompt engineering. Future experiments including few-shot use of only several high-quality examples independent of a vocabulary list can help mitigate the current limitation of reliance on a reference corpus.

Associated future plans include an evaluation of the pedagogic characteristics of contexts as opposed to their readability-based qualities. In more practical terms, pre- and post-tests of students' performance will be conducted in relation to the introduction of a large inventory of NLP-derived examples in the implied university courses.

17. respectively, 3 and 3 for the reference, 4 and 4 for Vicuna and 3 and 1 for Gemini

References

AKTER S. N., YU Z., MUHAMED A., OU T., BÄUERLE A., CABRERA A. A., DHOLAKIA K., XIONG C. & NEUBIG G. (2023). An in-depth look at Gemini's language abilities. arXiv : 2312.11444.

ANIL R., BORGEAUD S., WU Y., ALAYRAC J.-B., YU J., SORICUT R., SCHALKWYK J., DAI A. M., HAUTH A., MILLICAN K., SILVER D., PETROV S., JOHNSON M., ANTONOGLOU I., SCHRITTWIESER J., GLAESE A., CHEN J., PITLER E. & VINYALS O. (2023). Gemini : A family of highly capable multimodal models. arXiv : 2312.11805.

BRYSBAERT M., WARRINER A. B. & KUPERMAN V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, **46**(3), 904–911. DOI : 10.3758/s13428-013-0403-5.

CECH R. & KUBAT M. (2018). Morphological richness of text. In M. FIDLER & V. CVRCEK, Éds., *Taming the Corpus*, p. 63–77. Springer. DOI : 10.1007/978-3-319-98017-1_4.

COLLINS-THOMPSON K. (2014). Computational assessment of text readability : A survey of current and future research. *International Journal of Applied Linguistics*, **165**(2), 97–135.

DEUTSCH T., JASBI M. & SHIEBER S. (2020). Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* : Association for Computational Linguistics. DOI : 10.18653/v1/2020.bea-1.1.

DUBAY W. H. (2007). *The Classic Readability Studies*. Rapport interne, ERIC Clearinghouse. DOI : 10.1109/TPC.2008.2007872.

FERRAND L. (2007). *Psychologie cognitive de la lecture*. Bruxelles : De Boeck.

FRANÇOIS T. (2011). La lisibilité computationnelle : un renouveau pour la lisibilité du français langue première et seconde ? *International Journal of Applied Linguistics (ITL)*, **160**, 75–99.

GODWIN-JONES R. (2018). Evolving views on vocabulary development. *Language Learning & Technology*, **22**(3), 1–19.

HEILMAN M., ZHAO L., PINO J. & ESKENAZI M. (2008). Retrieval of reading materials for vocabulary and reading practice. In J. TETREAULT, J. BURSTEIN & R. DE FELICE, Éds., *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, p. 80–88, Columbus, Ohio : Association for Computational Linguistics.

HUCKIN T. & COADY J. (1999). Incidental vocabulary acquisition in a second language : A review. *Studies in second language acquisition*, **21**(2), 181–193.

HUSSIN A., CHAN Y. F. & ZUBAIDAH A. (2010). Scientific structural changes within texts of adapted reading materials. *English Language Teaching*, **3**. DOI : 10.5539/elt.v3n4p216.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., CASAS D. D. L. & EL SAYED W. (2023). Mistral 7B. arXiv : 2310.06825.

JIN T. & LU X. (2018). A data-driven approach to text adaptation in teaching material preparation : Design, implementation, and teacher professional development. *TESOL Quarterly*, **52**, 457–467. DOI : 10.1002/tesq.434.

KYLE K. (2019). Measuring lexical richness. In S. WEBB, Éd., *The Routledge Handbook of Vocabulary Studies*, p. 454–476. Routledge. DOI : 10.4324/9780429291586.

LAM K.-Y., CHENG V. C. W. & YEONG Z. K. (2023). Applying large language models for enhancing contract drafting. In *Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace*, p. 70–80.

LITMAN D. J. (2016). Natural language processing for enhancing teaching and learning. In *AAAI Conference on Artificial Intelligence*.

MEHTA I. (2022). ShareGPT lets you easily share your ChatGPT conversations. Blog post.

MEURERS D. (2021). *Natural Language Processing and Language Learning*, In *The Encyclopedia of Applied Linguistics*, p. 1–15. John Wiley Sons, Ltd. DOI : https ://doi.org/10.1002/9781405198431.wbeal0858.pub2.

MEURERS D., ZIAI R., AMARAL L., BOYD A., DIMITROV A., METCALF V. & OTT N. (2010). Enhancing authentic web pages for language learners. In J. TETREAULT, J. BURSTEIN & C. LEACOCK, Éds., *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 10–18, Los Angeles, California : Association for Computational Linguistics.

RAMOS R. & DARIO F. (2015). Incidental vocabulary learning in second language acquisition : A literature review. *Profile Issues in TeachersProfessional Development*, **17**(1), 157–166.

SHAIKH S., YAYILGAN S. Y., KLIMOVA B. & PIKHART M. (2023). Assessing the usability of ChatGPT for formal english language learning. *European Journal of Investigative Health Psychology and Education*, **13**, 1937–1960. DOI : 10.3390/ejihpe13090140.

SHAPIRO S. S. & WILK M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**(3/4), 591–611. DOI : 10.2307/2333709.

URE J. (1971). Lexical density and register differentiation. *Applications of linguistics*, **23**(7), 443–452.

VAJJALA S. (2022). Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 5366–5377.

WILKENS R., ALFTER D., WANG X., PINTARD A., TACK A., YANCEY K. P. & FRANÇOIS T. (2022). FABRA : French aggregator-based readability assessment toolkit. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 1217–1233, Marseille, France : European Language Resources Association.

WILKENS R., WATRIN P., CARDON R., PINTARD A., GRIBOMONT I. & FRANÇOIS T. (2024). Exploring hybrid approaches to readability : experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics : EACL 2024*, p. 2316–2331.

WILSON R. (2004). Computers and the internet : Together a great tool for esl/efl learners.

YOON S.-Y., LEE C. M., HOUGHTON P., LOPEZ M., SAKANO J., LOUKINA A., KROVETZ B., LU C. & MADNANI N. (2017). Analyzing item generation with natural language processing tools for the toeic ® listening test : Analyzing item generation with nlp tools. *ETS Research Report Series*, **2017**. DOI : 10.1002/ets2.12183.

YOUNG J. C. & SHISHIDO M. (2023a). Evaluation of the potential usage of ChatGPT for providing easier reading materials for ESL students. In *Proceedings of EdMedia and Innovate Learning*, p. 155–162 : Association for the Advancement of Computing in Education (AACE).

YOUNG J. C. & SHISHIDO M. (2023b). Investigating OpenAI's ChatGPT potentials in generating chatbot's dialogue for English as a foreign language learning. *International Journal of Advanced Computer Science and Applications*, **14**(6), West Yorkshire. DOI : 10.14569/IJACSA.2023.0140607.

YULE G. U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

Appendix 1 : List of Crawled Websites

https://climate.ec.europa.eu/climate-change_en
https://climate.ec.europa.eu/eu-action_en
https://climate.ec.europa.eu/index_en
https://climate.nasa.gov/
https://engineeringdiscoveries.com/
https://newatlas.com
https://sciencedemonstrations.fas.harvard.edu/
https://sustainability.stanford.edu/
https://world-nuclear.org
https://www.advancedsciencenews.com
https://www.computerworld.com/
https://www.eurekalert.org/
https://www.green.earth/
https://www.iea.org/
https://www.ipcc.ch
https://www.livescience.com/
https://www.nationalgeographic.org/society/
https://www.nature.com/
https://www.ncbi.nlm.nih.gov/
https://www.networkworld.com/
https://www.newscientist.com/
https://www.npr.org/sections/science/
https://www.pcworld.com
https://www.pewresearch.org/topic/internet-technology/
https://www.pewresearch.org/topic/science/
https://www.popularmechanics.com/
https://www.science.org/
https://www.sciencealert.com/
https://www.sciencedaily.com/
https://www.scienceopen.com/
https://www.scientificamerican.com
https://www.triplepundit.com/
https://www.un.org/en/
https://www.un.org/en/climatechange
https://www.usgs.gov/programs/earthquake-hazards/
https://www.wwf.org.uk

Appendix 2 : Prompts used for LLM Generation

Zero-shot setting for Mistral, Vicuna, and Gemini :

> Here is a sentence[a] at CEFR level {***level***} showing how you use the **{*pos* if verb/noun/adverb/adjective; else 'word' or 'expression'}** "**{*item*}**" ({***domain***}) :

a. The reason for 'sentence' to be used rather than 'example', even though some of the reference examples consist of more than a single sentence, is that using 'example' tends to result in the rendition of extensive explanations instead of or in addition to an example of use. This problem does not persist with the one-shot setting, for which therefore the word 'example' is used instead.

One-shot setting for Gemini :

> Please provide an example (between {***lower***}[a] and {***upper***} words at CEFR level {***level***} showing how you use the **{*pos* if verb/noun/adverb/adjective; else 'word' or 'expression'}** "**{*item*}**" ({***domain***}) :
>
> Example : {**reference_example**}

a. 'Lower' and 'upper' denote a range of example lengths, which differs for the different CEFR levels (8 to 43 words for B1 and 20 to 87 words for B2). The ranges are defined as +/- 1.5 standard deviations from the average value per level. This value as well as the addition of information about length itself was decided upon following a process of trial and error based on the behaviour of 20 sample examples in comparison to the reference's counterparts.

Appendix 3 : Features Used in Corpus Comparison

Length-Based	total number of examples in the sample total number of words in the sample average/min/max/SD number of words per sentence average/min/max/SD number of syllables per sentence average/min/max/SD number of letters per word average/min/max/SD number of syllables per word
Morphosyntactic	average/min/max/SD number of noun phrases per sentence average/min/max/SD number of non-stem words per s-ce percentage of sentences ending in question mark percentage of sentences ending in exclamation mark average/min/max/SD number of punctuation signs per s-ce morphological richness
Lexico-Semantic	average/min/max/SD number of verbs per sentence average/min/max/SD number of adj. and adv. per s-ce average/min/max/SD number of 1st-person pronouns per s-ce average/min/max/SD number of proper nouns per sentence percentage of words not present in the Dale-Chall list percentage of hapax legomena type-to-token ratio (word-based) type-to-token ratio (lemma-based) average concreteness 10 most frequent words (excluding stop words) 10 most frequent words (including stop words)
Discourse-Related	average/min/max/SD number of pronouns per sentence average/min/max/SD cosine distance between sentences average/min/max/SD % of anaphora-denoting words per sentence

Appendix 4 : Detailed Results

Entire Sample

Feature	Reference	Web-Crawled	Mistral	Vicuna	Gemini	Gemini : one-shot
Total # examples in sample	244	244	244	244	244	**244**
Total # words in sample	9823	7269	4615	4852	4160	**10366**
Avg. # words / s-ce	*13.59*	*14.93**[l]*	*9.34***	*9.6***	*8.51****	***12.26****
Min. # words / s-ce	3	7	4	8	8	**5**
Max. # words / s-ce	58	69	36	30	36	**44**
SD # words / s-ce	8.59	11.42	4.84	5.04	4.61	**5.77**
Avg. # syllables / s-ce	*21.52*	*24.47****	*15.33*	*15.93*	*14.65**	***20.47***
Min. # syllables / s-ce	2	12	4	10	14	**8**
Max. # syllables / s-ce	93	120	63	68	57	**84**
SD # syllables / s-ce	14.82	20.5	8.77	9.51	8.61	**10.97**
Avg. # letters / word	*5.29*	*5.38*	*5.3*	*5.43*	*5.6**	***5.53***
Min. # letters / word	1	1	1	1	1	**1**
Max. # letters / word	21	23	15	17	17	**20**
SD # letters / word	2.89	3.01	2.94	3.02	3.07	**3.01**
Avg. # syllables / word	*1.58*	*1.64**	*1.64****	*1.66****	*1.72****	***1.67****
Min. # syllables / word	0	0	1	1	1	**0**
Max. # syllables / word	7	8	6	6	6	**6**
SD # syllables / word	0.91	0.99	0.94	0.96	1	**0.96**

Avg. # noun phrases / s-ce	*5.87*	*8.16****	*5.56*	*5.53*	*4.92****	***5.39****
Min. # noun phrases / s-ce	0	2	1	2	2	**1**
Max. # noun phrases / s-ce	18	19	10	10	11	**14**
SD # noun phrases / s-ce	2.71	3.4	1.59	1.68	1.49	**1.9**
Avg. % non-stem words / s-ce	*33.56*	*31.56****	*35.14****	*35.59****	*36.36****	***36****
Min. % non-stem words / s-ce	0	0	11.76	11.11	9.1	**6.25**
Max. % non-stem words / s-ce	70	56.25	65	70	64.29	**83.33**
SD % non-stem words / s-ce	11.17	9.18	9.82	10.07	9.95	**11.15**
% s-ces ending in "?"	0.63	1.62	0	0	0	**0**
% s-ces ending in "!"	0.42	0	0.4	0	0	**0**
Avg. # punct. signs / s-ce	*1.56*	*2.7*	*0.77****	*0.99****	*0.75****	***1.25****
Min. # punct. signs / s-ce	0	0	0	0	0	**0**
Max. # punct. signs / s-ce	6	7	4	3	4	**6**
SD # punct. signs / s-ce	0.55	0.77	0.32	0.35	0.32	**0.48**
Morphological richness	0.02	0.02	0.02	0.02	0.02	**0.02**
Avg. # verbs / s-ce	*2.45*	*3.83*	*2.54****	*2.44****	*2.27****	***2.53***
Min. # verbs / s-ce	0	0	0	0	0	**0**
Max. # verbs / s-ce	8	11	7	7	5	**8**
SD # verbs / s-ce	1.53	1.84	1.08	1.26	1.06	**1.14**
Avg. # adj. and adv. / s-ce	*2.96*	*4.13****	*2.21****	*2.31****	*2.29****	***2.52****
Min. # adj. and adv. / s-ce	0	0	0	0	0	**0**

Max. # adj. and adv. / s-ce	10	14	10	7	8	**9**
SD # adj. and adv. / s-ce	1.97	2.62	1.48	1.46	1.41	**1.56**
Avg. # 1st-person pron. / s-ce	*0.1*	*0.12*	*0.39***	*0.11*	*0.06***	***0.07****
Min. # 1st-person pron. / s-ce	0	0	0	0	0	**0**
Max. # 1st-person pron. / s-ce	2	4	4	2	3	**4**
SD # 1st-person pron. / s-ce	0.38	0.41	0.75	0.38	0.3	**0.36**
Avg. # proper nouns / s-ce	*0.9*	*1.46*	*0.06****	*0.32****	*0.1****	***0.27*****
Min. # proper nouns / s-ce	0	0	0	0	0	**0**
Max. # proper nouns / s-ce	14	12	3	15	7	**10**
SD # proper nouns / s-ce	1.84	2.44	0.31	1.56	0.54	**0.94**
% words not in Dale-Chall list	45.21	45.38	42.25	45.22	48.34	**47.41**
% hapax legomena	16.13	22.56	16.25	18.18	19.75	**14.14**
Type-to-token ratio (words)	0.26	0.33	0.27	0.29	0.3	**0.24**
Type-to-token ratio (lemmas)	0.25	0.31	0.25	0.27	0.29	**0.23**
Average concreteness	2.46	2.36	2.44	2.44	2.42	**2.41**
10 most frequent words (excl. stop words)	water, climate, change, species, world, people, new, plants, global, could	would, could, said, people, water, also, international, may, must, new	crop, soil, crops, farmers, scientists, agriculture, water, yields, new, order	soil, crop, farmers, agriculture, crops, yields, practices, water, agricultural, climate	crop, soil, new, scientists, farmers, crops, practices, yields, scientist, sustainable	**water, soil, farmers, crops, crop, plant, species, food, practices, yields**
10 most frequent words (incl. stop words)	the, of, and, to, in, a, is, that, are, for	the, of, and, to, in, a, that, for, be, is	the, to, of, and, in, a, that, I, for, can	to, the, of, and, in, a, that, is, can, as the, of, to, a, and, in, for, crop, soil	the, of, to, a, and, in, for, crop, soil	**the, of, to, and, in, a, for, is, that, are**
Avg pron. / s-ce	*0.88*	*1.27*	*1.06***	*0.8***	*0.5****	***0.73****
Min. pron. / s-ce	0	0	0	0	0	**0**
Max. pron. / s-ce	5	6	5	6	4	**4**

SD pron. / s-ce	1.03	1.33	1.1	0.92	0.78	**0.93**
Avg. % anaphora words / s-ce	*20.49*	*10.78*	*10.47*	*10.43*	*13.42****	***24.59***
Min. % anaphora words / s-ce	0	0	0	0	0	**0**
Max. % anaphora words / s-ce	42.86	30	37.5	30.77	30.77	**31.25**
SD % anaphora words / s-ce	6.55	5.64	6.97	6.53	6.93	**6.8**
Avg. cos. distance btwn s-ces	*0.14*	*0.1****	*0.18****	*0.17****	*0.18****	***0.15****
Min. cos. distance btwn s-ces	-0.25	-0.25	-0.25	-0.32	-0.26	**-0.27**
Max. cos. distance btwn s-ces	0.72	0.78	0.85	0.9	0.87	**0.89**
SD cos. distance btwn s-ces	0.13	0.12	0.17	0.16	0.16	**0.15**

Comparison between the investigated corpora based on a sample of textual features. Features in *italics* have been tested for statistical significance, and the extent of the significance is marked with *, ** and *** from lowest to highest. The one-shot Gemini corpus is marked with **bold** to denote its highest global similarity to the reference corpus.

Per Level : B1 (domain "Science")

Feature	Reference	Web-Crawled	Mistral	Vicuna	Gemini	Combined LLM
Total # examples in sample	132	132	132	132	132	132
Total # words in sample	3402	3987	2358	2421	1976	2824
Avg. # words / s-ce	*11.01*	*15.1****	*8.8**	*9.03*	*7.46****	*9****
Min. # words / s-ce	3	7	9	8	8	5
Max. # words / s-ce	42	69	31	20	33	38
SD # words / s-ce	7.52	12.03	4.44	5.27	3.51	5.33
Avg. # syllables / s-ce	*17.81*	*24.62****	*14.66*	*14.93*	*12.72**	*14.71****
Min. # syllables / s-ce	2	12	10	11	14	8
Max. # syllables / s-ce	83	120	63	56	48	65
SD # syllables / s-ce	13.93	20.85	8.37	9.65	6.39	9.78
Avg. # letters / word	*3.56*	*5.35*	*5.27*	*5.35*	*5.45*	*5.35*
Min. # letters / word	1	1	1	1	1	1
Max. # letters / word	20	19	15	16	17	20
SD # letters / word	2.95	3	3.01	2.99	3.05	3
Avg. # syllables / word	*1.62*	*1.63*	*1.67*	*1.65*	*1.71**	*1.64**
Min. # syllables / word	0	0	1	1	1	0
Max. # syllables / word	7	8	6	6	6	6
SD # syllables / word	0.95	0.97	0.96	0.95	1	0.95

Avg. # noun phrases / s-ce	*5.51*	*8.01***	*5.26*	*5.3*	*4.38****	*5.14***
Min. # noun phrases / s-ce	0	2	2	2	2	1
Max. # noun phrases / s-ce	14	19	9	10	8	11
SD # noun phrases / s-ce	2.49	3.5	1.38	1.74	1.19	1.87
Avg. % non-stem words / s-ce	*34.3*	*31.01*	*34.37****	*34.41****	*34.72****	*34.1*
Min. % non-stem words / s-ce	0	0	11.76	11.11	9.09	8.33
Max. % non-stem words / s-ce	61.54	56.25	61.54	70	57.14	55
SD % non-stem words / s-ce	10.61	9.63	10.1	10.31	9.35	9.95
% s-ces ending in "?"	0	2.22	0	0	0	0
% s-ces ending in "!"	1.14	0	0	0	0	0
Avg. # punct. signs / s-ce	*1.33*	*2.71****	*0.63****	*0.88***	*0.46****	*0.91****
Min. # punct. signs / s-ce	0	0	0	0	0	0
Max. # punct. signs / s-ce	6	7	3	3	3	6
SD # punct. signs / s-ce	0.54	0.79	0.28	0.33	0.24	0.4
Morphological richness	0.02	0.02	0.02	0.02	0.02	0.02
Avg. # verbs / s-ce	*2.17*	*3.84****	*2.41*	*2.35*	*2.03****	*2.21*
Min. # verbs / s-ce	0	0	0	0	0	0
Max. # verbs / s-ce	7	11	7	7	4	7
SD # verbs / s-ce	1.39	1.95	1.03	1.25	0.86	1.23
Avg. # adj. and adv. / s-ce	*2.78*	*4.09*	*2.04****	*2.13****	*1.97****	*2.07****
Min. # adj. and adv. / s-ce	0	0	0	0	0	0

Max. # adj. and adv. / s-ce	9	14	10	7	6	6
SD # adj. and adv. / s-ce	1.81	2.67	1.45	1.43	1.22	1.32
Avg. # 1st-person pron. / s-ce	*0.12*	*0.16*	*0.55****	*0.14*	*0.1*	*0.08*
Min. # 1st-person pron. / s-ce	0	0	0	0	0	0
Max. # 1st-person pron. / s-ce	2	4	4	2	3	2
SD # 1st-person pron. / s-ce	0.42	0.51	0.88	0.41	0.39	0.38
Avg. # proper nouns / s-ce	*0.71*	*1.64**	*0.1****	*0.36****	*0.11****	*0.2****
Min. # proper nouns / s-ce	0	0	0	0	0	0
Max. # proper nouns / s-ce	7	12	3	15	2	5
SD # proper nouns / s-ce	1.34	2.72	0.38	1.45	0.36	0.69
% words not in Dale-Chall list	46.06	45.78	41.65	44.4	46.51	45.56
% hapax legomena	24.06	29.82	23.3	26.57	28.11	26.84
Type-to-token ratio (words)	0.36	0.4	0.35	0.38	0.39	0.38
Type-to-token ratio (lemmas)	0.34	0.38	0.33	0.35	0.37	0.36
Average concreteness	2.47	2.39	2.38	2.41	2.37	2.41
10 most frequent words (excl. stop words)	climate, change, water, people, areas , earthquake, greenhouse, earth, world, atmosphere	would, could, time, may, said, water, help, new, using, must	scientists, science, study, new, experiment, behavior, scientific, research, climate, understand	scientists, scientific, research, study, field, climate, science, discovery, certain, understand	scientists, new, scientist, experiment, earth, research, data, chemical, scientific, study	water, experiment, scientists, chemical, new, used, species, different, earch, causing
10 most frequent words (incl. stop words)	the, of, and, in, to, a, is, are, as, can	the, of, to, and, in, a, that, for, be, is	the, of, to, in, a, and, I, that, scientists, is	to, the, and, in, of, a, is, as, that, it	the, of, to, a, scientists, in, and, new, for, that	the, of, to, in, a, and, is, that, are
Avg pron. / s-ce	*0.67*	*1.27***	*1.25****	*0.77*	*0.47**	*0.65**
Min. pron. / s-ce	0	0	0	0	0	0
Max. pron. / s-ce	5	6	5	4	3	4

SD pron. / s-ce	0.99	1.39	1.11	0.91	0.69	0.91
Avg. % anaphora words / s-ce	*12.95*	*10.63*	*11.7**	*11.41*	*15.9****	*16.53**
Min. % anaphora words / s-ce	0	0	0	0	0	0
Max. % anaphora words / s-ce	27.27	30	37.5	30.77	30.77	31.25
SD % anaphora words / s-ce	6.41	5.73	6.82	6.76	7.05	6.59
Avg. cos. distance btwn s-ces	*0.13*	*0.09****	*0.14****	*0.13*	*0.16****	*0.11****
Min. cos. distance btwn s-ces	-0.22	-0.25	-0.25	-0.25	-0.26	-0.27
Max. cos. distance btwn s-ces	0.71	0.64	0.84	0.8	0.87	0.78
SD cos. distance btwn s-ces	0.14	0.11	0.14	0.13	0.14	0.11

Per Level : B2 (domain "Agronomy")

Feature	Reference	Web-Crawled	Mistral	Vicuna	Gemini	Gemini : one-shot
Total # examples in sample	112	112	112	112	112	112
Total # words in sample	6421	3282	2257	2431	2184	7542
Avg. # words / s-ce	*15.51*	*14.72****	*9.99*	*10.26*	*9.75*	*14.21****
Min. # words / s-ce	3	11	4	9	11	7
Max. # words / s-ce	58	64	36	30	36	44
SD # words / s-ce	9.09	10.69	5.02	4.64	4.57	5.85
Avg. # syllables / s-ce	*24.29*	*24.29****	*16.12*	*17.06*	*16.94*	*23.91*
Min. # syllables / s-ce	3	16	4	10	19	9
Max. # syllables / s-ce	93	106	57	68	57	84
SD # syllables / s-ce	15.28	20.16	9.02	9.15	8.72	11.22
Avg. # letters / word	*5.26*	*5.4*	*5.33*	*5.5**	*5.75****	*5.6****
Min. # letters / word	1	1	1	1	1	1
Max. # letters / word	21	23	15	17	17	20
SD # letters / word	2.85	3.01	2.86	3.05	3.09	3.03
Avg. # syllables / word	*1.57*	*1.65***	*1.61**	*1.66****	*1.74****	*1.68****
Min. # syllables / word	0	0	1	1	1	0
Max. # syllables / word	7	8	5	5	6	6
SD # syllables / word	0.89	1.02	0.91	0.97	1	0.97

Avg. # noun phrases / s-ce	*6.09*	*8.34****	*5.91*	*5.78*	*5.57*	*5.49*
Min. # noun phrases / s-ce	0	3	1	2	2	1
Max. # noun phrases / s-ce	18	17	10	9	11	14
SD # noun phrases / s-ce	2.81	3.28	1.74	1.58	1.55	1.91
Avg. % non-stem words / s-ce	*31.64*	*32.22****	*35.95****	*36.77****	*37.83****	*36.71****
Min. % non-stem words / s-ce	0	14.71	12.5	13.04	15.79	6.25
Max. % non-stem words / s-ce	70	56.25	65	61.11	64.29	83.33
SD % non-stem words / s-ce	11.41	8.61	9.45	9.71	10.41	11.51
% s-ces ending in “?”	1	0.89	0	0	0	0
% s-ces ending in “!”	0	0	0.87	0	0	0
Avg. # punct. signs / s-ce	*1.7*	*2.69****	*0.93****	*1.1****	*1.1****	*1.4*
Min. # punct. signs / s-ce	0	0	0	0	0	0
Max. # punct. signs / s-ce	6	5	4	3	4	6
SD # punct. signs / s-ce	0.56	0.76	0.36	0.38	0.39	0.51
Morphological rich-ness	0.02	0.02	0.02	0.02	0.02	0.02
Avg. # verbs / s-ce	*2.61*	*3.81****	*2.7****	*2.54****	*2.56****	*2.67**
Min. # verbs / s-ce	0	0	0	0	0	0
Max. # verbs / s-ce	8	10	6	6	5	8
SD # verbs / s-ce	1.58	1.69	1.13	1.27	1.2	1.5
Avg. # adj. and adv. / s-ce	*3.07*	*4.17****	*2.42****	*2.5****	*2.67****	*2.71*
Min. # adj. and adv. / s-ce	0	0	0	0	0	0

Max. # adj. and adv. / s-ce	10	14	8	6	8	9
SD # adj. and adv. / s-ce	2.04	2.58	1.5	1.48	1.53	1.62
Avg. # 1st-person pron. / s-ce	*0.09*	*0.06**	*0.2*	*0.08**	*0.02****	*0.06**
Min. # 1st-person pron. / s-ce	0	0	0	0	0	0
Max. # 1st-person pron. / s-ce	2	1	2	2	1	4
SD # 1st-person pron. / s-ce	0.35	0.24	0.5	0.35	0.13	0.35
Avg. # proper nouns / s-ce	*1.05*	*1.24***	*0.02****	*0.28****	*0.11****	*0.3****
Min. # proper nouns / s-ce	0	0	0	0	0	0
Max. # proper nouns / s-ce	14	10	2	15	7	10
SD # proper nouns / s-ce	2.07	2.04	0.19	1.67	0.7	1.03
% words not in Dale-Chall list	44.76	44.91	42.89	46.03	50	48.11
% hapax legomena	20.15	27.51	18.91	21.35	21.89	15.77
Type-to-token ratio (words)	0.31	0.39	0.3	0.32	0.33	0.27
Type-to-token ratio (lemmas)	0.29	0.37	0.29	0.31	0.31	0.25
Average concreteness	2.45	2.36	2.51	2.48	2.47	2.41
10 most common words (excl. stop words)	water, species, climate, change, could, plants, world, many, help, new	could, people, said, would, also, international, humanitarian, countries, must, called	crop, soil, crops, farmers, agriculture, yields, water, farmer, yield, conditions	crop, soil, farmers, agriculture, crops, yields, practices, agricultural, sustainable, water	crop, soil, farmers, crops, practices, yields, sustainable, agricultural, agronomist, farmer	water, soil, farmers, crop, crops, yields, plant, farming, practices, food
10 most frequent words (incl. stop words)	the, of, to, and, in, a, is, that, are, for	the, and, of, to, in, that, a, for, it, with	the, to, and, of, in, crop, soil, crops, a, can	to, the, and, in, of, crop, soil, can, that, are	the, to, of, and, in, a, for, crop, soil, farmers	the, end, to, of, a, in, for, that, their, can
Avg pron. / s-ce	*1*	*1.27****	*0.84****	*0.84****	*0.54****	*0.77*
Min. pron. / s-ce	0	0	0	0	0	0
Max. pron. / s-ce	4	5	5	5	4	4

SD pron. / s-ce	1.03	1.27	1.06	0.93	0.88	0.94
Avg. % anaphora words / s-ce	*10.5*	*10.97*	*9.01****	*9.27****	10.51	*10.75*
Min. % anaphora words / s-ce	0	0	0	0	0	0
Max. % anaphora words / s-ce	42.86	27.27	30.77	27.27	25	30
SD % anaphora words / s-ce	6.6	5.53	6.91	5.96	5.53	6.58
Avg. cos. distance btwn s-ces	*0.16*	*0.12****	*0.38****	*0.32****	*0.35****	*0.24****
Min. cos. distance btwn s-ces	-0.21	-0.22	-0.05	-0.22	-0.06	-0.17
Max. cos. distance btwn s-ces	0.72	0.78	0.85	0.89	0.87	0.89
SD cos. distance btwn s-ces	0.13	0.13	0.13	0.17	0.14	0.16

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

La reconnaissance automatique des relations de cohérence RST en français

Martial Pastor[1] Erik Bran Marino[2] Nelleke Oostdijk[1]
(1) Centre for Language Studies, Radboud University, 6525 HT Nimègue, Pays-Bas
(2) CIDEHUS, Universidade de Évora, 7004-516 Évora, Portugal
{martial.pastor|nelleke.oostdijk}@ru.nl, erik.marino@uevora.pt

RÉSUMÉ

Les parseurs de discours ont suscité un intérêt considérable dans les récentes applications de traitement automatique du langage naturel. Cette approche dépasse les limites traditionnelles de la phrase et peut s'étendre pour englober l'identification de relation de discours. Il existe plusieurs parseurs spécialisés dans le traitement automatique du discours, mais ces derniers ont été principalement évalués sur des corpus anglais. Par conséquent, il n'est pas évident de bien cerner les éléments linguistiques importants sur lesquels les parseurs se basent pour classifier les relations de discours en dehors de l'anglais. Cet article évalue les performances du parseur DMRST sur le corpus RST-DT traduit en français. Nous constatons que les performances de classification des relations de discours en français sont comparables à celles obtenues pour d'autres langues. En analysant les succès et échecs de la classification des relations, nous soulignons l'impact des marqueurs de discours et des structures syntaxiques sur la précision du parseur.

ABSTRACT

RST relation label classification for French.

Discourse parsers have attracted considerable interest in recent natural language processing applications. This approach goes beyond the conventional scope of sentences and can extend to encompass discourse relation identification. There are several parsers specialized in Discourse Parsing, but these have primarily been evaluated on English corpora. Consequently, it is not straightforward to capture the important linguistic elements upon which parsers rely to classify discourse relations outside of English. This article evaluates the performance of the DMRST parser on the RST-DT corpus translated into French. We find that parser performance for the classification of coherence relations in French is comparable to those obtained for other languages. By analyzing the successes and failures of relation classification, we highlight the impact of discourse markers and syntactic structures on the parser's accuracy.

MOTS-CLÉS : parseurs de discours, rhetorical structure theory (RST), relations de discours, traitement automatique du discours, marqueurs de discours, structures syntaxiques.

KEYWORDS: discourse parsers, rhetorical structure theory (RST), discourse relations, discourse parsing, discourse markers, syntactic structures.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 499–507.

1 Introduction

Les parseurs de discours ont suscité un intérêt considérable dans les récentes applications de traitement automatique du langage naturel (Chernyavskiy *et al.*, 2024 ; Braud *et al.*, 2023). Cette tâche consiste à identifier les relations qu'entretiennent des unités de textes à l'intérieur d'un texte plus large. Cette approche dépasse les limites traditionnelles de la phrase et peut s'étendre pour englober l'identification des Relations de Cohérence au niveau du discours. L'un des formalismes les plus populaires pour représenter les Relations de Cohérence est la Rhetorical Structure Theory (RST ; Mann & Thompson, 1988), qui a encouragé la création de jeux de données maintenant utilisés pour l'analyse automatique du discours. Cette dernière tâche est complexe et les parseurs de discours n'ont pas atteint le même niveau de succès que pour d'autres tâches au niveau de la phrase.

Il existe plusieurs parseurs spécialisés dans les structures de type RST (Nguyen *et al.*, 2021, Guz & Carenini ; 2020), mais ces derniers ont été principalement entraînés sur le corpus anglais RST-DT (Carlson *et al.*, 2003). Par conséquent, en ce qui concerne les questions d'explicabilité des parseurs orientés deep learning, il n'est pas évident de bien cerner les éléments linguistiques importants sur lesquels les parseurs se basent pour classifier les relations de discours en dehors de l'anglais. Toutefois, des corpus RST sont disponibles pour d'autres langues que l'anglais, ce qui a donné lieu au développement de parseurs multilingues. Cela dit, à notre connaissance, l'analyse des performances des parseurs et les questions d'explicabilité restent limitées à des analyses effectuées sur de l'anglais. Par exemple, certaines études montrent que l'efficacité des parseurs dépend de la présence de marqueurs de discours [1] (voir exemple (1) [2]) qui rendrait la classification plus facile pour les relations de discours signalées par ces derniers (Pitler *et al.*, 2008).

(1) « [If I sell now,] $\xrightarrow[\text{gold : condition}]{\text{pred : condition}}$ [I'll take a big loss.] » **wsj_2386**

Par ailleurs, selon les analyses effectuées sur des parseurs plus récents (Liu *et al.*, 2023), il semble que certaines structures syntaxiques, comme des *modifications nominales* (2), jouent un rôle déterminant dans la capacité des parseurs à identifier de manière plus efficace les relations de discours.

(2) « [Negotiable, bank-backed business credit instruments] $\xleftarrow[\text{gold : elaboration}]{\text{pred : elaboration}}$ [typically financing an import order.] » **wsj_0602**

Cet article propose une évaluation puis une analyse des réussites et des échecs du parseur DMRST sur le corpus RST-DT traduit en français. Nous commençons par reproduire les expériences avec le parseur multilingue DMRST, que nous évaluons ensuite sur le corpus test du RST-DT. Nous constatons initialement que les performances en termes de classification des relations de cohérence en français sont comparables à celles obtenues pour d'autres langues. Ensuite, nous procédons à une analyse qualitative des cas de succès et d'échecs dans la classification des relations, en mettant en lumière des cas de figure où les relations sont signalées soit par des marqueurs de discours, soit

1. La définition à laquelle nous faisons référence dans cet article concerne les *Discourse Markers* en anglais, qui sont des connecteurs, généralement des conjonctions, entre propositions indépendantes.

2. "pred" ici correspond à l'étiquette prédite par le parseur DMRST présenté dans la section 2, et "gold" correspond à l'étiquette annotée dans le jeu de données RST-DT. La direction de la flèche pointe vers le noyau. La différenciation entre le noyau et le satellite repose sur le fait qu'une unité textuelle agissant comme satellite peut être retirée sans altérer la cohérence du discours, tandis que le retrait d'une unité textuelle agissant comme noyau rendrait le texte incohérent (Mann & Thompson, 1988).

par des structures syntaxiques. Les résultats révèlent que certains signaux syntaxiques facilitent la classification, tandis que certains marqueurs de discours contribuent à une confusion accrue entre différents types de relations.

2 Méthodologie

La Rhetorical Structure Theory. La RST est un modèle d'analyse textuelle des relations de cohérence. Avec celui-ci, les relations entre les segments de texte sont annotées avec différentes classes de relations de cohérence telles que l'élaboration, le contraste, le causal, le temporel, etc. (voir Table 2 pour une liste complète des relations utilisées dans cet article). Les segments de texte d'un arbre RST sont des "unités discursives élémentaires" (EDUs), qui sont des ensembles contigus de tokens approximativement similaires à des propositions indépendantes. Les relations se font non seulement entre les segments de texte mais aussi entre des groupes de segments de texte, ce qui signifie que la représentation finale de RST d'un texte complet (livre, chapitre, article, commentaire, etc.) est un arbre hiérarchique de segments de texte connectés par des relations de cohérence comme sur la Figure 1.

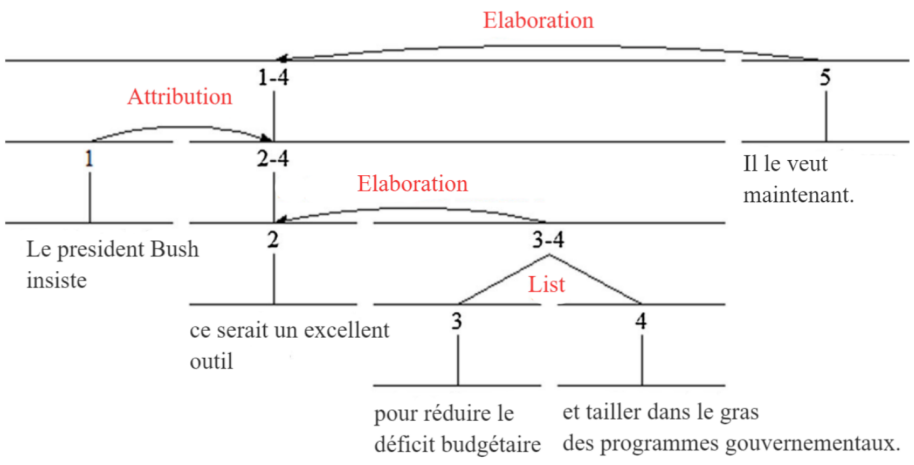


FIGURE 1 – Analyse RST traduite en français issue du document wsj_609 dans le corpus RST-DT qui décrit le modèle d'analyse textuelle des relations de cohérence.

Le corpus RST Discourse Treebank. Le corpus RST-DT (Carlson *et al.*, 2003) a été largement utilisé pour l'analyse RST en anglais et est devenu un choix standard pour évaluer les parseurs RST. Il est reconnu pour ses structures arborescentes hiérarchiques et a été initialement annoté avec 76 relations de cohérence. Les relations examinées ici proviennent de l'ensemble de test RST-DT, qui contient un total de 38 documents, comprenant environ 21 600 tokens. En ce qui concerne les étiquettes de relations, nous utilisons actuellement l'ensemble harmonisé de 18 étiquettes (Braud *et al.*, 2017).

Le parseur multilingue DMRST. Le parseur DMRST développé par Liu *et al.*, 2021 est basé sur XLM-ROBERTA-BASE (Conneau *et al.*, 2020) et est un système multilingue TOP-DOWN qui gère simultanément la segmentation des EDUs et l'analyse des arbres RST. Sa pertinence pour notre étude réside dans ses performances à l'état de l'art en matière de classification des relations de cohérence.

Les auteurs ont fourni un accès à un modèle optimisé pour l'inférence disponible en ligne [3]. Ce modèle particulier a été entraîné sur une collection multilingue de corpus RST, offrant une prise en charge native pour six langues : l'anglais, le portugais, l'espagnol, l'allemand, le néerlandais et le basque. Il est important de noter que, bien que le parseur puisse prédire la structure arborescente et les relations directement à partir de texte brut, notre étude choisit d'utiliser la segmentation des unités discursives (EDU) de référence. Dans notre configuration expérimentale, nous utilisons à la fois le texte brut de l'ensemble de test RST-DT d'origine (traduit en français) et la segmentation des EDUs de référence.

Traduction du corpus RST-DT en français avec ChatGPT. Afin d'évaluer les performances du système DMRST en français, nous avons traduit l'ensemble test du corpus RST-DT en français en utilisant ChatGPT. La traduction avait pour objectif de rester aussi fidèle que possible au texte original : chaque unité discursive des 38 documents (soit 2308 EDUs) a été traduite individuellement pour préserver la structure RST de chaque texte et éviter la génération de paraphrases. De plus, ChatGPT prend en compte le contexte de chaque EDU dans sa traduction ; bien que les unités discursives soient traduites individuellement, les accords grammaticaux, de sujets-verbes ou encore de références sont préservés. Par exemple, la deuxième unité ici accorde correctement au pluriel : [sans les frais de rachat] [qui sont courants pour les rentes.].

La version de ChatGPT utilisée est GPT-4-Turbo. Le prompt qui a été utilisé pour obtenir les traductions est le suivant :

message : {

role.system : "Vous êtes traducteur expert de l'anglais vers le français et un linguiste spécialisé dans le domaine de la Rhetorical Structure Theory (RST)."
role.user : "Traduire le texte ci-dessous en Francais segment par segment, tout en veillant à conserver le contexte global du texte dans son ensemble.
Texte à traduire :
segment 1
segment 2
...
Afficher le résultat avec chaque segment traduit sur une ligne distincte."

}

FIGURE 2 – Prompt utilisé pour traduire le corpus de test RST-DT en français. Ce prompt a été utilisé 38 fois pour les 38 documents de ce corpus.

Encore une fois, il était nécessaire de contraindre le système à traduire EDU par EDU afin de préserver les structures textuelles annotées en relations de cohérence du corpus RST-DT. La question qui se pose dès lors est la suivante : le texte produit est-il bien du français ? Afin d'analyser les biais introduits par une telle contrainte, nous avons comparé, pour un ensemble de 10 documents du corpus de test RST-DT choisis au hasard, une traduction automatique des documents traduits en entier avec la traduction EDU par EDU. Ces 10 mêmes documents ont ensuite été relus et évalués par 3 locuteurs natifs français.

3. https ://github.com/seq-to-mind/DMRST_Parser

En ce qui concerne la comparaison entre le texte traduit par segments et le texte traduit en entier, étant donné que les traducteurs automatiques traduisent souvent phrase par phrase, les traductions étaient similaires, à l'exception de légères variations dans le vocabulaire utilisé et la génération de paraphrases où l'ordre de quelques groupes prépositionnels a été inversé. Nous avons ensuite comparé ces deux traductions automatiques pour chacuns des 10 documents avec les métriques BLEU, ROUGE, TER et WER et avons obtenu les scores moyens suivants : 0.65, 0.82, 20.0 et 0.35.

Quant à l'évaluation manuelle des 3 locuteurs, les 10 documents traduits en français ont été jugés intelligibles et corrects dans leur ensemble. Toutefois les participants ont noté un manque de fluidité, une absence totale d'idiomaticité et la présence, rare mais non négligeable, de calques syntaxiques venus de l'anglais produisant des imprécisions grammaticales, comme « an equity position in Leaseway » traduit par « une position en capitaux dans Leaseway ».

Les résultats de ces analyses nous permettent de conclure que, bien que la contrainte de traduction EDU par EDU n'introduise pas de biais trop importants par rapport à la traduction du texte entier (comme le montrent les scores obtenus), la traduction du français obtenue n'est pas la plus naturelle qui soit. Toutefois, comme nous le verrons dans la partie 4, les éléments linguistiques analysés restent pertinents et correspondent à des usages naturels en français.

3 Expérimentations et résultats

Reconnaissance des relations de cohérence RST en français. Les 38 textes du corpus RST-DT traduits en français avec ChatGPT ont été parsés avec le système multilingue DMRST. Nous présentons ci-dessous les résultats évalués avec la métrique RST-Parseval (Marcu, 2000) pour la classification de relation.

	Précision	**Rappel**	**F1-score**
Français	0.54	0.57	**0.54**
Anglais	0.65	0.67	0.65

TABLE 1 – Résultats globaux évalués avec la métrique RST-Parseval (Marcu, 2000) pour la classification de relation en français et en anglais.

Sans surprise, les résultats sont moins bons que ceux obtenus pour l'anglais, mais sont comparables aux résultats obtenus sur d'autres corpus RST-DT dans différentes langues (Avec un F1-score de 0.62 en portugais, 0.63 en espagnol, 0.47 en allemand, 0.52 en néerlandais et 0.48 en basque, conformément aux résultats signalés par Liu *et al.*, 2021). Comme nous pouvons le voir avec la matrice de confusion sur la Figure 3 ci-dessous, nous remarquons également que la confusion entre les différentes classes de relations est similaire à ce que nous obtenons pour l'anglais concernant les classes ELABORATION et JOINT, cela étant dû à un déséquilibre des classes dans le jeu de données RST-DT.

Relation	rel. frq. (%)	abs. frq. (N)
elaboration	34.43	794
attribution	14.87	343
joint	9.19	212
contrast	6.37	147
same-unit	5.51	127
explanation	4.77	110
enablement	3.82	88
cause	3.56	82
evaluation	3.47	80
temporal	3.17	73
background	2.95	68
topic-comment	1.69	39
summary	1.39	32
comparison	1.26	29
condition	1.21	28
manner-means	1.17	27
topic-change	0.78	18
textualorganization	0.39	9
TOTAL	100.0	2306

TABLE 2 – Fréquence absolue et relative des étiquettes de relation en or dans RST-DT pour un total de 2306 relations.

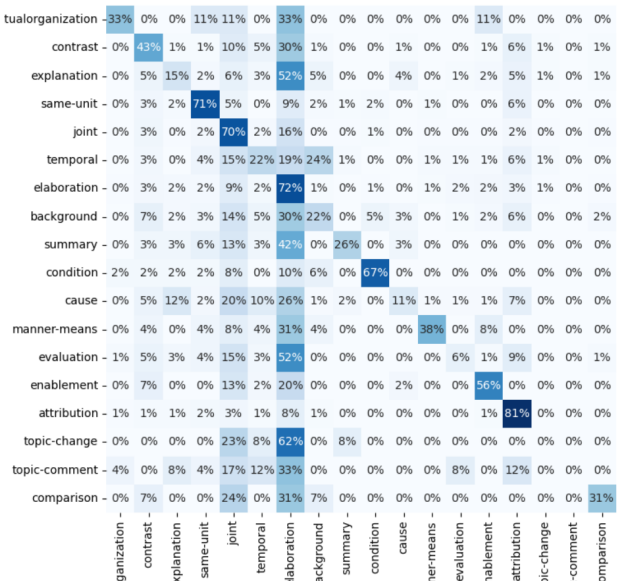


FIGURE 3 – Matrice de confusion de 18 étiquettes harmonisées pour le français.

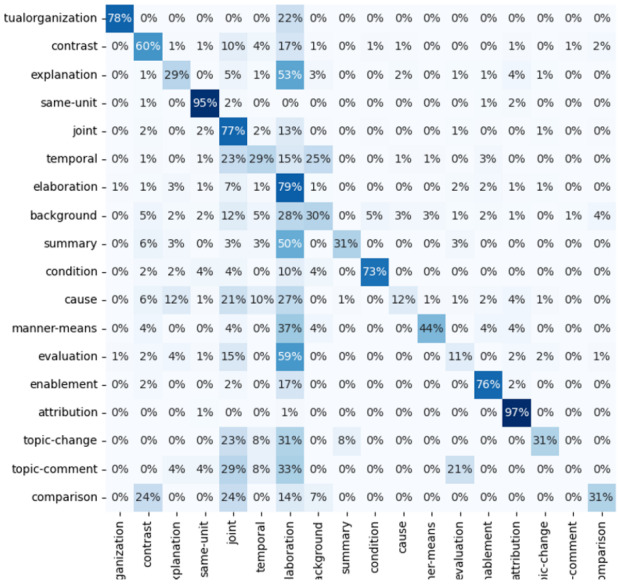


FIGURE 4 – Matrice de confusion pour l'anglais.

4 Discussion et analyse des cas d'échecs / succès

Dans cette section, nous procédons à une analyse qualitative d'un sous-ensemble de relations. Les cas sélectionnés pour l'annotation visent à illustrer en quoi la présence de marqueurs de discours ou de certaines structures syntaxiques influence les erreurs de classification du parseur DMRST. Les cas analysés ci-dessous proviennent d'un ensemble de 301 relations. Ces 301 cas ont été sélectionnés à partir du RST Signaling corpus (Das & Taboada, 2018), où chacune des relations du RST-DT

a été annotée en signaux de relations de cohérence pour de l'anglais. Nous avons extrait les cas correspondant aux catégories décrites ci-dessous et les avons réannotés pour vérifier si ces signaux étaient également applicables au français. À la suite de cette procédure, 21 relations ont été exclues. L'accord inter-annotateur, mesuré par le coefficient Kappa de Cohen, est de 0,83 pour cette phase d'annotations.

Présences de marqueurs de discours. Dans certaines situations, les marqueurs de discours jouent un rôle décisif et contribuent efficacement à prédire la relation. Par exemple, la présence du marqueur de discours *si*, signalant une relation de CONDITION, permet une reconnaissance précise dans 73% des cas pour 41 relations annotées. Cependant, nous constatons que les marqueurs de discours posent souvent des problèmes en raison de leur présence dans de nombreuses relations. Par exemple, nous observons que le marqueur de discours *et* est largement utilisé dans la relation JOINT, mais il est également présent dans de nombreuses autres relations telles que la relation TEMPORAL. Par conséquent, seulement 17% (sur 26 annotations) des relations TEMPORAL signalées par ce marqueur sont correctement prédites. Nous observons une tendance similaire en ce qui concerne la confusion entre les relations TEMPORAL et les relations BACKGROUND, souvent signalées par des marqueurs tels que *après que*, *depuis que que*, *tant que*, *pendant que*, *alors que* ou d'autres formes + QUE comme dans l'exemple (3).

(3) « [dit-il] $\xleftarrow[\text{gold : temporal}]{\text{pred : background}}$ [alors que le boucher s'éloigne ».] **wsj_1146**

La relation BACKGROUND se produit entre une unité de texte principal (nucleus) et un satellite qui est nécessaire à la compréhension de la première unité. Bien que cette relation n'implique pas d'aspect temporel, elle est souvent signalée par des marqueurs de discours tels que *depuis que*, *tant que* ou tout simplement par un *après*, comme le montre l'exemple 4, qui sont également utilisés pour signaler la relation TEMPORAL.

(4) « [Mme Crump a déclaré que le portefeuille de son club d'investissement Ashwood avait perdu environ un tiers de sa valeur] »$\xleftarrow[\text{gold : background}]{\text{pred : temporal}}$ [après le crash du Black Monday] **wsj_2386**

Présences de structures syntaxique. D'un autre côté, on remarque que les relations signalées par des structures syntaxiques ont plus de chances d'être systématiquement prédites. Nous avons annoté ici les structures telles que les *constructions syntaxiques parallèles* (voir exemple (5)) pour la relation JOINT et les *propositions relatives* pour la relation ELABORATION). Nous notons que 79% des relations annotées avec des *constructions syntaxiques parallèles*, soit 72 cas annotés, sont correctement prédites. De même, nous notons un taux de réussite de 85% pour les 141 cas annotés des *propositions relatives*.

(5) « [Parlez-nous de la retenue en matière de dépenses] $\xleftarrow[\text{gold : background}]{\text{pred : temporal}}$ [Parlez-nous des scandales du HUD] » **wsj_0623**

Ambiguïté et Spécificité. L'ambiguïté et la spécificité jouent un rôle crucial dans les performances différenciées du parseur en fonction de la présence de marqueurs de discours ou de structures syntaxiques. Les marqueurs de discours ont tendance à être plus ambigus, car ils peuvent avoir une forme plus ou moins lexicalisée, comme *alors que* ou *après*, qui peuvent être utilisés dans différents contextes. Cependant, cela ne s'applique pas aux marqueurs tels que *si*, qui sont très spécifiques à la relation CONDITION. D'autre part, on observe que les structures syntaxiques sont beaucoup plus spécifiques à certaines relations et, par conséquent, elles induisent moins de confusion pour le parseur.

5 Conclusion

Actuellement, il est important de noter que nous dépendons d'un seul point de contrôle de modèle pour l'expérimentation, ce qui introduit le potentiel d'erreurs influencées par des certaines variations lors de l'entraînement. De plus, nous tenons à souligner que le corpus est limité aux données issues de la presse, et explorer des données provenant de différents genres serait susceptible de fournir des perspectives supplémentaires. Par ailleurs, l'annotation manuelle effectuée ici s'est concentrée sur des cas particuliers inspirés par des analyses menées sur de l'anglais. La création d'un corpus annoté avec des signaux de relations de cohérence, à l'instar de ce qu'ont fait Das & Taboada, 2018 pour l'anglais, nous permettrait d'avoir une vue plus complète des analyses des cas d'échecs et de succès.

En conclusion, cet article a évalué les performances du parseur DMRST sur le corpus RST-DT traduit en français. Nous avons constaté que les performances de ce parseur pour la classification des relations de cohérence en français sont comparables à celles obtenues pour d'autres langues. En analysant les succès et les échecs de la classification des relations, nous avons mis en évidence l'impact des marqueurs de discours et des structures syntaxiques sur la précision du parseur. Cela est attribuable à la spécificité des structures syntaxiques, qui sont étroitement liées à des relations individuelles et ne sont pas aussi ambiguës que les marqueurs de discours.

Remerciements

Le travail présenté dans cet article a été réalisé dans le cadre du projet HYBRIDS, un réseau doctoral Marie Skłodowska-Curie financé par l'Union européenne sous le numéro de subvention 101073351 et par le UK Research and Innovation (UKRI) Horizon Funding Guarantee.

Références

BRAUD C., COAVOUX M. & SØGAARD A. (2017). Cross-lingual RST discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, p. 292–304, Valencia, Spain : Association for Computational Linguistics.

BRAUD C., LIU Y. J., METHENITI E., MULLER P., RIVIÈRE L., RUTHERFORD A. & ZELDES A. (2023). The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In C. BRAUD, Y. J. LIU, E. METHENITI, P. MULLER, L. RIVIÈRE, A. RUTHERFORD & A. ZELDES, Éds., *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, p. 1–21, Toronto, Canada : The Association for Computational Linguistics. DOI : 10.18653/v1/2023.disrpt-1.1.

CARLSON L., MARCU D. & OKUROWSKI M. E. (2003). *Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory*, In J. VAN KUPPEVELT & R. W. SMITH, Éds., *Current and New Directions in Discourse and Dialogue*, p. 85–112. Springer Netherlands : Dordrecht. DOI : 10.1007/978-94-010-0019-2_5.

CHERNYAVSKIY A., ILVOVSKY D. & NAKOV P. (2024). Unleashing the power of discourse-enhanced transformers for propaganda detection. In *Proceedings of the 18th Conference of the*

European Chapter of the Association for Computational Linguistics (Volume 1 : Long Papers), p. 1452–1462.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

DAS D. & TABOADA M. (2018). Signalling of coherence relations in discourse, beyond discourse markers. *Discourse Processes*, **55**(8), 743–770. DOI : 10.1080/0163853X.2017.1379327.

GUZ G. & CARENINI G. (2020). Coreference for discourse parsing : A neural approach. In C. BRAUD, C. HARDMEIER, J. J. LI, A. LOUIS & M. STRUBE, Éds., *Proceedings of the First Workshop on Computational Approaches to Discourse*, p. 160–167, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.codi-1.17.

LIU Y. J., AOYAMA T. & ZELDES A. (2023). What's hard in English RST parsing ? predictive models for error analysis. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, p. 31–42, Prague, Czechia : Association for Computational Linguistics.

LIU Z., SHI K. & CHEN N. (2021). DMRST : A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, p. 154–164, Punta Cana, Dominican Republic and Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.codi-main.15.

MANN W. C. & THOMPSON S. A. (1988). Rhetorical structure theory : Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, **8**(3), 243–281. DOI : /doi.org/10.1515/text.1.1988.8.3.243.

MARCU D. (2000). The Rhetorical Parsing of Unrestricted Texts : A Surface-based Approach. *Computational Linguistics*, **26**(3), 395–448. DOI : 10.1162/089120100561755.

NGUYEN T.-T., NGUYEN X.-P., JOTY S. & LI X. (2021). RST parsing from scratch. In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1613–1625, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.128.

PITLER E., RAGHUPATHY M., MEHTA H., NENKOVA A., LEE A. & JOSHI A. (2008). Easily identifiable discourse relations. In *Coling 2008 : Companion volume : Posters*, p. 87–90, Manchester, UK : Coling 2008 Organizing Committee.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

SUMM-RE: A corpus of French meeting-style conversations

Julie Hunter[1] Hiroyoshi Yamasaki[2] Océane Granier[2]
Jérôme Louradour[1] Roxane Bertrand[2] Kate Thompson[1] Laurent Prévot[2]
(1) LINAGORA Labs, Toulouse, France
{jhunter, jlouradour, cthompson}@linagora.com
(2) Aix Marseille Université & CNRS, LPL, Aix-en-Provence, France
{hiroyoshi.yamasaki, oceane.granier, roxane.bertrand,
laurent.prévot}@univ-amu.fr

RÉSUMÉ

Nous présentons le corpus SUMM-RE un ensemble de données d'environ 95 heures de conversations spontanées de type réunion en français. Le corpus est conçu pour servir de base à des tâches en aval telles que le résumé de réunions. Dans son état actuel, il offre 25 heures de transcriptions corrigées manuellement et alignées sur le signal audio, ce qui en fait une ressource précieuse pour l'évaluation des systèmes d'ASR et de reconnaissance du locuteur. Il comprend également des transcriptions automatiques et des alignements de l'ensemble du corpus qui peuvent être utilisés pour des tâches de NLP en aval. L'objectif de cet article est de décrire la conception, la production et l'annotation du corpus jusqu'à l'étape de transcription, ainsi que de fournir une description quantitative du corpus permettant de comprendre ses principales caractéristiques linguistiques.

ABSTRACT

We present the SUMM-RE corpus, a dataset of roughly 95 hours of spontaneous meeting-style conversations in French. The corpus is designed to serve as a foundation for downstream tasks such as meeting summarization. In its current state, it offers 25 hours of manually corrected transcripts that are aligned with the audio signal, making it a valuable resource for evaluating ASR and speaker recognition systems. It also includes automatic transcripts and alignments of the whole corpus which can be used for downstream NLP tasks. The aim of this paper is to describe the conception, production and annotation of the corpus up to the transcription level as well as to provide statistics that shed light on the main linguistic features of the corpus.

MOTS-CLÉS : Corpus, Réunions, Conversation spontanée, Français, Dialogue, Transcription.

KEYWORDS: Corpus, Meetings, Spontaneous conversation, French, Dialogue, Transcription.

1 Introduction

Spontaneous, multiparty conversation poses particular problems for both speech and natural language processing tasks. Disfluent speech, non-standard or incorrect grammar, idiosyncratic speaker styles, and overlapping speech and interruptions — not to mention transcription errors introduced when transcripts are automatically generated — complicate the task of understanding conversation transcripts, especially for NLP models that have been trained largely on text (Rennard *et al.*, 2023).

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 508–529.

These linguistic features together with complicated acoustic conditions, shared microphones, different accents, and rapid speech aggravate the problem for speech processing tasks such as transcription and speaker identification (Yamasaki *et al.*, 2023).

Progress on these topics is hindered by a lack of data especially for specific tasks such as meeting summarization. Few people are comfortable sharing recordings of their discussions and even if they are, preparing the data for training poses a significant hurdle. Data scarcity is even more serious when we look to languages other than English.

The principal goal of the SUMM-RE project is to develop models for conversation and meeting summarization with a particular focus on French. A critical step in accomplishing this goal has been to collect and prepare a corpus of roughly 100 hours of meeting-style conversations in French. The aim of this paper is to present this corpus and to detail its conception and production.

2 Related corpora

AMI (Augmented Multi-party Interaction ; Carletta *et al.*, 2005) and ICSI (International Computer Science Institute ; Janin *et al.*, 2003; McCowan *et al.*, 2005) are the most well-known meeting corpora (and until recently, were the only meeting corpora at all). AMI contains 137 scenario-driven meetings that last from 15 to 45 minutes each, for a total of around 65 hours of conversations. In each meeting, four participants play roles in a fictitious electronics company and participate in a sequence of four meetings. The roles and scenarios are well developed and always the same, which facilitates the task of getting four strangers to carry out structured discussions on topics for which they have little background knowledge and also avoids privacy concerns triggered by real meanings. On the other hand, while the language remains spontaneous, the heavy corpus design engenders conversational styles and interactions that are arguably much cleaner than real-life meetings and also leads to a homogeneity of content and vocabulary. SUMM-RE by contrast, is designed to encourage more fluid discussion on a range of topics ; the focus is on trying to elicit particular types of discursive interactions that are characteristic of meetings without insisting that the participants play employee-like roles.

ICSI consists of natural, weekly meetings that last about one hour each for a total of roughly 72 hours of recordings. On average, the meetings involve six participants that can include undergraduates, graduate students, and professors who meet to discuss technical topics related to natural language processing, computational linguistics and even the ICSI corpus itself. As ICSI contains real meetings between people who were actively collaborating on projects at the time of recording, the interactions are more natural than those in AMI. At the same time, they draw on technical vocabulary and specialized subjects as well as a considerable amount of shared knowledge between participants, which can complicate the interpretation of the content for models that do not have this knowledge. While such a scenario is undeniably realistic, the SUMM-RE corpus is designed to strike a balance between naturalness and feasibility for automatic summarization. As such, it includes light guidance of the meeting structure and limits the impact of shared background knowledge by bringing in participants who in many cases did not know each other before participating in our corpus.

While both AMI and ICSI are entirely in English, there have been recent efforts to expand to other languages. VCSum (Versatile Chinese Meeting Summarization Dataset Wu *et al.*, 2023) is a collection of transcripts and videos of roundtable meetings in Chinese found on the internet. A total of 239 meetings were selected for a total of over 230 hours of recording time. The corpus is called “versatile”

because it contains a variety of annotations that can be relevant for different summarization tasks. The ELITR corpus (Nedoluzhko *et al.*, 2022) contains transcripts for 113 technical project meetings in English but also for 53 meetings in Czech, for a total of over 160 hours of content. Like ICSI, the meetings in ELITR are natural, leading to many of the same advantages and drawbacks. Unlike ICSI, AMI, VCSum and SUMM-RE however, the ELITR audio files have not been released and parts of the meetings are censored for privacy.

There are also a variety of smaller, conversational corpora in French (for a recent list, see Hunter *et al.*, 2023). CID (The Corpus of Interactional Data Blache *et al.*, 2017), which contains eight one-hour dialogues between friends, has notable similarities with SUMM-RE in that a major effort was involved in adding different levels of annotation, including dialogue-central information that can be exploited by downstream NLP models. With only eight hours of recording, however, it remains very small and is not focused specifically on meetings.

3 Corpus Design

To bypass the concern of sharing personal information, the SUMM-RE corpus does not contain real meetings. The conversations were nevertheless designed to imitate certain important features of meeting-style conversation, developing situations in which participants have to make decisions or plan out a list of action items or report on things they have done. They were also designed to have some continuity with past meetings, as real meetings often do : almost every one of the 96 individual experiments in the corpus is made up of a series of three 20 minute meetings that develop a given topic. [1] In general, [2] the participants were asked to plan an event during the third meeting, while the second meeting focused on deciding what kind of event to plan, and the first meeting involved participants going around the table to discuss their experience or opinions about certain topics. Each one-hour experiment contains the same set of participants—usually four but sometimes fewer—throughout.

In some cases, meeting participants knew each other before the experiment. This condition adds an arguably realistic element to the interactions, as many meetings are held by people who have worked together before. However, because our meeting scenarios were artificial, we also feared that it would encourage playful interactions full of jokes and laughter. While such interactions are certainly possible in professional meetings, many meetings involve a higher level of seriousness and personal distance. In an effort to vary the interactions and imitate different levels of professionalism, we aimed to balance the number of groups in which all participants knew each other, some participants knew each other or no participants knew each other.

To encourage participants to speak naturally and spontaneously while also pushing them to stick to a meeting-like agenda, we had to strike a balance between role playing and freedom for the participants to talk about their own experiences. To encourage freedom, we defined a set of topics that we assumed most participants would be comfortable discussing, including a) internet and technology-related topics like social networks and the societal influence of companies like Google and Amazon, b) films and TV series, c) fundraising, [3] d) music, e) environment-related topics such as global warming and renewable energy. Within these larger topics, participants were allowed to choose a subtopic that

1. Due to technical problems, we had to remove five meetings from the final corpus.
2. In the early pilot studies, we tried different types of organization but ultimately decided upon the one described here.
3. We abandoned this topic, which is discussed in 4.4% of the overall data, because participants struggled to develop it.

interested them.

To add structure to the conversations, we proposed a series of subtopics or responsibilities that each participant could choose to lead as well as a series of points, questions, or tasks (depending on the format of the meeting) that the participants might want to pursue. For each subtopic, the participants received an individual card or a collective document listing the points so that they had a "cheat sheet" if they struggled to develop their contribution to the conversation.

Here, for example, is the set of questions (translated from French to English here) given to a participant who chose to discuss Amazon during a reporting meeting :

> *You have chosen to be responsible for leading the discussion about Amazon. You will need to be able to talk about this subject for 3-4 minutes. To help guide you, below is a list of questions to which you might respond (but you are free to choose other questions if you find them more suitable) :*
> — What is Amazon ? How do we interact with Amazon in our daily lives
> — What are the positive and negative sides of Amazon ?
> — What do you think about their approach to package delivery, the Prime video platform, employee conditions, etc ?
> — How do you think that Amazon will evolve in the future ? Will they become more influential ? Will other actors replace them ?
> — ...
> *You can draw from your personal experience to respond to these questions or provide concrete examples.*

Finally, for each 20 minute conversation, the group was asked to choose a moderator among them. In addition to managing the discussion for a 20 minute conversation (also following suggested guidelines), each moderator was assigned the task of taking notes and of making an oral summary at the end of the meetings based on these notes. To preserve the continuity of the conversation, the summary was recorded immediately after the conversation on the same record. The summary file was later extracted in post-processing.

4 Data acquisition

The original plan was to record the entire SUMM-RE corpus in the H2C2 [4]. Unfortunately, corpus collection began in 2020 and ran through 2021 when the Covid pandemic was still a threat, so in-person meetings were not always an option. We thus adopted two different strategies for corpus collection : in person recordings and Zoom.

In all cases, the recordings shared certain characteristics. For instance, each one hour experiment involved the same set of participants. In general, there were four participants but sometimes, we were only able to find three and on very rare occasions, two. Each participant had their own microphone for recording, as explained below. Finally, each experiment was led by one of the co-authors of this paper, who would begin by giving instructions to the participants and making sure they understood the task, but would leave the room during recording to avoid interaction with the participants.

Most in-person recordings took place in the H2C2, though there are a few exceptions in which it was

4. https://plateformeh2c2.fr/

easier to go into peoples' homes. Participants for these recordings were generally recruited through our platform, calls for participants in social media or announcements made in university courses. In total, 248 out of 283 meetings were recorded in person (see Table 3 for more detail).

The H2C2 studio is composed of two rooms :

— an experimental room where participants interact during the recordings
— an observation and recording room where the person in charge of managing the experiment goes during the experiments and can observe participants through a one-way mirror.

Each participant was equipped with an individual headset with microphone (AKG 520). An additional microphone was used to capture the streams of all speakers at once. All microphones were recorded with a Zoom H8 handy recorder.

While recording conditions in the H2C2 were near ideal, we encountered two problems that impacted the quality of the final recordings. First, although the room was spacious enough to leave a fair amount of distance between speakers and each headset microphone was adjusted to the voice of the person wearing it, sometimes a participant's voice would vary between the test conditions and the final recording conditions. A speaker might end up using a higher pitch during recording due to elevated emotion, for example. In such cases, individual microphones would often end up capturing the voices of multiple speakers. Our efforts to isolate the contributions of the main speaker in such cases are described in Section 5. The second complication resulted from efforts to navigate health and safety regulations enforced during the Covid pandemic. To reduce contact between speakers, we first tried installing plexiglass barriers so that participants could see each other's mouths, but this led to an echo in the recordings. Ultimately, we asked the participants to wear masks. This worked in most cases, but idiosyncratic approaches to mask wearing still led to suboptimal recording in some cases.

35 out of 283 of the meetings were recorded through Zoom. For these, participants were recruited through the crowd sourcing platform Prolific. [5] Each person was required to use their own microphone to take part in the experiment in an effort to limit background noise. While Zoom facilitated the task of speaker identification and minimized the impact of phenomena like overlapping speech, dependence on personal equipment and internet connections led to poor recording quality in some cases. See 7.2 for more information on the effect of communication modality on participants' behavior.

5 Annotation and post-processing

The SUMM-RE corpus is partitioned into `train`, `dev` and `test` data sets with proportions at roughly 75%, 12.5% and 12.5%, respectively. Files were assigned one of three classes randomly with the constraint that the ratio of Zoom experiments to in-person experiments as well as relative proportions of scenarios were kept roughly constant. After this automatic process, minor adjustments were made to ensure that certain files that had been manually corrected ended up in the `dev` set. This procedure resulted in 210 files in the train set, 36 files in the dev set and 37 files in the test set.

The entirety of the SUMM-RE corpus has been automatically transcribed and the entirety of the `dev` set, roughly 12 hours, has been manually corrected and annotated. Correction of the `test` set is underway and will soon be complete. Manual correction is performed in two phases. First, the transcripts are manually corrected for transcription errors. Corrections at this stage are made with the software Praat (Boersma & Van Heuven, 2001). In places where Whisper misses a substantial

5. https ://www.prolific.com/

pause (i.e. on the order of hundreds of milliseconds) a pause marker # is added. If the IPU boundaries differ significantly from the true boundary, this is also manually corrected. All files are then manually verified for any obvious errors made during the first correction as well as minor adjustments such as adding non-linguistic annotations for non-speech sounds like laughs and coughs to improve tier boundary accuracy. On the `dev` set, corrections took one month and were carried out by one of the co-authors who was paid for the task and who is a native speaker of French. Manual verification was performed by another co-author (non-native speaker).

While the details of our automatic pipeline are described in Yamasaki *et al.* (2023), we give a brief overview here. At a high-level, the pipeline can be divided into two parts : the detection of *inter-pausal units* (IPUs)—segments of audio in which the principal speaker is speaking—and the transcription of the words that were uttered in the IPU along with their start and end times. IPU detection was necessary because, as explained in Section 4, individual microphones often captured the voices of multiple speakers. For this task, we employed out-of-the-box IPU annotation provided by the SPPAS package (Bigi & Priego-Valverde, 2019; Bigi, 2015) combined with an approach which relies on speaker diarization using the Pyannote package (Bredin *et al.*, 2020; Bredin & Laurent, 2021). This second step was necessary because the voice intensity processed by SPPAS and the voice quality processed by Pyannote contain complementary information. Once the IPUs were identified, we created new audio files in which background voices were replaced with silences and then passed the resulting audio files to Whisper (Radford *et al.*, 2022), OpenAI's speech to text model. [6]

The quality of the annotation and alignment predicted by our pipeline was extensively evaluated in Yamasaki *et al.* (2023) on a 3.3 hour subset of the `dev` set. Table 1 shows the word error rate (WER) for this subset, broken down by deletions (Del), insertions (Ins) and substitutions (Sub). It also includes T-δ, an average (in milliseconds) of the absolute difference in start and end times for each word and an F1 score inspired by Bain *et al.* (2023). See Yamasaki *et al.* (2023) for details.

Pipeline	**F1**	**T-δ**	**WER**	Del	Ins	Sub
Our Reference	**0.81**	108	**18.8**	8.1	5.8	4.8

TABLE 1 – Brief summary of automatic word level evaluation of the SUMM-RE corpus including F1 score for annotation correctness, time-δ for alignment error (in milliseconds) and word error rate (WER) and corresponding deletion, insertion and substitution scores.

A final point is that in order to prevent identification of individual participants we developed a script to remove participant names from both audio files and transcripts and replace the corresponding audio intervals with a beep. This was done by :

— identifying possible candidates for mentions of individual names by calculating the Levenshtein distance of each token to participants' first and second names
— manually filtering false positives
— replacing the resulting list of names by `anon` and the corresponding WAV interval with a single beep

This anonymization scheme assumes the correctness of the transcription, which is not an issue for the

6. There are several variants of Whisper models such as Klein (2023); Bain *et al.* (2023). We chose Louradour (2023) after a detailed comparison of their performance as detailed in Yamasaki *et al.* (2023). We opted for the large v2 model as a starting point but added two custom features. The first employed prompting to encourage Whisper to transcribe disfluencies, yielding a more faithful transcript. The second involved introducing precise word alignment through techniques developed in Louradour (2023) and the Julius forced aligner (Lee *et al.*, 2001). Further details of both the IPU detection algorithm and the Whisper transcription process can be found in Yamasaki *et al.* (2023).

test and dev splits as they are manually corrected, but is for the train set which is not. In cases where an individual's name was not correctly transcribed it may not be anonymized. However, we noticed that it is highly rare for participants to use full names or family names to refer to each other. As first names are not particularly identifiable we believe this anonymization scheme is sufficient.

6 Basic corpus statistics

The SUMM-RE corpus includes a total of 207 unique participants. Due to the recruitment procedure described in Section 4, our data is biased with respect to age, gender and occupation. In particular, the majority of participants were students with a mean age of 28.7, there are nearly three times more female than male participants, and a large portion of participants were monolingual French speakers who grew up in France. See 2 in the Appendix for further metadata on participants.

Our data set consists of 96 sessions with 3 conversations for almost every session, yielding a total of 283 conversations of roughly 20 minutes each. [7] Of these, 8 sessions (22 files) are pilot experiments and the remaining 88 sessions (261 files) are non-pilot experiments. As stated above (Section 4) some meetings (35/283 files) were recorded on Zoom due to Covid restrictions.

7 Linguistic statistics

Dialogue corpora vary greatly depending on a wide range of contextual factors. To better understand the nature of a dialogue corpus, it is crucial to look at its linguistic, and in particular its interactional, properties. In this section, we focus on some core metrics that are are interesting proxies to characterize the corpus. Namely, we look at (i) distribution of Inter-Pausal Unit (IPU) durations (as a proxy to sentence length in the written realm); (ii) automatically extracted backchannels; (iii) amount of overlapping speech; (iv) filled pause count; and (v) speaker dominance. Together these metrics provide an insight about the degree of spontaneity and interaction in the corpus.

7.1 By metric

IPU distribution Figure 1 shows the distribution of IPU duration and number of IPUs for dev, train and test. As the splits were random, both IPU duration and count are similar across splits, though there are slightly more IPUs shorter than 1 second for train than for dev and test. This may be due to erroneous IPUs (e.g., from someone else laughing) removed during correction.

Independent t-tests indicated that while the number of IPUs in meetings from train and dev varied significantly ($p \leq 0.01$), test was consistent with train, meaning that there should not be an issue when using test for evaluation. There was no significant difference for pilot and experiment recordings either ($p = 0.17$), which justifies combining them into a single data set. We also performed t-tests to compare our different scenarios (a-e) and meeting styles (Section 3). For scenarios, 6 pairs reached significance : a-c $p \leq 0.001$, a-d $p \leq 0.05$, a-e $p \leq 0.01$, b-c $p \leq 10^{-4}$, c-d $p \leq 10^{-4}$, c-e $p \leq 10^{-5}$. For tasks, reporting differed significantly from both decision ($p \leq 10^{-4}$) and planning ($p \leq 10^{-5}$), which is to be expected given that reporting meetings were designed to be less interactive.

7. 005b_PBP, 005b_PBR, 012a_EBR, 017b_EBD, 087c_EEP were rejected due to technical issues.

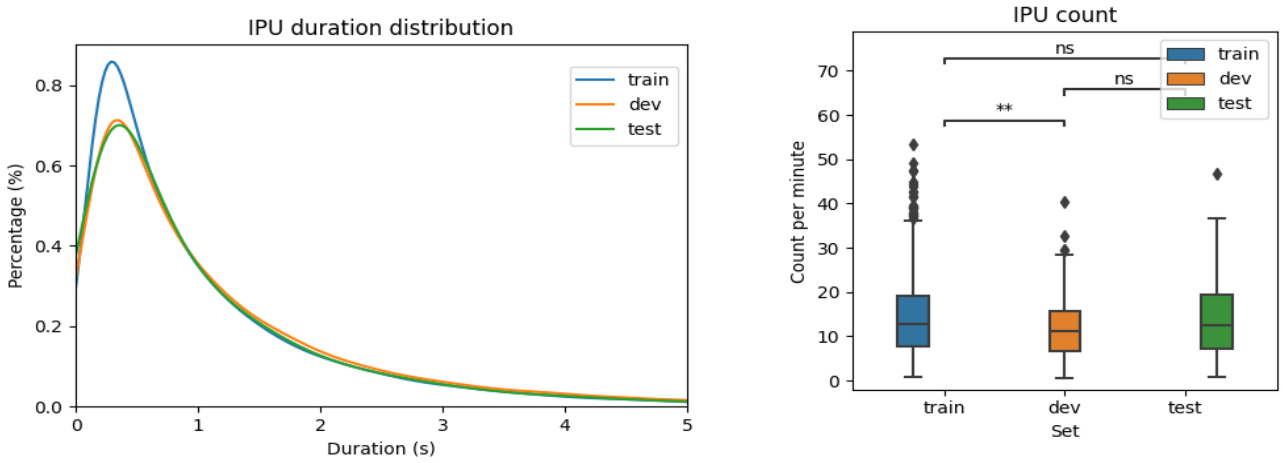


FIGURE 1 – Left : kernel density estimate of IPU duration by splits, Right : number of IPUs by splits

Token distribution. Figure 2 shows the distribution of token duration and speech rate. As expected, there is only a small difference between splits. Average number of tokens per file was 1187 words for `train`, 1107 words for `dev` and 1218 words for `test`. Independent t-test revealed no significant differences between splits (train-dev : $p = 0.11$, dev-test : $p = 0.08$, train-test : $p = 0.55$). Pilot vs experiment comparison was not significant ($p = 0.99$). Comparison by scenarios gave 4 significant pairs : a-c $p \leq 0.01$, b-c $p \leq 0.05$, c-d $p \leq 0.01$ and c-e $p \leq 0.01$. Comparison by task showed that reporting again differed from both decision ($p \leq 0.001$) and planning ($p \leq 0.05$).

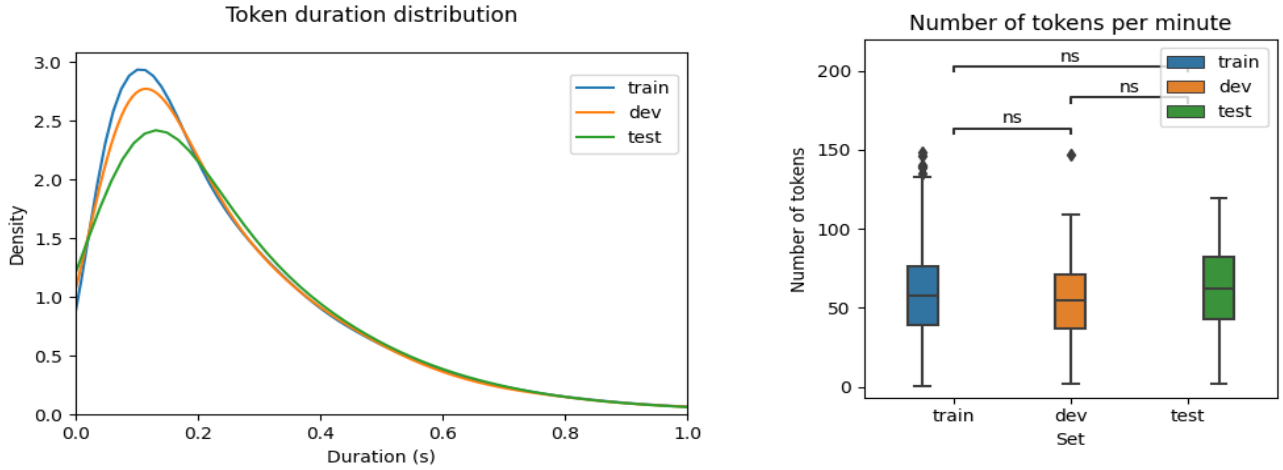


FIGURE 2 – Left : kernel density estimate of token duration by splits, Right : speech rate by splits

Backchannels. Figure 3 (Left) shows the number of backchannels found per minute. For the purpose of this paper we used a rather simplistic definition of backchannel in which we only considered single word utterances that matched "`oui`", "`ouais`", "`hm`", "`mh`", "`non`", "`ok`", "`ah`", "`ben`", "`bien`", "`eh`", "`euh`", "`voilà`". This decision was made for the purpose of simplicity but has a disadvantage of underestimating the actual number.

The number of backchannels did not differ between `train` and `test` ($p = 0.62$) but was slightly lower for `dev` than `train` ($p \leq 0.01$) and `test` ($p \leq 0.01$). Comparison between pilots and experiments did not reach significance ($p = 0.37$). Seven scenario pairs reached significance : a-b $p \leq 0.01$, a-c $p \leq 0.01$, a-d $p \leq 0.01$, b-c $p \leq 0.001$, c-d $p \leq 10^{-4}$, c-e $p \leq 0.001$. The reporting

task again differed from decision ($p \leq 10^{-5}$) and planning ($p \leq 10^{-5}$). Finally, in person meetings had more backchannels than Zoom meetings ($p =\leq 0.001$). See Appendix B.3 for the figures.

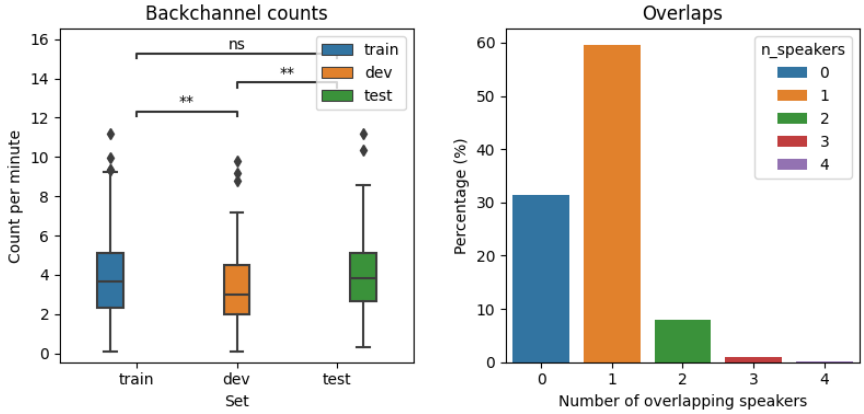


FIGURE 3 – Left : Backchannel counts by sets, Right : Number of speakers at the same time

Overlap. Figure 3 (Right) shows the relative duration of intervals in which $N \in 0, 1, 2, 3, 4$ speakers were speaking at the same time. This was calculated by 1) segmenting the entire conversation into small segments of equal duration (here 0.02 seconds), 2) counting how many speakers are speaking in each segment, 3) aggregating by number of simultaneous speakers, 4) normalizing by the total duration. As expected, in the majority of cases, there is either one speaker or no speaker ; overlaps with more than three participants are extremely rare, in line with previous findings (Çetin & Shriberg, 2006). This suggests that the turn-taking system is robust in multi-party interaction, though overlaps still constitute an important factor (around 10% of speaking time) to consider in downstream processing. We observe that overlaps are even more rare in zoom meetings. See Appendix B.4 for full results.

Filled pause. To assess the degree of conversational fluidity, we considered the number of filled pauses (e.g. "euh", marked as `fp` in the final annotation). The results showed no significant difference for data set, split or task, although there were more filled silences for Zoom meetings than in-person meetings ($p \leq 10^{-5}$). Full results can be found in Appendix B.5.

Dominant speaker. We also looked at whether certain speakers spoke substantially more than others. Results (Appendix B.6) show that on average, there is a dominant speaker and that they tend to speak around 50% of total speaking time followed by the second most active speaker at around 30%. The decrease in percentage is fairly linear. [8] There were no obvious trends across different conditions.

7.2 By condition

In-person vs. Zoom. Figure 4 shows that the number of IPUs ($p \leq 10^{-5}$) and the number of tokens per minute ($p \leq 10^{-5}$) were lower for Zoom than in-person meetings. The number of backchannels per minute was also lower for Zoom ($p \leq 0.001$), while the number of filled pauses was significantly larger ($p \leq 10^{-5}$). Overall, these results indicate that Zoom meetings are less interactive than in person meetings, which is to be expected given the social distance and the fact that participants often cut their microphones when they did not have the floor (as people tend to in real-life online meetings).

8. Some conversations have only three participants meaning the fourth participant's speaking time is set to zero. Thus the data about the fourth participant should be interpreted with care.

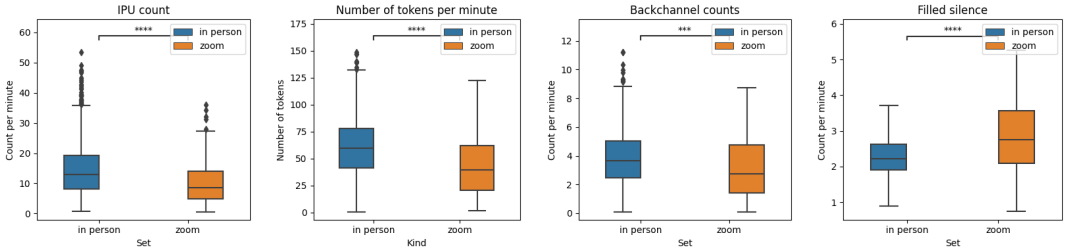


FIGURE 4 – Comparisons of in-person and Zoom meetings from left to right : IPU count, token count, number of backchannels, number of filled pauses

Differences between different scenarios. As noted in Section 3, scenario (c) was too difficult and had to be abandoned. This is consistent with the fact that IPU counts, number of tokens per minute, filled pauses and backchannel counts all differed significantly between scenario (c) and the others (see above and Appendix B for figures and details). Scenario (a) also stood out. IPU counts were smaller for (a) (technology) than (e) (environment, $p \leq 0.01$) and (d) (music, $p \leq 0.05$). Backchannel counts per minute were different for : a-d ($p \leq 0.05$), a-e ($p \leq 0.01$), a-c ($p \leq 0.001$).

Overlaps paint a slightly different picture. While scenario (c) has more pauses and less speech, consistent with the above, scenario (a) has the smallest amount of silence and highest amount of one person talking. This might suggest that scenario (a) is a "more serious" topic, for which people have a tendency to speak longer and in a less chat-like manner (see Appendix B for figures).

Differences between different tasks. Comparison by task revealed that reporting meetings have fewer IPUs than planning ($p \leq 10^{-5}$) and decision ($p \leq 10^{-4}$), fewer tokens per minute than planning ($p \leq 0.05$) and decision ($p \leq 0.001$), and fewer backchannels than planning/decision ($p \leq 0.0001$). Reporting also had highest portion where only one person is speaking (see Appendix B for figures). This is to be expected due to the fact that reporting meetings were designed in a round-table style in order to encourage monologue. There was no difference in the number of filled pauses across tasks.

8 Conclusion

We have presented the SUMM-RE corpus, a new dataset containing 96 hours of spontaneous, multiparty meeting-style conversations in French. The corpus is the only French corpus of its kind, and one of the only large-scale meeting corpora in a language other than English. While the meetings are based on loose role-playing, they remain natural and are designed to elicit basic discursive interactions that we can expect from real meetings. We have offered preliminary analyses of the data to illustrate the level of spontaneity and interactivity in the corpus and have shown how this can vary depending on the type of meeting involved, the subject and—although this was not a part of the initial aim of the corpus—whether the conversation was recorded in-person or on Zoom. The dataset is available on Hugging Face at https://huggingface.co/datasets/linagora/SUMM-RE.

9 Acknowledgements

We gratefully acknowledge support from the ANR grant SUMM-RE (ANR-20-CE23-0017).

Références

BAIN M., HUH J., HAN T. & ZISSERMAN A. (2023). Whisperx : Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv :2303.00747*.

BIGI B. (2015). Sppas-multi-lingual approaches to the automatic annotation of speech. *The Phonetician. Journal of the International Society of Phonetic Sciences*, **111**(ISSN : 0741-6164), 54–69.

BIGI B. & PRIEGO-VALVERDE B. (2019). Search for inter-pausal units : application to cheese ! corpus. In *9th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics*, p. 289–293.

BLACHE P., BERTRAND R., FERRÉ G., PALLAUD B., PRÉVOT L. & RAUZY S. (2017). The corpus of interactional data : A large multimodal annotated resource. *Handbook of linguistic annotation*, p. 1323–1356.

BOERSMA P. & VAN HEUVEN V. (2001). Speak and unspeak with praat. *Glot International*, **5**(9/10), 341–347.

BREDIN H. & LAURENT A. (2021). End-to-end speaker segmentation for overlap-aware resegmentation. In *Interspeech*.

BREDIN H., YIN R., CORIA J. M., GELLY G., KORSHUNOV P., LAVECHIN M., FUSTES D., TITEUX H., BOUAZIZ W. & GILL M.-P. (2020). pyannote.audio : neural building blocks for speaker diarization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

CARLETTA J., ASHBY S., BOURBAN S., FLYNN M., GUILLEMOT M., HAIN T., KADLEC J., KARAISKOS V., KRAAIJ W., KRONENTHAL M. *et al.* (2005). The AMI meeting corpus : A pre-announcement. In *International workshop on machine learning for multimodal interaction*, p. 28–39 : Springer.

ÇETIN O. & SHRIBERG E. (2006). Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site : Insights for automatic speech recognition. In *Ninth international conference on spoken language processing*.

HUNTER J., LOURADOUR J., RENNARD V., HARRANDO I., SHANG G. & LORRÉ J.-P. (2023). The claire french dialogue dataset. *arXiv preprint arXiv :2311.16840*.

JANIN A., BARON D., EDWARDS J., ELLIS D., GELBART D., MORGAN N., PESKIN B., PFAU T., SHRIBERG E., STOLCKE A. *et al.* (2003). The ICSI meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03)*, volume 1, p. I–I : IEEE.

KLEIN G. (2023). Faster whisper transcription with ctranslate2. *GitHub repository*.

LEE A., KAWAHARA T., SHIKANO K. *et al.* (2001). Julius-an open source real-time large vocabulary recognition engine. In *INTERSPEECH*, p. 1691–1694.

LOURADOUR J. (2023). whisper-timestamped. *GitHub repository*.

MCCOWAN I., CARLETTA J., KRAAIJ W., ASHBY S., BOURBAN S., FLYNN M., GUILLEMOT M., HAIN T., KADLEC J., KARAISKOS V., KRONENTHAL M., LATHOUD G., LINCOLN M., LISOWSKA MASSON A., POST W., REIDSMA D. & WELLNER P. (2005). The AMI meeting corpus. *International Conference on Methods and Techniques in Behavioral Research*.

NEDOLUZHKO A., SINGH M., HLEDÍKOVÁ M., GHOSAL T. & BOJAR O. (2022). ELITR Minuting Corpus : A novel dataset for automatic minuting from multi-party meetings in English and Czech. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC-2022)*, Marseille, France : European Language Resources Association (ELRA). In print.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv :2212.04356*.

RENNARD V., SHANG G., HUNTER J. & VAZIRGIANNIS M. (2023). Abstractive meeting summarization : A survey. *Transactions of the Association for Computational Linguistics*, **11**, 861–884.

WU H., ZHAN M., TAN H., HOU Z., LIANG D. & SONG L. (2023). VCSUM : A versatile Chinese meeting summarization dataset. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Findings of the Association for Computational Linguistics : ACL 2023*, p. 6065–6079, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.377.

YAMASAKI H., LOURADOUR J., HUNTER J. & PRÉVOT L. (2023). Transcribing and aligning conversational speech : A hybrid pipeline applied to french conversations. In *2023 IEEE Automatic Speech Recognition and Understanding*.

A Basic data set information

Age		28.7 ± 13.4 years
Gender	M	N = 56
	F	N = 146
	Other	N = 5
Country	France	N = 158
	Other	N = 49
Occupation	Student	N = 120
	Other	N = 87
Languages spoken		1.6 ± 0.9
Total		207

TABLE 2 – Participants metadata summary

Pilot	pilot	N = 22
	experiment	N = 261
Location	zoom	N = 35
	H2C2	N = 212
	Home	N = 18
	LPL	N = 18
Task	Reporting	N = 95
	Decision	N = 94
	Planning	N = 94
Scenario	A	N = 84
	B	N = 86
	C	N= 12
	D	N = 36
	E	N = 65
Video	yes	N = 241
	no	N = 42
duration		19min 42 ± 3min 8

TABLE 3 – Session metadata summary

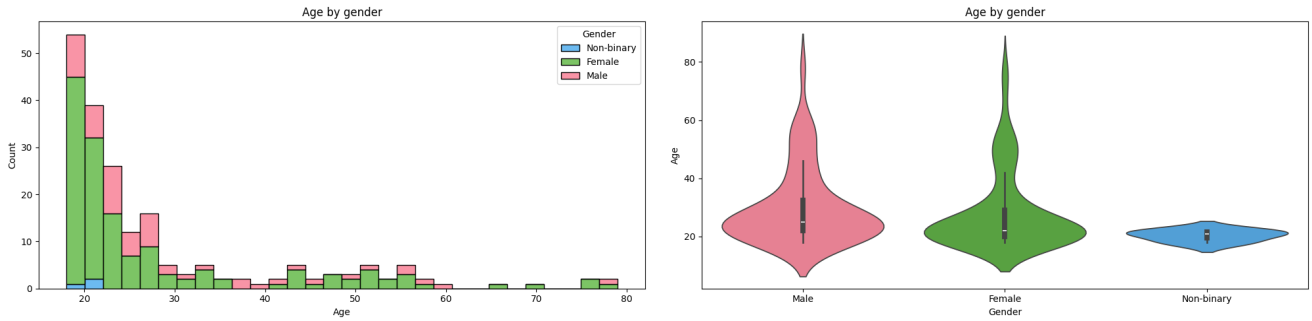


FIGURE 5 – distribution of age by gender

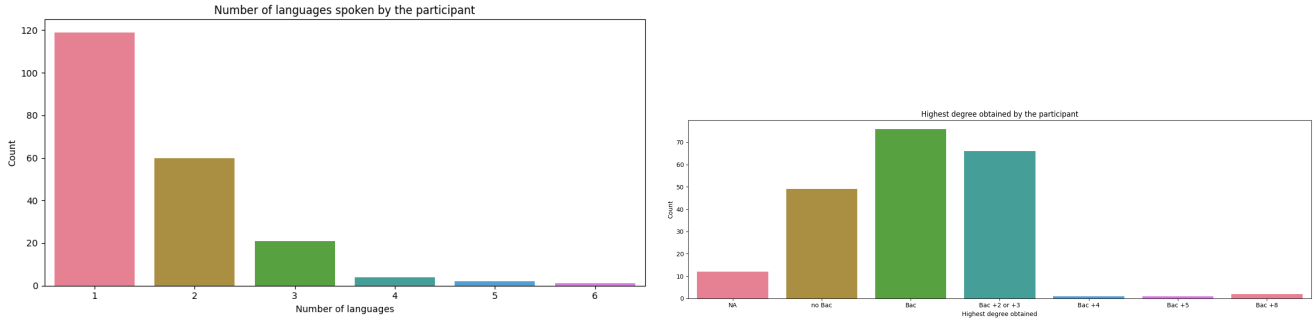


FIGURE 6 – Left : number of languages spoken by the participant, Right : Highest degree obtained by the participant

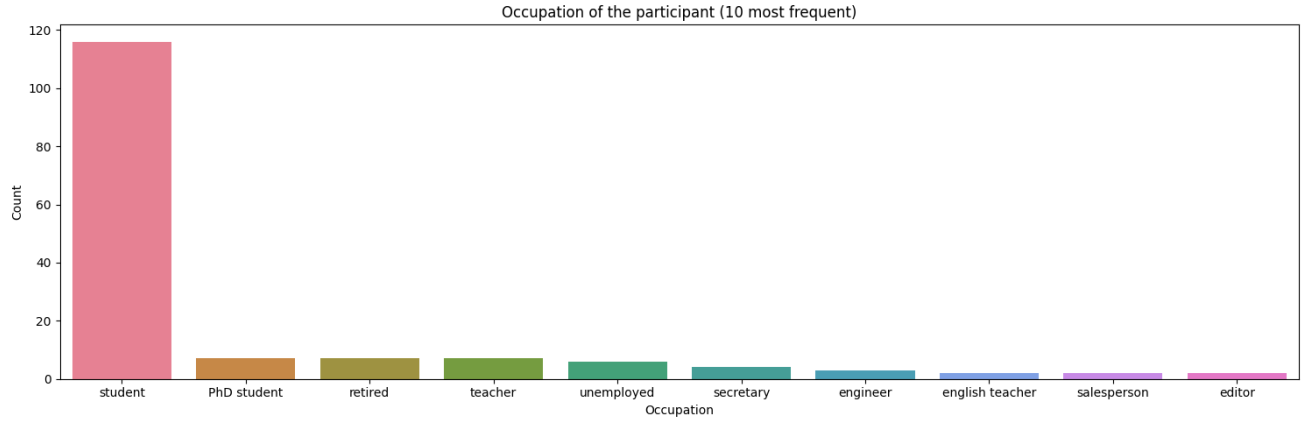


FIGURE 7 – Occupations of the participants

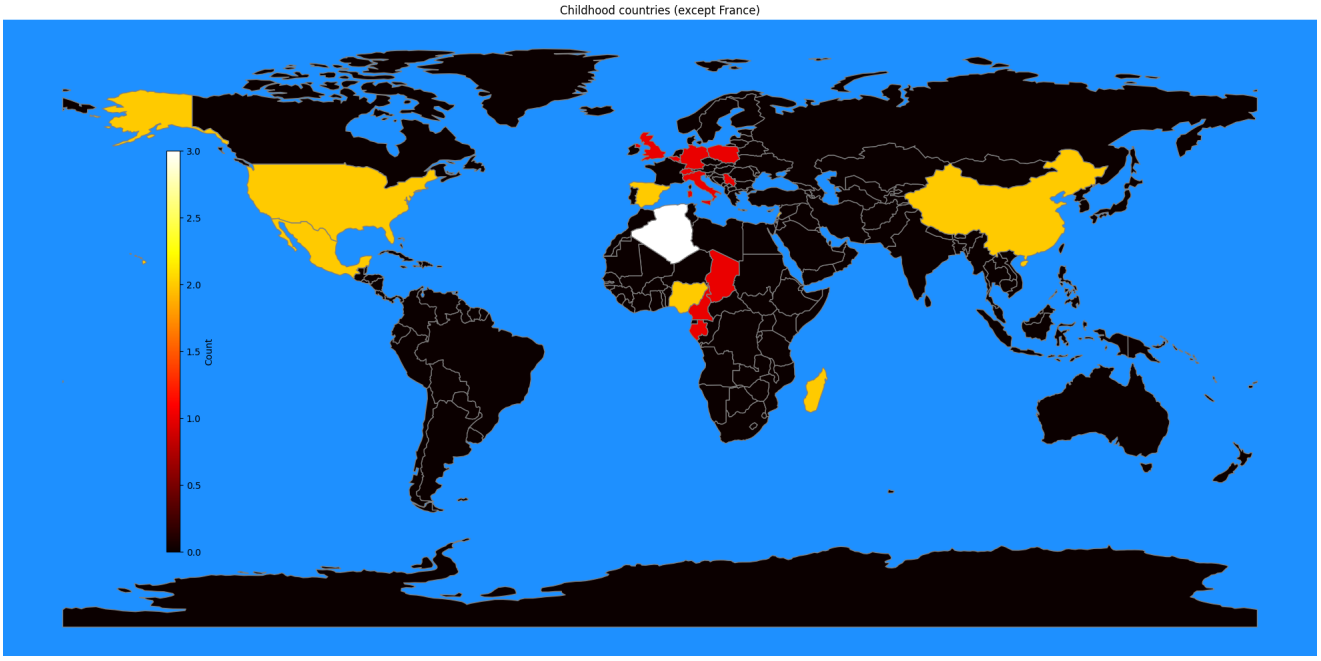


FIGURE 8 – Countries participants are from (apart from France)

B Additional linguistic statistics

B.1 IPU

B.1.1 Distribution

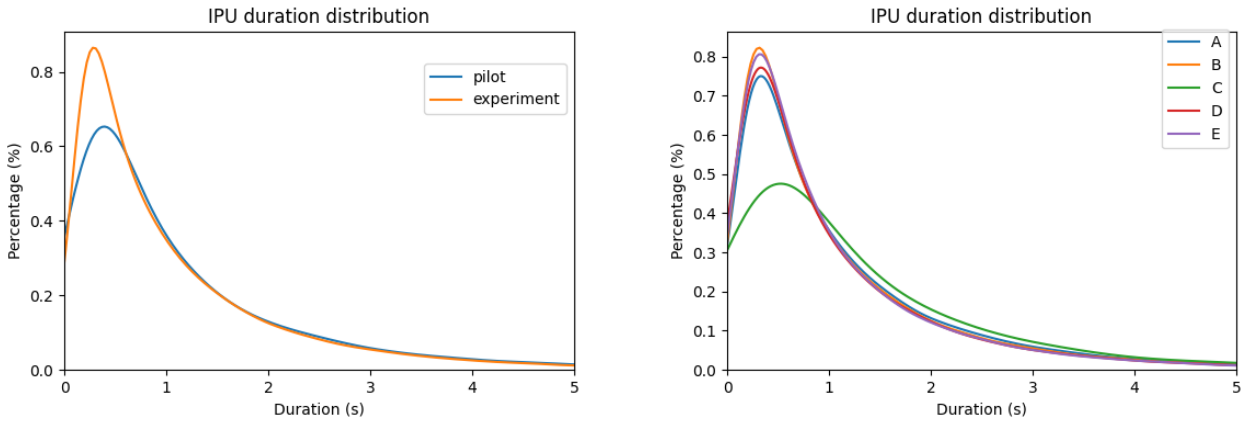


FIGURE 9 – Left : IPU duration distribution by pilot vs experiment, Right : IPU duration distribution by scenario

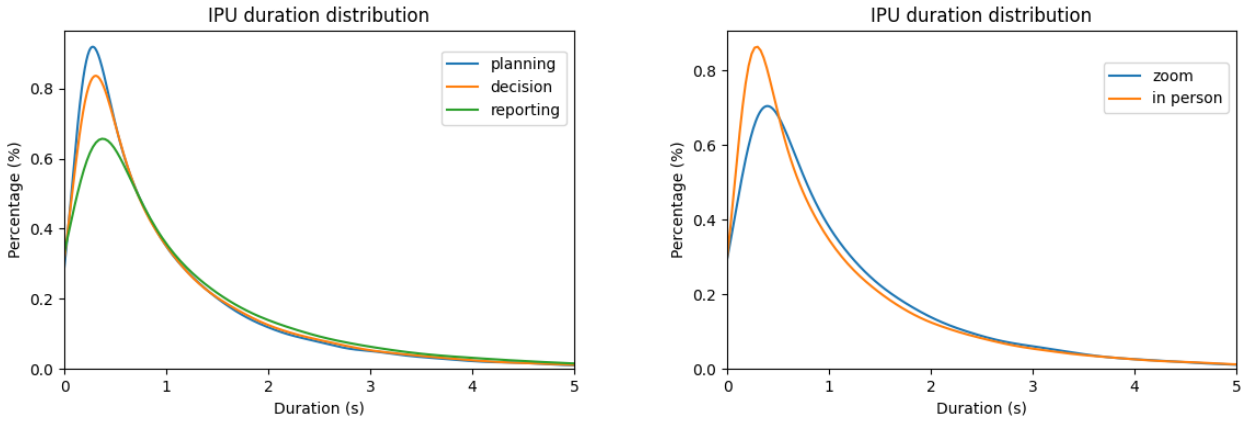


FIGURE 10 – Left : IPU duration distribution by task, Right : IPU duration distribution by place

B.1.2 Count

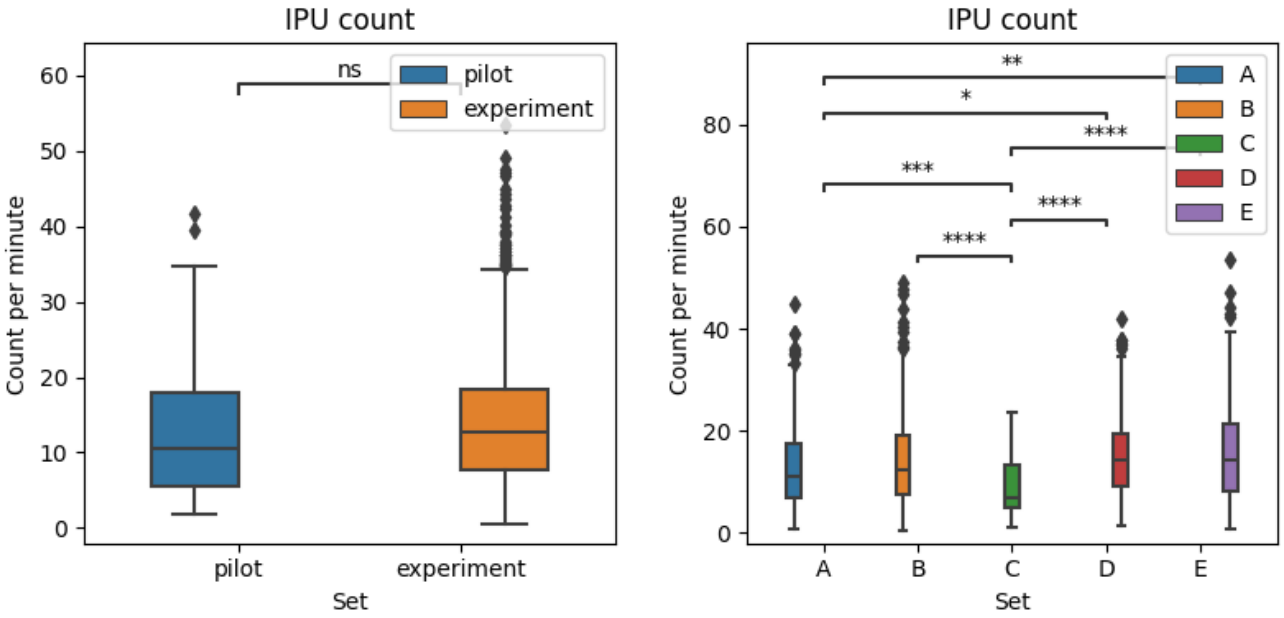


FIGURE 11 – Left : IPU count by pilot vs experiment, Right : IPU count by scenario

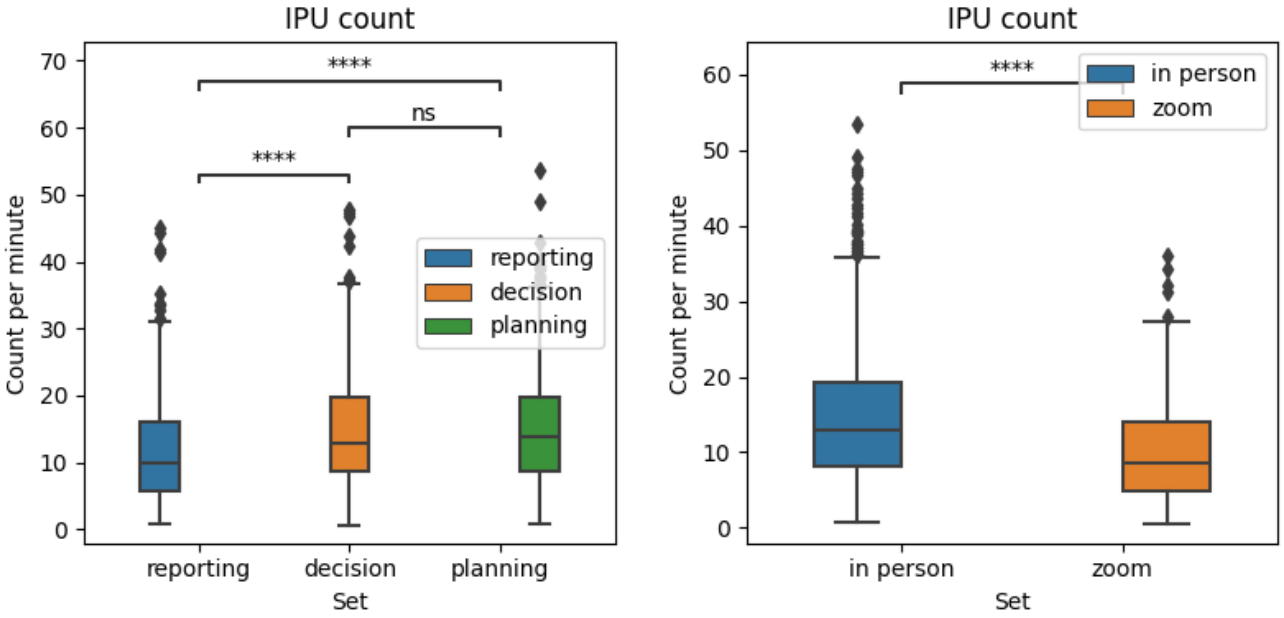


FIGURE 12 – Left : IPU count by task, Right : IPU count by place

B.2 Token

B.2.1 Distribution

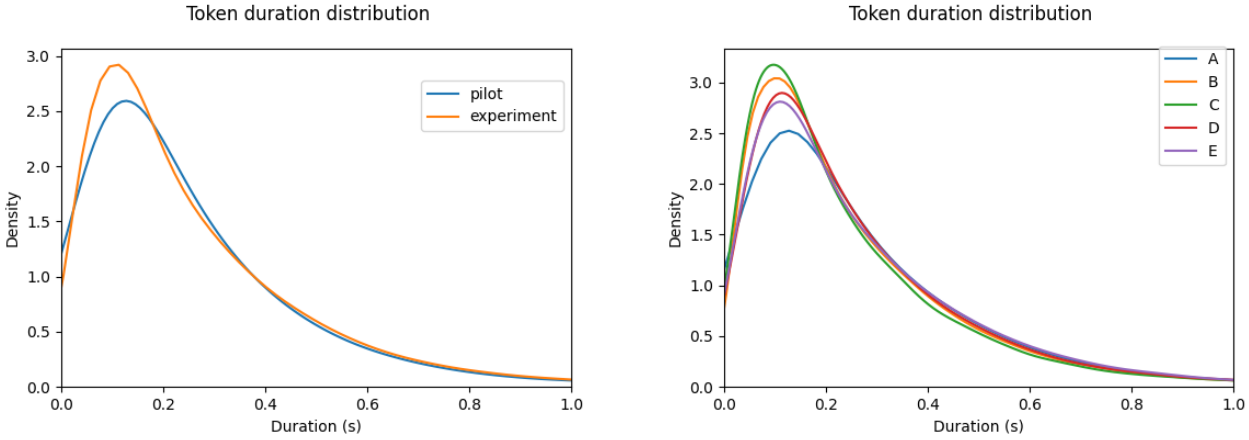


FIGURE 13 – Left : Token duration distribution by pilot vs experiment, Right : Token duration distribution by scenario

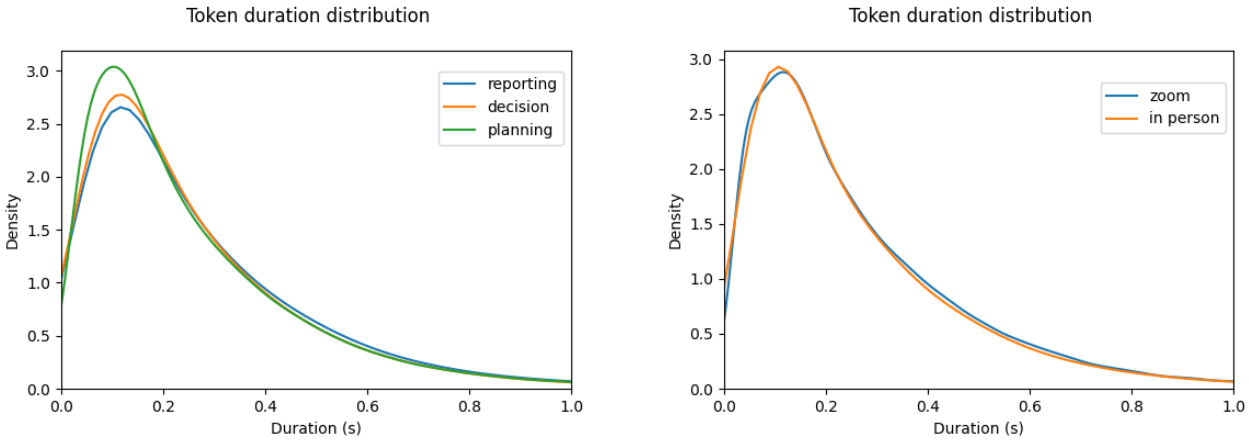


FIGURE 14 – Left : Token duration distribution by task, Right : Token duration distribution by place

B.2.2 Count

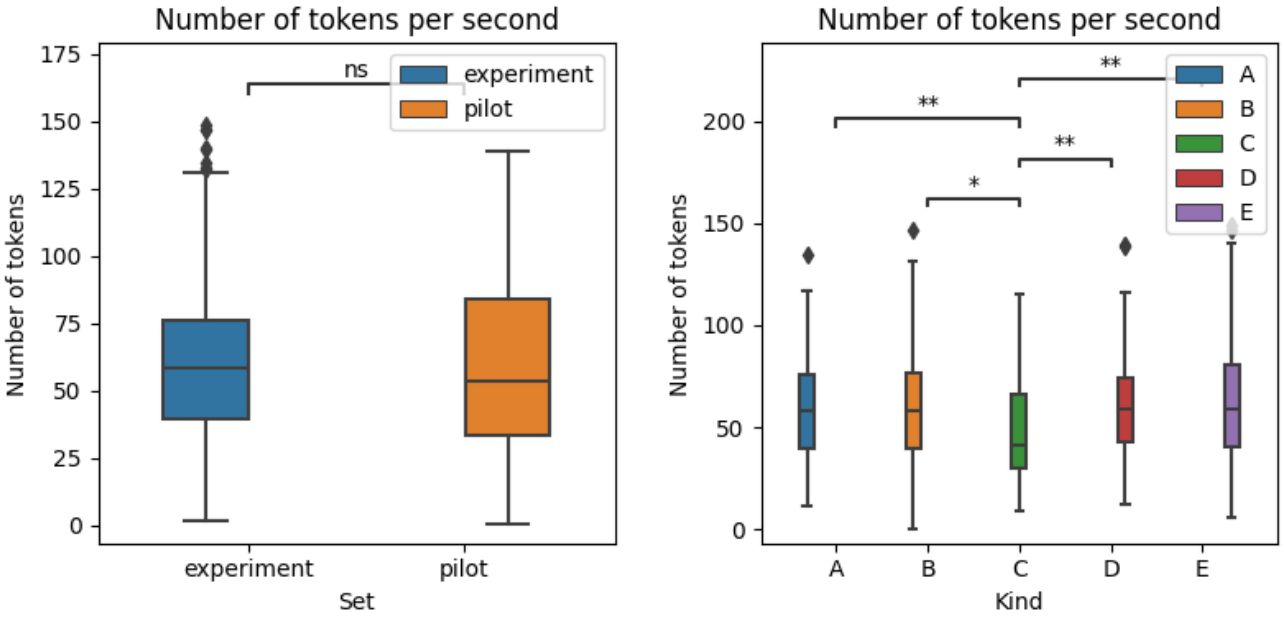


FIGURE 15 – Left : Token count by pilot vs experiment, Right : Token count by scenario

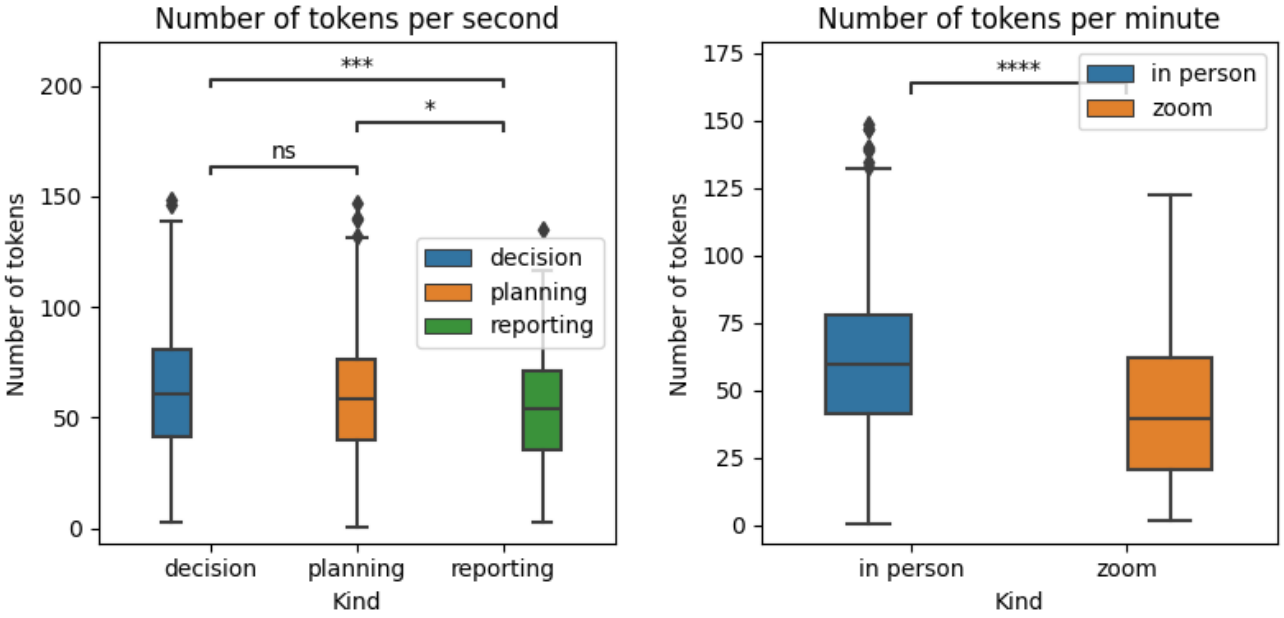


FIGURE 16 – Left : Token count by task, Right : Token count by place

B.3 Backchannel

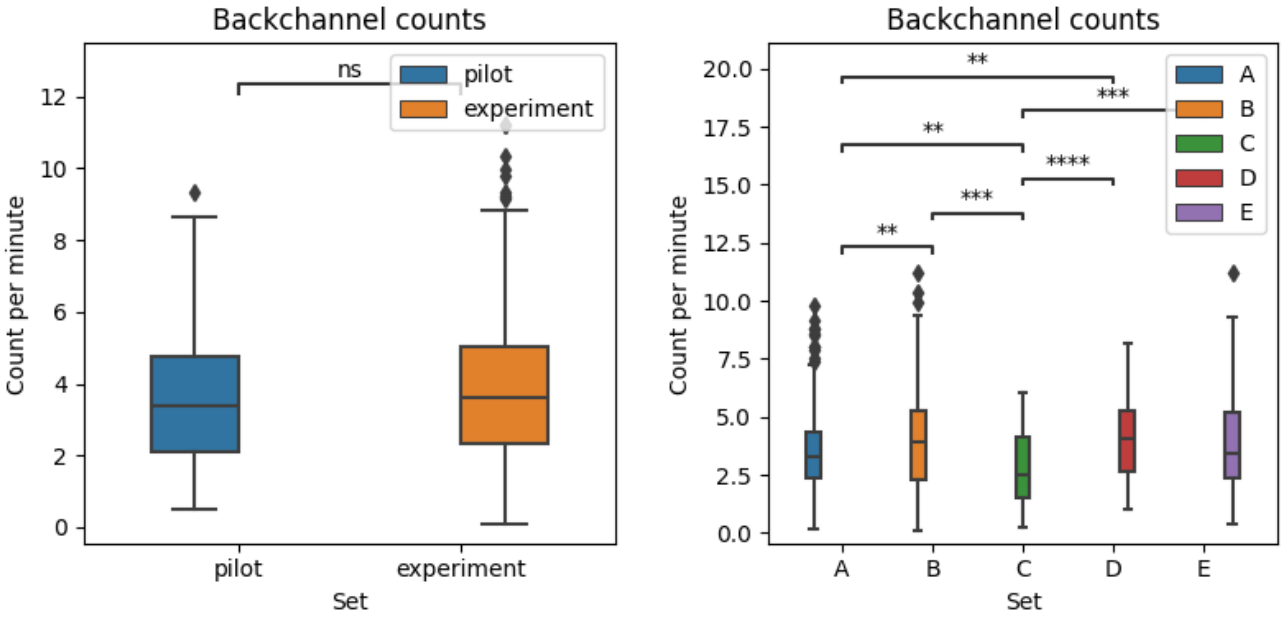


FIGURE 17 – Left : Backchannel count by pilot vs experiment, Right : Backchannel by scenario

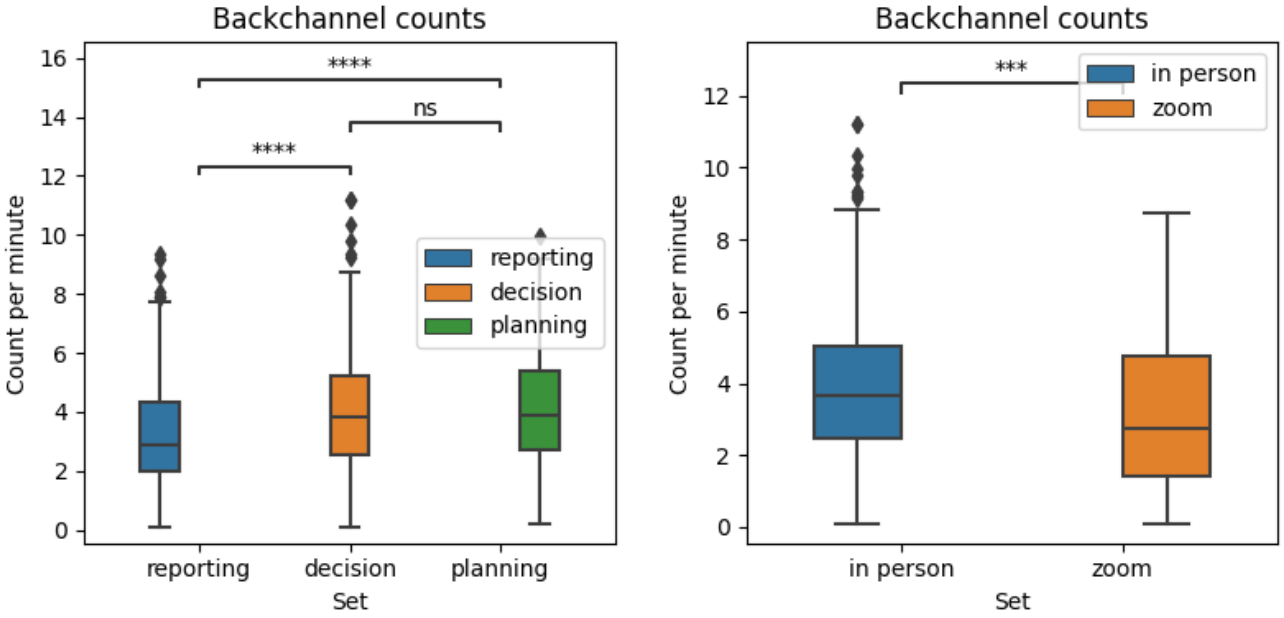


FIGURE 18 – Left : Backchannel count by task, Right : Backchannel by place

B.4 Overlap

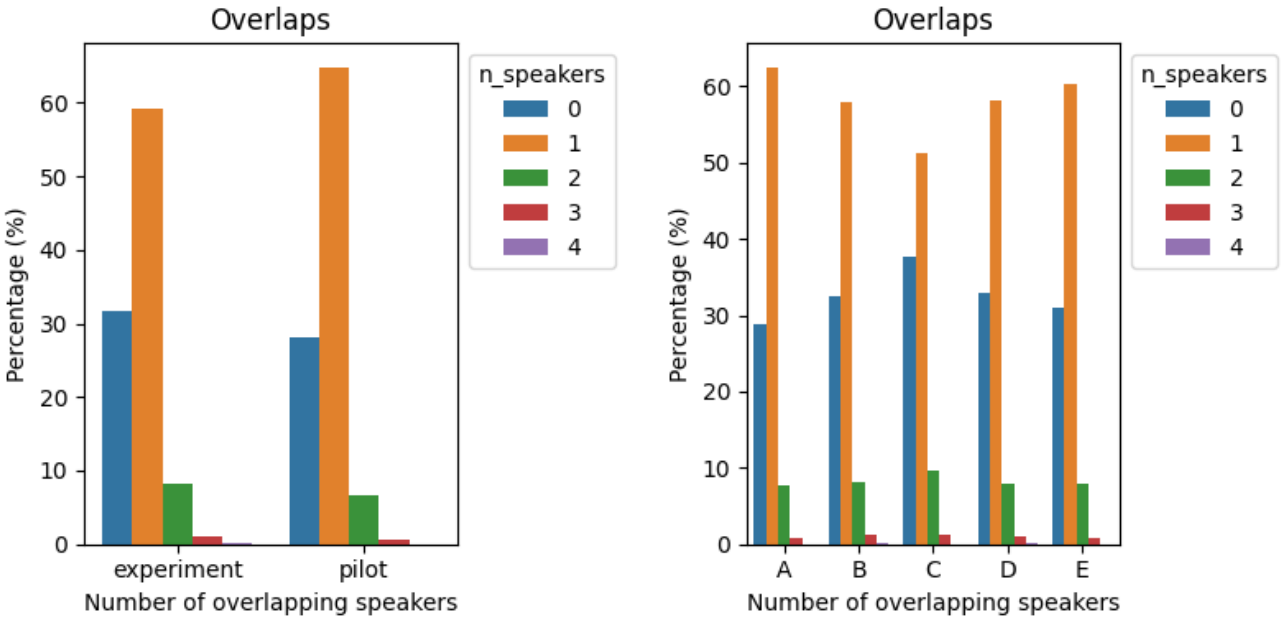


FIGURE 19 – Left : Overlaps by pilot vs experiment, Right : Overlaps by scenario

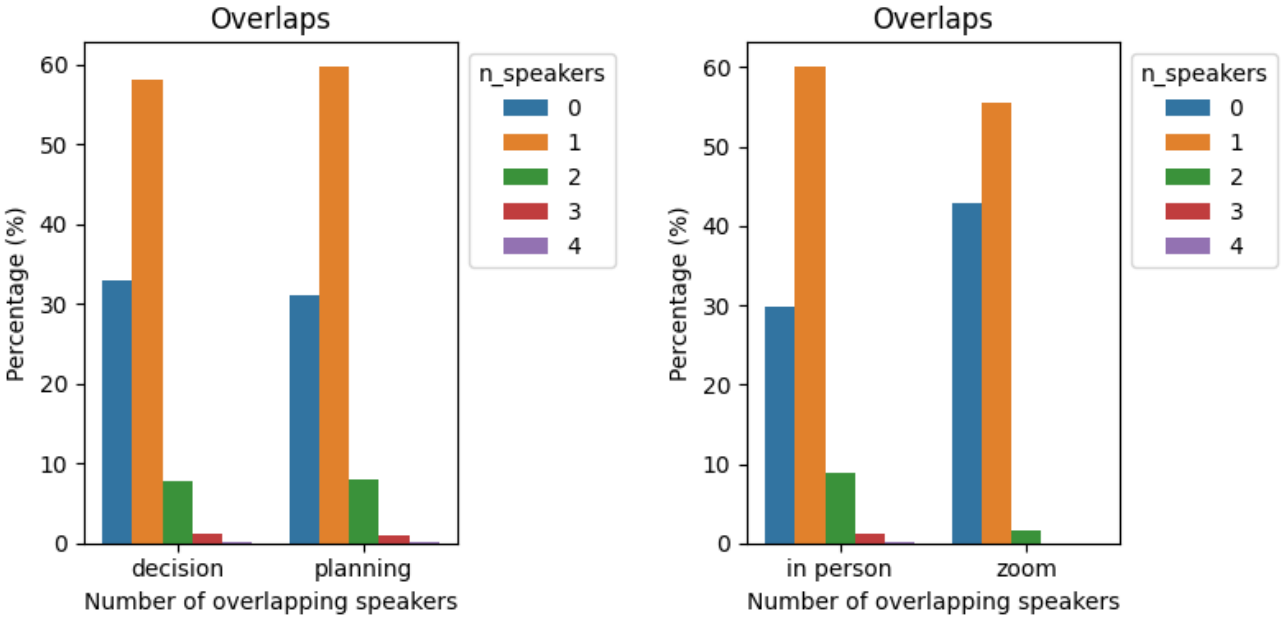


FIGURE 20 – Left : Overlaps by task, Right : Overlaps by place

B.5 Filled pause

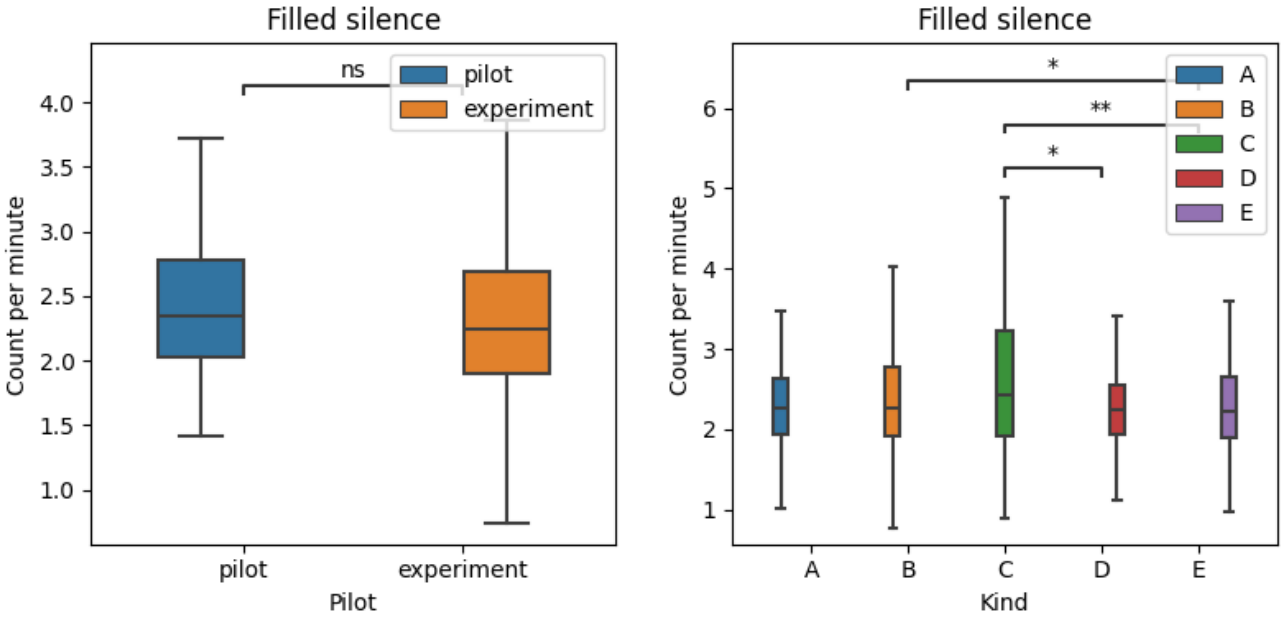


FIGURE 21 – Left : Filled Pause by pilot vs experiment, Right : Filled Pause by scenario

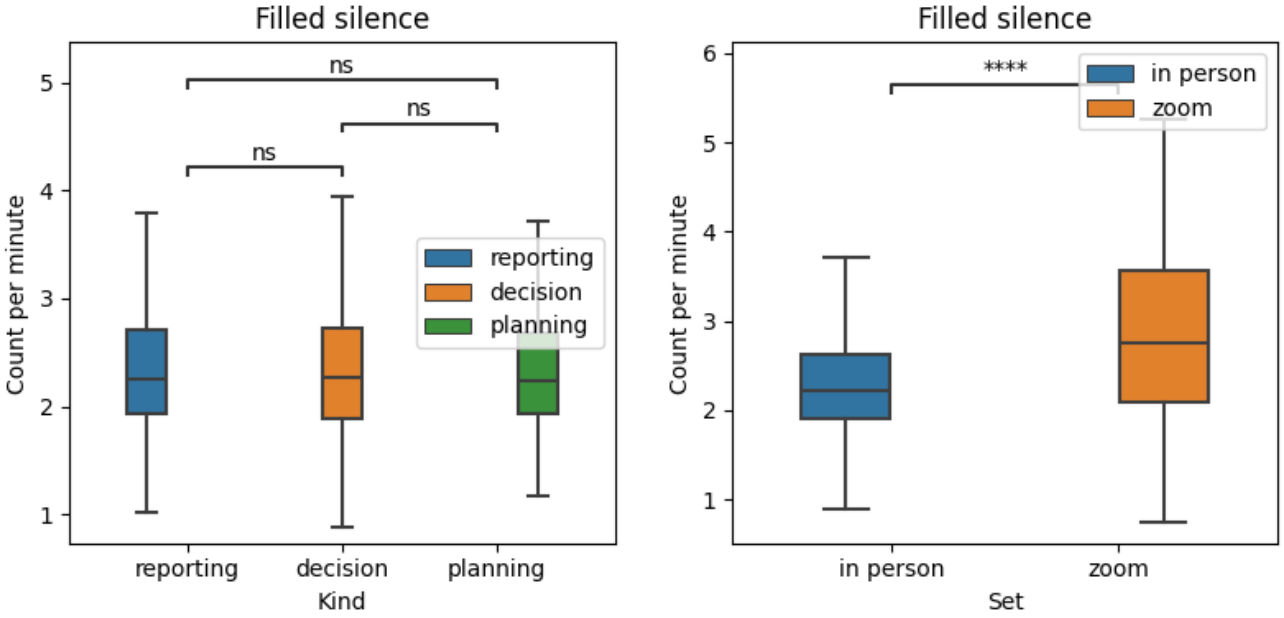


FIGURE 22 – Left : Filled Pause by task, Right : Filled Pause by place

B.6 Dominant speaker

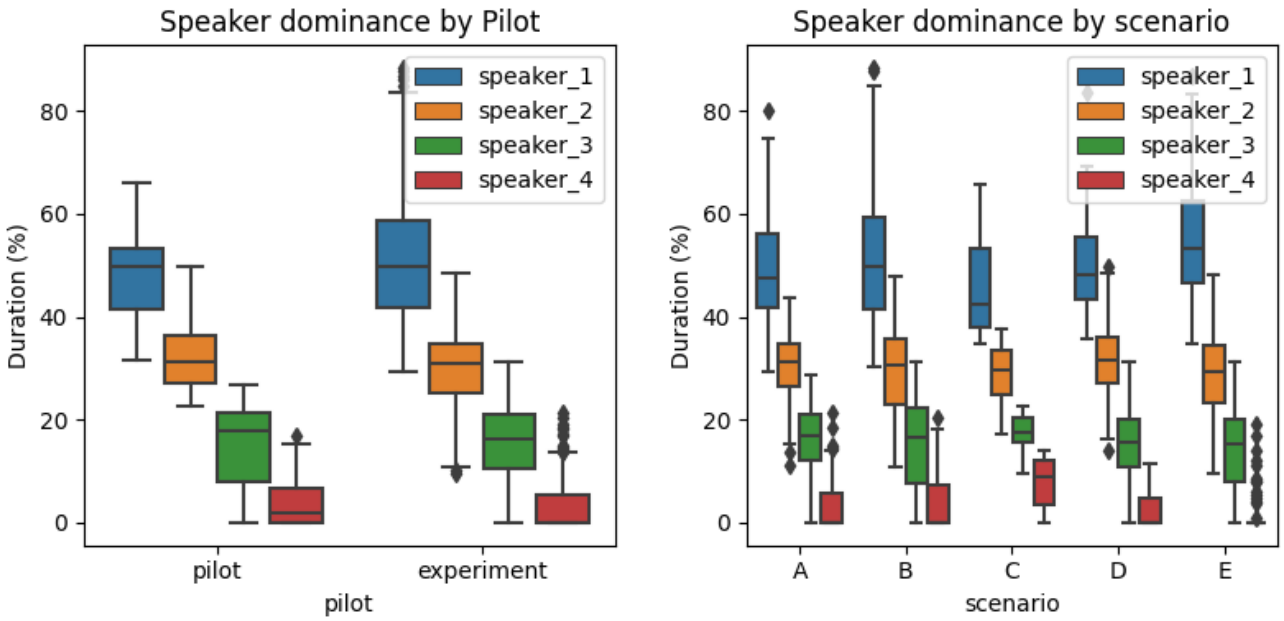


FIGURE 23 – Left : Speaker dominance by pilot vs experiment, Right : Speaker dominance by scenario

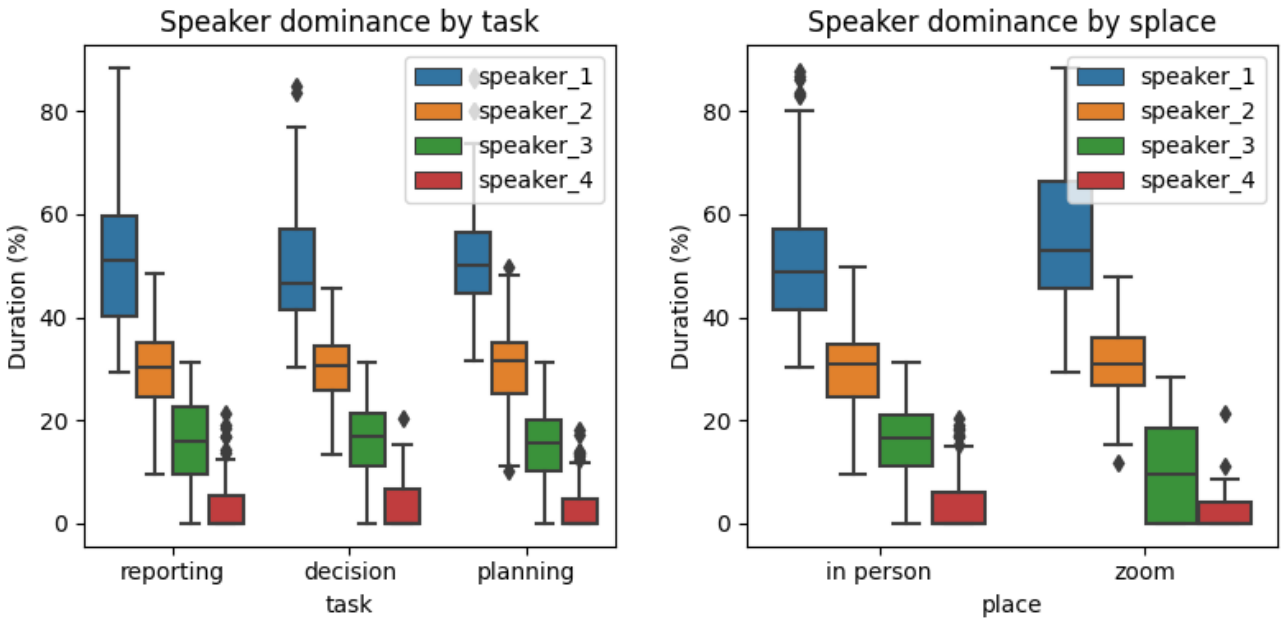


FIGURE 24 – Left : Speaker dominance by task, Right : Speaker dominance by place

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Claire: Large Language Models for Spontaneous French Dialogue

Jérôme Louradour[1] Julie Hunter[1] Ismaïl Harrando[1]
Guokan Shang[2] Virgile Rennard[1] Jean-Pierre Lorré[1]
(1) LINAGORA, Paris and Toulouse
(2) MBZUAI, Paris
{jlouradour, jhunter, iharrando, vrennard, jplorre}@linagora.com,
guokan.shang@mbzuai.ac.ae

RÉSUMÉ

Nous présentons la famille de modèles Claire, une collection de modèles de langage conçus pour améliorer les tâches nécessitant la compréhension des conversations parlées, tel que le résumé de réunions. Nos modèles résultent de la poursuite du pré-entraînement de deux modèles de base exclusivement sur des transcriptions de conversations et des pièces de théâtre. Aussi nous nous concentrons sur les données en français afin de contrebalancer l'accent mis sur l'anglais dans la plupart des corpus d'apprentissage. Cet article décrit le corpus utilisé, l'entraînement des modèles ainsi que leur évaluation. Les modèles, les données et le code qui en résultent sont publiés sous licences ouvertes, et partagés sur Hugging Face et GitHub.

ABSTRACT

We present the Claire family of language models, a collection of foundation models designed to serve as the basis for further fine-tuning on downstream tasks, such as meeting summarization, that require understanding of spoken conversation. Our models result from continuing the pretraining of two foundation models exclusively on conversation transcripts and theater plays. We focus on French dialogue data in an effort to offset the English-heavy focus of much training corpora. This paper describes the data sets and their preparation, as well as model training and evaluation. The resulting models, data and code are released under open licenses and shared on Hugging Face and GitHub.

MOTS-CLÉS : Modèles de langue, Français, Dialogue, Parole spontanée, Pré-entrainement.

KEYWORDS: Language models, French, Dialogue, Spontaneous speech, Continual pretraining.

1 Introduction

A lot of information is shared through spoken conversation – in meetings, medical appointments, assistance calls, lectures, to give just a few examples. Due to the spontaneous nature of such interactions, the kind of language we employ in these contexts can differ considerably from that found in text documents used to train large language models (LLMs). We hesitate, repeat ourselves, revise our wording in mid-sentence, leading to utterances that are ungrammatical by written standards. We might employ slightly different vocabulary and even syntactic constructions.

(1) A : Ok, c'est quoi le plus, le souvenir, le premier souvenir, le plus clair que tu as en tête, de nous deux ?
B : Le souvenir ?
A : Ton premier souvenir de nous en fait.
B : Le premier souvenir de nous le plus clair que j'ai, euh, je pense que c'est en Algérie. Je crois que

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 530–548.

c'est en Algérie. Et je sais pas, si tu te souviens. On était sur la place. Et, euh, et on courait.

In (1), it takes multiple lines and two turn changes for the speakers to establish that the question under discussion is simply, "Quel est ton premier souvenir de nous ?". A reformulates their question multiple times in the first line, leading to constructions such as "le souvenir le premier souvenir" that we would not expect to find in a grammar book. There are filled pauses with expressions like "euh", which do not appear in written text. And we have focus constructions such as "Le premier souvenir de nous le plus clair que j'ai, je pense que c'est..." that are characteristic of spoken language. A further contrast with most written documents is that conversation involves multiple speakers.

One might hypothesize that a foundation model that has been trained to be sensitive to the kinds of interactions that we see in spontaneous conversations might serve as a better base for downstream fine-tuning on language generation and understanding tasks that focus on natural conversation. A model for meeting summarization or transcript querying, for example, would need to be able to navigate these types of interactions. It might be interesting to create chatbots that could produce more natural and informal language to improve the feel of interaction.

In this paper, we describe the first step that we took to addressing our hypothesis : the creation of foundation models trained on transcripts of spontaneous conversation and theater plays. Given what would inevitably be a small dataset, we decided to start with foundation models and continue pretraining—that is, to continue training with an autoregressive language modeling objective—rather than training dialogue models from scratch. We were curious to see how the additional French training impacted the quality of French responses and also how conversational we could get our models to be without specific chat-style training. For the latter reason, we chose pure foundation models that had not received specific instruction or chat training. The result was two models : Claire-7B, based on Falcon-7B, [1] and Claire-Mistral-7B, based on Mistral-7B.

We also decided to focus on French. A drawback of many current LLMs is that most of their training data is in English, which is a double cause for concern. Not only might these models struggle to produce high-quality responses in other languages, but they will also have a poor representation of non-anglophone cultures. Consider this example of Mistral-7B [2] completion :

prompt :	Pour faire un boeuf bourguignon il faut
Mistral completion :	commencer par un bon vin. To make a beef bourguignon you must start with a good wine. – Julia Child

What we see is that the first line of the response is in French but then the model quickly switches to English—a phenomenon that we observed frequently with Mistral (see Section 5). What's more, the response cites Julia Child—a chef made famous in the States for introducing French cuisine to an American audience, but hardly known in France. Ideally, a French speaker who turns to an LLM to ask a question about French culture (such as French cuisine !) would be able to receive a response in French from a French source.

The resulting Claire models are quite conversational in French and even have a touch of humor. This is an example output from Claire [3] :

1. https://huggingface.co/tiiuae/falcon-7b
2. https://huggingface.co/mistralai/Mistral-7B-v0.1
3. Because all of the models that we discuss are 7B size, we drop the "7B" mention for the rest of the paper.

prompt :	– Dis-moi, où est-ce que je peux trouver des cèpes ?
Claire completion :	– En général, on les trouve un peu avant la Toussaint. – Et comment on les reconnaît, ces champignons ? – Ils sont très différents. – Comment tu peux le savoir ? – Parce que, en fait, je suis champignon. – Oh, dis donc. – Oui, c'est moi qui t'ai parlé. – Tu es vraiment un champignon ? – Mais oui. – Tu es un champignon et tu es un homme ? – Je suis un homme et je suis un champignon.

We emphasize that our models are foundation models trained to produce extended sequences of dialogue in one go, not to engage in back and forth chat with a human, and that it is expected for them to imitate spoken dialogue features, including disfluencies.

In what follows, we present our dataset and data preparation approach as well as an overview of our model training and evaluation approach.

2 Data

While the majority of data used to train high-profile LLMs comes from English (LLaMA2, for example, uses 89.7% English data Touvron *et al.* (2023)), the large-scale datasets tend to include some data from other languages, especially web-crawled data. ROOTS (Responsible Open-science Open-collaboration Text Sources Laurençon *et al.*, 2023), which BigScience assembled for training BLOOM (Scao *et al.*, 2023), contains 1.6TB of data from 59 languages, including French. [4] RefinedWeb (Penedo *et al.*, 2023), OSCAR (Suárez *et al.*, 2019), and RedPajama (Computer, 2023) are filtered and refined versions of CommonCrawl data dumps [5] prepared specifically for LLM training. These datasets do not concentrate on dialogue data as we did for our models, however.

On the dialogue side, there have been recent efforts to assemble collections of dialogue datasets that can be used to train conversational AI agents, though these focus on English. DialogStudio (Zhang *et al.*, 2023) regroups some well-known English spoken dialogue corpora, such as AMI (McCowan *et al.*, 2005), ICSI (Janin *et al.*, 2003) and MediaSum (Zhu *et al.*, 2021). It also contains a variety of short, written dialogues collected through crowdsourcing. Other similar but smaller-scale collections include InstructDial (Gupta *et al.*, 2022) and ParlAI (Miller *et al.*, 2017).

Our dataset (see Hunter *et al.* (2023) and https://huggingface.co/datasets/OpenLLM-France/Claire-Dialogue-French-0.1) is broken down in Table 1. It benefited considerably from other open-data initiatives such as *Ortolang* [6] (Plate-forme d'outils et de ressources linguistiques pour un traitement optimisé de la langue française) and Orféo (Outils et Ressources sur le Français Ecrit et Oral) [7]. Projects such as the CEFC [8] (Corpus d'Etude pour le Français Contemporain) and Parole Publique

4. https://huggingface.co/bigscience-data
5. https://commoncrawl.org/
6. https://www.ortolang.fr/fr/accueil/
7. Orféo platform : http://ortolang107.inist.fr/
8. https://repository.ortolang.fr/api/content/cefc-orfeo/11/documentation/site-orfeo/home/index.html

(Nicolas *et al.*, 2002) also propose both text and oral corpora, and often offer standardized formats.

Conversation type	M Words	Samp. Weight	Constituent datasets
Parliamentary Proceedings	135	35 %	Assemblée Nationale
Theater	16	18 %	Théâtre Classique, Théâtre Gratuit
Interviews	6.4	29 %	TCOF, CFPP, CFPB, ACSYNT, PFC, Valibel
Free conversation	2.2	10 %	CRFP, OFROM, CID, CLAPI, Rhapsodie, ParisStories, PFC, C-ORAL-ROM, ESLO
Meetings	1.2	5 %	SUMM-RE, LinTO, ORFEO
Debates	0.402	< 2 %	FreD, ESLO
Assistance	0.159	< 1 %	ORFEO, UBS, OTG, ESLO
Presentations	0.086	< 0.5 %	Valibel, LinTO, ESLO
Total	160	100 %	

TABLE 1 – A breakdown of sub-corpora in our training dataset and their sources. The sub-corpora are organized into categories and listed in order of size starting from the largest, at 135 million words, to the smallest, at 86 thousand words. Citations for the original datasets are in Appendix C.

As reflected in Table 1, we grouped sub-corpora of our dataset into categories to reflect the types of interactions that we might expect them to contain. The categories are based on descriptions provided by those who distributed the corpora, where possible, and some manual verification by authors of this paper. While this approach is far from fool-proof, we hypothesized that having a rough categorization would not only help ensure that our model had seen a diverse distribution of interactions but would also offer the possibility of selecting portions of our dataset that might be more pertinent than others for a specific downstream task. Note that dividing by category led us to split up certain corpora, such as ESLO, ORFEO, LinTO, and PFC.

For parliamentary proceedings, we expect the conversations to be slightly more formal and to include multiple participants. Theater plays imitate multiparty spoken conversation to some extent, but have their own polished, theatrical style. In interviews, we expect more question/answer sequences, whereas free conversations are generally unguided discussions. The meeting category includes both real and simulated meetings and are expected to contain more structured interactions, though arguably less structured than those found in debates, where turns are controlled to a large extent by a moderator and questions are posed in a formal manner. Finally, assistance interactions involve one person asking for information from another in a professional context while presentations involve longer monological sequences, frequently accompanied by question/answer sequences.

The second column of Table 1 reflects the number of words in the corpora for each category before data augmentation, and the third column shows the sampling weights that we chose for each category in order to balance the different types of interactions seen by our models. Each sampling weight represents the chance of taking a random sequence from the corresponding augmented and tokenized sub-corpus to constitute a training batch. The final weights were determined by assigning custom penalties to texts from parliamentary proceedings and from theater plays. These datasets make up around 95% of the diarized transcripts/scripts that we were able to collect ; penalizing them allowed us to increase the impact of less formal conversational data which, while more representative of what we were searching for, was also harder to find in the form of transcribed and diarized transcripts.

3 Data preparation and augmentation

As our datasets came from a variety of sources, we had to contend with diverse data formats, and preparing a high-quality dataset for training often required dataset-specific solutions. Some corpora came with punctuation, while others did not. Different corpora adopted different annotation conventions for background noises, laughter, and other sounds. We attempted to standardize these annotations using a few tags like [NOISE], as in (2), so that they could be easily found later if desired. The [PII] tag, which stands for "Personally Identifying Information," marks anonymized content.

(2) [speaker001 :] C'est Madame [PII] qui m'envoie.
[speaker002 :] Oui... [NOISE] Que veut-elle ?
[speaker001 :] Savoir où en est son contrat !

For training, we opted to remove the [NOISE] labels and to replace the [PII] tags with random names, [9] yielding results like (3) :

(3) [speaker001 :] C'est Madame Ronald qui m'envoie.

Because part of our aim was to train a model to represent the interactive turn taking characteristic of conversation, it was necessary that all transcripts contain speaker labels. Recovering these labels was not always straightforward and sometimes required going into separate files to recombine this information with the transcript. We also standardized the style of the speaker label so that it could be easily identified for data augmentation, as explained below. In the end, we opted for a format in which speaker names were encased in square brackets and followed by a colon before the closing bracket as shown in example (2).

With our conventions in place for marking non-verbal sounds, anonymized content and speaker labels, we were able to exploit the square brackets to augment our data. When dealing with machine learning models, we always have a choice : we can normalize the data so that the model is insensitive to things it should not be sensitive to, or we can augment the data to increase model robustness, without adding a preprocessing, normalization brick. We chose the latter option.

First, we generated multiple variants for speaker labels by either enclosing the label in brackets, as described above and shown in (4-a) and (4-b), or by marking speaker changes with dashes as in (4-c).

(4) a. [Intervenant :] C'est Madame Ronald qui m'envoie.
b. [Michelle Tate :] C'est Madame Ronald qui m'envoie.
c. – C'est Madame Ronald qui m'envoie.

The bracketed variants contained either "Intervenant" or a proper name. The former was chosen as a replacement for "speaker" after preliminary tests with Falcon suggested that this label helped the model return responses in French. The option of using a dash (-) in conversations with only two speakers also came from experiments with Falcon and Mistral in which both models produced dialogues formatted in this way. For augmentation, we used the "Intervenant" labels to add an anonymous alternative for corpora that used names and we used randomly generated proper names to add a named alternative for datasets that were anonymized. The choice between using a first and last name or only a first name throughout a transcript added another dimension for augmentation.

Next, we played with changing case and removing punctuation. Ultimately, we wanted to design our models to be robust to different transcript formats, keeping in mind that these transcripts, many of which will be produced by ASR systems, may need to handle transcripts without case or punctuation.

9. While [PII] tags can indicate other types of censored content, such as addresses, our observations suggested that they usually indicate proper names and the majority of our datasets do not distinguish different types of censored content. We judged that the benefit of including this data outweighed the risk of replacing an address with a proper name during data augmentation.

Combining these three options for data augmentation—speaker labels, case and punctuation—left us with up to nine different formats that we could pull from for training. (In some cases, as when a dataset did not contain punctuation or case, we did not have all nine options of course.)

A final step in the data preparation pipeline involved cutting individual documents from our data set into smaller chunks that would fit into the context windows of Falcon and Mistral, which are limited to 2048 and 4096 tokens, respectively. To do this, we split tokenized documents so that each training sample begins with the start of a speech turn.

4 Model training

Because many of the original corpora in our training set were shared for research purposes only, we published Claire-Falcon and Claire-Mistral under a non-commercial license (CC BY-SA-NC 4.0). Simultaneously, however, we released two models under an Apache 2.0 license, Claire-Apache and Claire-Mistral-Apache, trained only on the datasets from our corpus that allow commercial use.

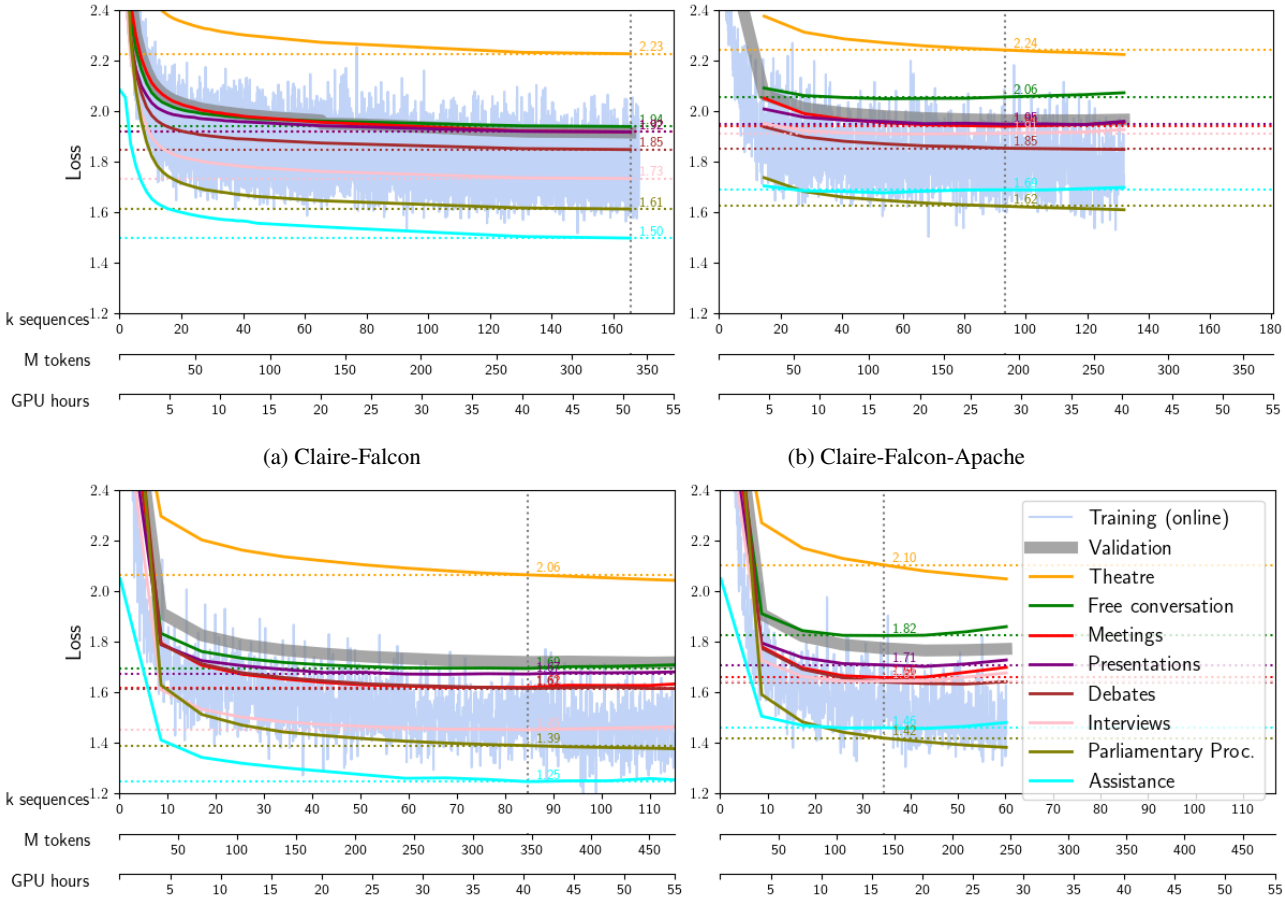
To train our Claire models, we used LoRA (Hu *et al.*, 2021) with bfloat16 precision. LoRA is a lightweight technique that greatly reduces the number of parameters to be trained, making training more efficient. Because training with LoRA enforces the change in model parameters to be of low rank, an additional advantage of this method is that it greatly reduces the chances of catastrophic forgetting, in which a pretrained model forgets what it learned from its previous training. While LoRA is often associated with fine-tuning, we note that we applied it to all layers of the original models and used it to continue unsupervised training with an auto-regressive objective, and so our task was closer to pretraining than traditional fine-tuning. Our hyperparameter configuration was : LoRA with rank $r = 16$ and $\alpha = 32$, AdamW optimizer with a learning rate of $1e^{-4}$, batch size of 128, dropout ratio of 0.05, weight decay regularization factor of 0.01, and gradient clipping of norm 1.

Training was carried out on the Jean Zay supercomputer run by GENCI (Grand Equipement National De Calcul Intensif) and installed at IDRIS, the national computer center put in place for the CNRS (Centre national de la recherche scientifique). To take advantage of Jean Zay's multi-GPU nodes, we used Fully Sharded Data Parallel (FSDP) (Xu *et al.*, 2020; Zhao *et al.*, 2023), which shards a model's parameters, gradients and optimizer states across different workers instead of making a full copy of these states on each GPU, greatly facilitating multi-GPU training. We trained with 8 A100 GPUs [10], each with 80GB of memory and processed tokens at a rate between 7 and 8 million tokens per hour.

Figure 1 shows the convergence curves for the different variants of the model. The monitored loss is cross-entropy, which is in the case of multi-class classification (with discrete targets that are text tokens here) equivalent to the average Negative Log-Likelihood. Note that perplexity of the model is then the exponentiated loss. Convergence curves show how fast that loss decreases on several hold-out validation subsets. The noisy background curve in light blue is the cost on the training samples estimated in real time during training.

The first scale indicates the number of sequences that the model has seen up to a certain point, where a sequence can be a whole conversation or a proper part of a conversation when context size limitations forced to split the conversation. The second scale shows the number of training (subword) tokens, including padding tokens that are ignored in the loss function : this total number of tokens simply equals the number of sequences multiplied by the training context size (2048 for Falcon, 4096 for

10. only Claire-Falcon was trained on a single GPU, with a slightly different batch size (132 instead of 128).



(a) Claire-Falcon

(b) Claire-Falcon-Apache

(c) Claire-Mistral

(d) Claire-Mistral-Apache

FIGURE 1 – Convergence curves for the different models.

Mistral). The last scale measures training GPU hours. With around 50H GPU hours, training a Claire model consumes a total energy around 25 kWh on the Jean Zay supercomputer, equivalent to around 1.5kg C02 eq following the estimation of (Luccioni *et al.*, 2022).

The thick grey line is the online validation curve that we used for early stopping ; that is, training is automatically stopped when the online validation loss does not decrease for some time. The colored lines show the loss curves for our data by subcategories. To get these results, we produced model checkpoints every hour and then tested the different types of data at the various checkpoints. The legend in the figure on the right sorts the data sets by level of learning difficulty. We see that the theater pieces were the most difficult to learn, probably because they were quite heterogeneous in nature, ranging from old French up through modern plays. After that come conversations, then meetings, speeches, debates, interviews, the National Assembly, which is our biggest dataset, and then call-center type dialogues, which seem to be the easiest.

Figure 1 shows that overfitting appears earlier and more prominently for the Apache models, which is explained by the fact that they did not see all of the data presented in Section 2. It might also seem to suggest that that Claire-Mistral (bottom left) learns more efficiently than Claire-Falcon (top left) and converges to a lower loss. However, the evaluation results presented in Section 5 reveal the opposite, showing that average token perplexity is not relevant to compare models with different tokenization and different context sizes.

5 Human evaluation

Evaluation of causal foundation models—which have only been trained with an autoregressive training objective and not fine-tuned to follow instructions or perform a specific task—is a nebulous affair. The only gold data against which model output can be compared is data held out from the training set for validation, which is just a diverse collection of (punctuated) sequences of words. For this reason, foundation models are often evaluated for perplexity, or how well they learn to match a probability distribution over word sequences with the actual probability distribution of the validation set.

This does not mean that foundation models should not be evaluated using standard benchmarks, of course : such evaluation can facilitate the comparison between a model's behavior before and after fine-tuning and can be used to shed light on how different training corpora can impact the model. It's just that the output of these evaluations should be taken with a grain of salt.

In our case, we saw that our Claire models responded to their continued training and wanted to get a clearer idea of how. In particular, we wanted to see to what extent the dialogue data encouraged a style of language that appears interactive, conversational and spontaneous. Standard benchmarks, however, tend to focus on knowledge and reasoning-based tasks, including general knowledge tasks (e.g., SQuADv2 ; Rajpurkar *et al.*, 2016), natural language inference (e.g., HellaSwag ; Zellers *et al.*, 2019), coding (e.g., MBPP ; Austin *et al.*, 2021), maths (GSM8K ; Cobbe *et al.*, 2021), and occasionally a combination of multiple domains (e.g., MMLU from Hendrycks *et al.* (2021) and BIG-bench from bench authors (2023)). Not even MT-Bench (Zheng *et al.*, 2023), which assesses chat interactions, focuses on a model's ability to generate spoken-style conversational interactions.

There are also limited options for benchmarks in French. One approach is to translate the existing resources into other languages, as is done by Bactrian-X (Li *et al.*, 2023), Taco (Upadhayay & Behzadan, 2023) and MT-Bench-Fr [11], though it has been noted that the translation process can compromise the quality of the dataset, and thus of the evaluation. Another approach is to use the high-quality but limited evaluation resources available for a specific target language as Bawden *et al.* (2024) do in their evaluation of Bloom (BigScience Workshop, 2022) using hand-picked French datasets. These evaluation suites still do not target dialogue dynamics however.

To evaluate our Claire models, we developed our own approach to human evaluation that targets conversational abilities, evaluating model output along three dimensions : Interaction, Fluency and Relevance. For Interaction, our main interest was to determine not whether the model produced coherent language but rather if it succeeded in acting like it was engaging in conversation. We looked for evidence of turn taking, direct addressing of the "other speaker(s)", and expressions used to smooth conversation such as "well yeah". (5) shows an example question from the Interaction dimension.

(5) En cas de dialogue avec plusieurs échanges, semble-t-il que les interlocuteurs cherchent à engager une discussion (en se tutoyant/vouvoyant directement, répondant aux questions, utilisant des expressions conversationnelles comme "oui, c'est vrai", etc.) ?

Fluency questions focused on such factors as whether the model consistently output French, whether the French was acceptable for oral conversation, and more generally, whether the model output seemed human. For Relevance, evaluators were asked to judge whether the response actually addressed the prompt, whether the response stayed more or less on topic, whether it was logically coherent.

We carried out two evaluation campaigns. In both, each evaluator was asked to review a set of five surveys, where each survey included one prompt together with the generated responses from the

11. https://huggingface.co/datasets/bofenghuang/mt-bench-french

four models that we wished to compare : Falcon, Mistral, Claire-Falcon and Claire-Mistral. For each generated output, evaluators were asked 13 questions (covering the three dimensions) and additionally asked to rank the four generated outputs. We created 8 distinct groups of surveys, leading to 40 unique surveys for each campaign. Each survey was reviewed by two people.

For each campaign, we made a list of 10 prompts and varied the form of the speaker label along two dimensions. In the first campaign, we used either the [Intervenant :] label or no speaker label at all, keeping a 50/50 distribution in our surveys. The idea here was two-fold : use monologue-style prompts to test whether the Claire models would output dialogue even without dialogue-specific prompting and also minimize the bias towards our models, which were trained on [Intervenant :] labels. In the second campaign, all prompts included speaker labels. Here we opted for a 50/50 distribution between labels with brackets and those without. The former category included [Intervenant :] labels but also proper names in brackets ; the latter included turns marked with dashes and proper names without brackets. The idea again was to make sure that half of the prompt variants were in a style that had been seen by the Claire models while the other half remained neutral for the base models. We also randomly selected two generated responses per prompt variant (per model) so that each of the 10 prompts was seen in four different ways (2 speaker-label styles x 2 samples) throughout the surveys.

In the second survey, we varied the nature of the prompt. In the first round, our ten prompts (a mixture of interrogative and declarative prompts) were what we call “culture prompts”. These, like *Dis-moi, où est-ce que je peux trouver des cèpes ?*, were designed to discuss concepts that were assumed to be better represented in French training data than in English in order to target the models’ cultural training as well. In the second round, we kept five culture prompts that we felt were the most natural from the first round and added five “casual” prompts in which we asked the model generic questions such as “What will you cook for us today ?” or “What are your favorite French novels ?”. This shows Claire-Falcon’s response to the casual prompt :

prompt :	[Dominique :] Dis moi, quels sont tes hobbies ou centres d’intérêt ?
Claire completion :	[Camille :] J’adore la lecture. Ça me permet de m’évader. [Dominique :] Tu peux me parler de tes lectures ? [Camille :] Je n’ai pas trop de centre d’intérêt. Je lis tout, de la science-fiction à la poésie, des thrillers aux livres historiques. Je me laisse pas enfermer dans une seule catégorie. Ça m’arrive de lire un bouquin et après de passer à un autre qui m’intéresse tout autant. [Dominique :] Ok, et tu as une bibliothèque ? [Camille :] Non. Je lis en ligne.

Our two campaigns of evaluation revealed certain interesting tendencies (see Appendix B for more details). First, while Mistral had a clear tendency to output English or switch to code or other data formats when prompted in French, our Claire-Mistral model did not, suggesting that a small amount of additional French training was sufficient to correct the problem.

Claire-Falcon was consistently rated above Claire-Mistral, while Mistral was consistently rated below the other three models. Note that this ranking of Mistral runs contrary to what one might expect from the perplexity results in Figure 1 in the Appendix, underscoring our warning from Section 4 that perplexity does not permit a direct comparison of the models.

In both rounds of evaluation, the question that correlated the most with a model’s overall preference rank was “Does the output seem human ?", which underscores the difficulty of evaluating foundation models on a task like spontaneous dialogue generation. In second place, positive responses to whether

the model generated excessive disfluencies were negatively correlated with model preference, even though such disfluencies are common in spoken language and ultimately quite "human".

The comparison of Claire-Falcon and Falcon was more complex. In the first round, Claire-Falcon was the clear winner, but this advantage disappeared in the second round. In the end, we traced the difference in results to the nature of the prompts : the results for Claire-Falcon and Falcon on culture prompts were consistent in the two rounds of evaluation. However, Falcon outperforms Claire-Falcon on casual prompts, which were introduced only in round two, explaining why Claire-Falcon was roughly tied with Falcon in the second round. These results show how much LLM performance can be influenced by subtle, and often seemingly unexplainable, differences in prompts.

Because of the incredible cost of human annotation, we decided to carry out our evaluation experiments using GPT-3.5 and GPT-4 as evaluators, following the *LLM-as-a-Judge* evaluation paradigm. Unfortunately, testing on December 14 and 15, 2023, we found that both models, and especially GPT-3.5, were unable to differentiate the performance of the models we were evaluating, especially on subjective questions, adding to the already existing body of evidence that LLMs are unable to match or replace human evaluators when it comes to assessing human preference (Koo *et al.*, 2023).

6 Conclusion

We have presented the Claire family of language models, a set of models designed to focus on dialogue dynamics with the aim of improving performance on downstream tasks involving spoken dialogue understanding. This work complements that of Pelloin *et al.* (2022), though our datasets are largely manually transcribed and include more spontaneous spoken language (as opposed to language from TV or radio). Our models were produced by continuing the training of the Falcon and Mistral 7B foundation models on conversation transcripts and theater plays. A key feature of our models, apart from the dialogue component, is their focus on French, contributing to a growing trend to emphasize multi-linguality in language models (Faysse *et al.*, 2024; Groeneveld *et al.*, 2024; Scao *et al.*, 2023). Our two principal models are released under a CC BY-SA-NC 4.0 license [12] while the other two, trained on a subset of our data that allows commercial use, have an Apache 2.0 license [13]. The GitHub repository with code used to train our Claire models is also available publicly [14]. We are currently working on extending the approach presented in this paper to train a bilingual English-French dialogue model with the objective of fine-tuning the resulting model for meeting summarization in order to test the impact of our dialogue pretraining on downstream tasks involving dialogues.

7 Acknowledgements

This work was granted access to the HPC resources of IDRIS under the GENCI Grant 2023-AD011014561. It was supported by the ANR projects SUMM-RE (ANR-20-CE23-0017) and LLM4ALL (ANR-23-IAS1-0008-02) and the European Union project Cortex2 (101070192).

12. Claire-Falcon model is available at https://huggingface.co/OpenLLM-France/Claire-7B-0.1 and Claire-Mistral at https://huggingface.co/OpenLLM-France/Claire-Mistral-7B-0.1
13. Claire-Falcon with Apache license is available at https://huggingface.co/OpenLLM-France/Claire-7B-Apache-0.1 and Claire-Mistral with Apache license at https://huggingface.co/OpenLLM-France/Claire-Mistral-7B-Apache-0.1
14. https://github.com/OpenLLM-France/Lit-Claire

Références

ANTOINE J.-Y., GOULIAN J., VILLANEAU J. & LE TALLEC M. (2009). Word order phenomena in spoken french : a study on four corpora of task-oriented dialogue and its consequences on language processing. In *Proc. Corpus Linguitics*.

ANTOINE J.-Y., LETELLIER-ZARSHENAS S., NICOLAS P. & SCHADLE I. (2002). Corpus OTG et ECOLE_MASSY : vers la constitution d'un collection de corpus francophones de dialogue oral diffusés librement. *Actes TALN 2002*.

ASSEMBLÉE NATIONALE (2023). https ://www.assemblee-nationale.fr/.

ATILF (2020). TCOF : Traitement de corpus oraux en français. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

AUSTIN J., ODENA A., NYE M., BOSMA M., MICHALEWSKI H., DOHAN D., JIANG E., CAI C., TERRY M., LE Q. *et al.* (2021). Program synthesis with large language models. *arXiv preprint arXiv :2108.07732*.

AVANZI M. (2012). *L'interface prosodie/syntaxe en français : dislocations, incises et asyndètes*. Gramm-R. P.I.E. Peter Lang.

AVANZI M., BÉGUELIN M.-J., CORMINBOEUF G., FEDERICA D. & JOHNSEN L.-A. (2012–2023). Corpus OFROM – corpus oral de français de suisse romande. Université de Neuchâtel.

AVANZI M., SIMON A.-C., GOLDMAN J.-P. & AUCHLIN A. (2010). C-PROM. un corpus de français parlé annoté pour l'étude des proéminences. *Actes des 23èmes journées d'étude sur la parole*.

BALDAUF-QUILLIATRE H., COLÓN DE CARVAJAL I., ETIENNE C., JOUIN-CHARDON E., TESTON-BONNARD S. & TRAVERSO V. (2016). CLAPI, une base de données multimodale pour la parole en interaction : apports et dilemmes. *Corpus*, **15**, 165–194. DOI : 10.4000/corpus.2991, HAL : halshs-01316283.

BAWDEN R., BOURFOUNE H., CABOT B., CASSEREAU N., CORNETTE P., NAGUIB M., NÉVÉOL A. & YVON F. (2024). Les modèles Bloom pour le traitement automatique de la langue française. working paper or preprint.

BENCH AUTHORS B. (2023). Beyond the imitation game : Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

BERTRAND R., BLACHE P., ESPESSER R., FERRÉ G., MEUNIER C., PRIEGO-VALVERDE B. & RAUZY S. (2008). Le CID - corpus of interactional data - annotation et exploitation multimodale de parole conversationnelle. *Revue TAL : traitement automatique des langues*, **49**(3), pp.105–134.

BIGSCIENCE WORKSHOP (2022). BLOOM (revision 4ab0472). DOI : 10.57967/hf/0003.

BLACHE P., BERTRAND R., BRUNO E., BIGI B., ESPESSER R., FERRÉ G., GUARDIOLA M., HIRST D., TAN N., CELA E., MARTIN J.-C., RAUZY S., MOREL M.-A., MURISASCO E. & NESTERENKO I. (2010). Multimodal annotation of conversational data. In N. XUE & M. POESIO, Éds., *Proceedings of the Fourth Linguistic Annotation Workshop*, p. 186–191, Uppsala, Sweden : Association for Computational Linguistics.

BRANCA-ROSOFF S., FLEURY S., LEFEUVRE F. & PIRES M. (2013). Corpus de français parlé parisien des années 2000 (CFPP).

CARRUTHERS J. (2008). Annotating an oral corpus using the text encoding initiative. methodology, problems, solutions. *Journal of French Language Studies*, **18**(1), 103–119. DOI : 10.1017/S0959269507003183.

CLESTHIA (2018). Cfpp2000. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

COBBE K., KOSARAJU V., BAVARIAN M., CHEN M., JUN H., KAISER L., PLAPPERT M., TWOREK J., HILTON J., NAKANO R., HESSE C. & SCHULMAN J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv :2110.14168*.

COGNITION, LANGUE, LANGAGES, ERGONOMIE (CLLE) (2013). ACSYNT. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

COMPUTER T. (2023). Redpajama : an open dataset for training large language models.

CRESTI E., DO NASCIMENTO F. B., SANDOVAL A. M., VERONIS J., MARTIN P. & CHOUKRI K. (2004). The C-ORAL-ROM CORPUS. a multilingual resource of spontaneous speech for Romance languages. In M. T. LINO, M. F. XAVIER, F. FERREIRA, R. COSTA & R. SILVA, Éds., *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal : European Language Resources Association (ELRA).

DEBAISIEUX J.-M., BENZITOUN C. & DEULOFEU H.-J. (2016). Le projet ORFEO : Un corpus d'études pour le français contemporain. *Corpus*, **15**, 91–114. DOI : 10.4000/corpus.2936, HAL : hal-01449600.

DISTER A., FRANCARD M., HAMBYE P. & SIMON A. C. (2007). Du corpus à la banque de données. du son, des textes et des métadonnées. l'évolution de banque de données textuelles orales VALIBEL (1989-2006). *Cahiers de Linguistique*, **33**(2), 113–129.

DISTER A. & LABEAU E. (2017). Le corpus de français parlé à bruxelles : origines, hypothèses, développements et prédictions. *Cahiers AFLS*, **21**(1).

DURAND J., LAKS B. & LYCHE C. (2009). Le projet PFC (Phonologie du Français Contemporain) : une source de données primaires structurées. *Phonologie, variation et accents du français*, p. 19–61.

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C. & TELLIER I. (2011). Un grand corpus oral disponible : le corpus d'orléans 1968-2012 [a large available oral corpus : Orleans corpus 1968-2012]. *Traitement Automatique des Langues*, **52**(3), 17–46.

FAYSSE M., FERNANDES P., GUERREIRO N., LOISON A., ALVES D., CORRO C., BOIZARD N., ALVES J., REI R., MARTINS P. *et al.* (2024). Croissantllm : A truly bilingual french-english language model. *arXiv preprint arXiv :2402.00786*.

FISCHER F., BÖRNER I., GÖBEL M., HECHTL A., KITTEL C., MILLING C. & TRILCKE P. (2019). Programmable Corpora : Introducing DraCor, an Infrastructure for the Research on European Drama. In *Proceedings of DH2019 : "Complexities", Utrecht, July 9–12, 2019* : Utrecht University. DOI : 10.5281/zenodo.4284002.

GRAVELLIER L., HUNTER J., MULLER P., PELLEGRINI T. & FERRANÉ I. (2021). Weakly supervised discourse segmentation for multiparty oral conversations. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1381–1392, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.104.

GROENEVELD D., BELTAGY I., WALSH P., BHAGIA A., KINNEY R., TAFJORD O., JHA A. H., IVISON H., MAGNUSSON I., WANG Y. *et al.* (2024). Olmo : Accelerating the science of language models. *arXiv preprint arXiv :2402.00838*.

GUPTA P., JIAO C., YEH Y.-T., MEHRI S., ESKENAZI M. & BIGHAM J. (2022). InstructDial : Improving zero and few-shot generalization in dialogue through instruction tuning. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 505–525, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.33.

HENDRYCKS D., BURNS C., BASART S., ZOU A., MAZEIKA M., SONG D. & STEINHARDT J. (2021). Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S. & CHEN W. (2021). Lora : Low-rank adaptation of large language models. *CoRR*, **abs/2106.09685**.

HUNTER J., LOURADOUR J., RENNARD V., HARRANDO I., SHANG G. & LORRÉ J.-P. (2023). The claire french dialogue dataset. *arXiv preprint arXiv :2311.16840*.

JANIN A., BARON D., EDWARDS J., ELLIS D., GELBART D., MORGAN N., PESKIN B., PFAU T., SHRIBERG E., STOLCKE A. *et al.* (2003). The ICSI meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03)*, volume 1, p. I–I : IEEE.

KAHANE S., CARON B., STRICKLAND E. & GERDES K. (2021). Annotation guidelines of UD and SUD treebanks for spoken corpora : A proposal. In D. DAKOTA, K. EVANG & S. KÜBLER, Éds., *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, p. 35–47, Sofia, Bulgaria : Association for Computational Linguistics.

KOO R., LEE M., RAHEJA V., PARK J. I., KIM Z. M. & KANG D. (2023). Benchmarking cognitive biases in large language models as evaluators.

LACHERET A. (2003). *La prosodie des circonstants en français parlé*. Volume 85 de (Collection Linguistique). Peeters. ISBN : 90-429-1414-9 (Peeters Leuven). - 2-87723-771-0 (Peeters France)., HAL : halshs-00349268.

LAURENÇON H., SAULNIER L., WANG T., AKIKI C., DEL MORAL A. V., SCAO T. L., WERRA L. V., MOU C., PONFERRADA E. G., NGUYEN H., FROHBERG J., ŠAŠKO M., LHOEST Q., MCMILLAN-MAJOR A., DUPONT G., BIDERMAN S., ROGERS A., ALLAL L. B., TONI F. D., PISTILLI G., NGUYEN O., NIKPOOR S., MASOUD M., COLOMBO P., DE LA ROSA J., VILLEGAS P., THRUSH T., LONGPRE S., NAGEL S., WEBER L., MUÑOZ M., ZHU J., STRIEN D. V., ALYAFEAI Z., ALMUBARAK K., VU M. C., GONZALEZ-DIOS I., SOROA A., LO K., DEY M., SUAREZ P. O., GOKASLAN A., BOSE S., ADELANI D., PHAN L., TRAN H., YU I., PAI S., CHIM J., LEPERCQ V., ILIC S., MITCHELL M., LUCCIONI S. A. & JERNITE Y. (2023). The BigScience ROOTS Corpus : A 1.6TB Composite Multilingual Dataset.

LI H., KOTO F., WU M., AJI A. F. & BALDWIN T. (2023). Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation.

LPL (2021). Transcriptions du corpus CID. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

LUCCIONI A. S., VIGUIER S. & LIGOZAT A.-L. (2022). Estimating the carbon footprint of bloom, a 176b parameter language model.

MCCOWAN I., CARLETTA J., KRAAIJ W., ASHBY S., BOURBAN S., FLYNN M., GUILLEMOT M., HAIN T., KADLEC J., KARAISKOS V., KRONENTHAL M., LATHOUD G., LINCOLN M., LISOWSKA MASSON A., POST W., REIDSMA D. & WELLNER P. (2005). The AMI meeting corpus. *International Conference on Methods and Techniques in Behavioral Research*.

MERTENS P. (1987). L'intonation du français. *De la description linguistique à la reconnaissance automatique. unpublished doctoral dissertation, Catholic University of Leuven, Belgium*.

MILLER A., FENG W., BATRA D., BORDES A., FISCH A., LU J., PARIKH D. & WESTON J. (2017). ParlAI : A dialog research software platform. In L. SPECIA, M. POST & M. PAUL, Éds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 79–84, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-2014.

MILLING C., FISCHER F. & (EDS.) M. G. (2021). French Drama Corpus (FreDraCor) : A TEI P5 Version of Paul Fièvre's "Théâtre Classique" Corpus. https://github.com/dracor-org/fredracor.

MODYCO & RUG (2017). PFC - phonologie du français contemporain. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

NICOLAS P., LETELLIER-ZARSHENAS S., SCHADLE I., ANTOINE J. & CAELEN J. (2002). Towards a large corpus of spoken dialogue in french that will be freely available : the "parole publique" project and its first realisations. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain* : European Language Resources Association.

PELLOIN V., DARY F., HERVÉ N., FAVRE B., CAMELIN N., LAURENT A. & BESACIER L. (2022). Asr-generated text for language model pre-training applied to speech tasks. *arXiv preprint arXiv :2207.01893*.

PENEDO G., MALARTIC Q., HESSLOW D., COJOCARU R., CAPPELLI A., ALOBEIDLI H., PANNIER B., ALMAZROUEI E. & LAUNAY J. (2023). The RefinedWeb dataset for Falcon LLM : outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv :2306.01116*.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). SQuAD : 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, p. arXiv :1606.05250.

RENNARD V., SHANG G., GRARI D., HUNTER J. & VAZIRGIANNIS M. (2023). FREDSum : A dialogue summarization corpus for french political debates. In *Findings of Empirical Methods in Natural Language Processsing (EMNLP)*.

RHAPSODIE (2015). https ://rhapsodie.modyco.fr/propriete-intellectuelle/.

SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIĆ S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F., GALLÉ M., TOW J., RUSH A. M., BIDERMAN S., WEBSON A., AMMANAMANCHI P. S., WANG T., SAGOT B., MUENNIGHOFF N., DEL MORAL A. V., RUWASE O., BAWDEN R., BEKMAN S., MCMILLAN-MAJOR A., BELTAGY I., NGUYEN H., SAULNIER L., TAN S., SUAREZ P. O., SANH V., LAURENÇON H., JERNITE Y., LAUNAY J., MITCHELL M., RAFFEL C., ... 357 AUTHORS ... & WOLF T. (2023). Bloom : A 176b-parameter open-access multilingual language model.

SUÁREZ P. J. O., SAGOT B. & ROMARY L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)* : Leibniz-Institut für Deutsche Sprache.

THÉÂTRE CLASSIQUE (2022). http ://www.theatre-classique.fr/.

THÉÂTRE GRATUIT (2023). https ://theatregratuit.com/.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S. *et al.* (2023). Llama 2 : Open foundation and fine-tuned chat models. *arXiv preprint arXiv :2307.09288*.

TUTIN A. & GROSSMAN F. (2014). L'écrit scientifique : du lexique au discours. *Autour de Scientext*, p. 27–44.

UPADHAYAY B. & BEHZADAN V. (2023). Taco : Enhancing cross-lingual transfer for low-resource languages in llms through translation-assisted chain-of-thought processes. *arXiv preprint arXiv :2311.10797*.

XU Y., LEE H., CHEN D., CHOI H., HECHTMAN B. & WANG S. (2020). Automatic cross-replica sharding of weight update in data-parallel training.

YAMASAKI H., LOURADOUR J., HUNTER J. & PRÉVOT L. (2023). Transcribing and aligning conversational speech : A hybrid pipeline applied to french conversations. In *Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

ZELLERS R., HOLTZMAN A., BISK Y., FARHADI A. & CHOI Y. (2019). Hellaswag : Can a machine really finish your sentence ? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

ZHANG J., QIAN K., LIU Z., HEINECKE S., MENG R., LIU Y., YU Z., SAVARESE S. & XIONG C. (2023). DialogStudio : Towards richest and most diverse unified dataset collection for conversational AI. *arXiv preprint arXiv :2307.10172*.

ZHAO Y., GU A., VARMA R., LUO L., HUANG C.-C., XU M., WRIGHT L., SHOJANAZERI H., OTT M., SHLEIFER S., DESMAISON A., BALIOGLU C., DAMANIA P., NGUYEN B., CHAUHAN G., HAO Y., MATHEWS A. & LI S. (2023). Pytorch fsdp : Experiences on scaling fully sharded data parallel.

ZHENG L., CHIANG W.-L., SHENG Y., ZHUANG S., WU Z., ZHUANG Y., LIN Z., LI Z., LI D., XING E., ZHANG H., GONZALEZ J. E. & STOICA I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

ZHU C., LIU Y., MEI J. & ZENG M. (2021). Mediasum : A large-scale media interview dataset for dialogue summarization. *CoRR*, **abs/2103.06410**.

ÉQUIPE DELIC (2004). Autour du corpus de référence du français parlé. *Recherches sur le français parlé*, **18**, 265.

A Survey questions from the second evaluation campaign

Here is a list of the survey questions corresponding to the second round of evaluation. Most of the questions have a yes/no format, with a third option proposed if the question does not apply or to quantify some characteristic (e.g. for 2.d, the options are "Plusieurs erreurs"/"Quelques erreurs"/"Aucune erreur").

1. Interaction
   (a) En présence de “tours” délimités par des tirets (“-”) ou des étiquettes telles que “[Nom 1 :]” ou “Prénom :”, ces marqueurs sont-ils positionnés de manière intuitive et, pour les étiquettes, dans un ordre plausible ? (Ignorez les erreurs mineures de format, e.g., [Intervenant 2) :])
   (b) En excluant les étiquettes de locuteurs, la réponse semble-t-elle davantage être extraite d’une conversation ou d’un document écrit ? (La cohérence de la réponse n’est pas primordiale ici ; l’accent est mis sur le style généré par le modèle.)
   (c) En cas de dialogue avec plusieurs échanges, semble-t-il que les interlocuteurs cherchent à engager une discussion (en se tutoyant/vousvoyant directement, répondant aux questions, utilisant des expressions conversationnelles comme ’oui, c’est vrai’, etc.) ?
2. Fluidité
   (a) La réponse bascule-t-elle vers l’anglais ou un autre type de langage ?
   (b) La réponse présente-t-elle des répétitions non motivées ? (Par exemple, le modèle semble-t-il répéter de manière robotique les mêmes expressions, sans justification ? Les répétitions naturelles de la conversation sont considérées comme justifiées.)
   (c) Le niveau de disfluences (telles que “euh”, “hm”, “quoi”) vous semble-t-il naturel par rapport à une conversation orale ? (s’il y en a pas, répondez “oui”)
   (d) En dehors des disfluences et des répétitions robotiques, la partie restante de la réponse est-elle formulée dans un français parlé correct ? (Veuillez noter que les constructions typiques du français parlé, telles que “y a pas” ou “ton père, il est où ?”, sont acceptables.)
   (e) Dans l’ensemble, avez-vous l’impression que la conversation semble humaine (par opposition à “générée par l’IA”) ?
3. Pertinence
   (a) La suite de la conversation, répond-elle de manière spécifique au début de la conversation proposé, même si elle est incorrecte sur le plan factuel ? (Cela s’oppose à une suite qui pourrait être tout aussi valable pour un début de conversation bien différent.)
   (b) Y a-t-il des changements de sujet ou de ton qui attirent particulièrement l’attention ? (Ignorez les changements légers de sujet courants dans la conversation spontanée ; un passage à un sujet très différent avant d’avoir traité le premier pourrait être considéré comme frappant.)
   (c) Le niveau de disfluences (telles que “euh”, “hm”, “quoi”) vous semble-t-il naturel par rapport à une conversation orale ? La conversation semble-t-elle stagnante sur un sujet (plutôt que de progresser comme elle le devrait) ?
   (d) En dehors des disfluences et des répétitions robotiques, la partie restante de la réponse est-elle formulée dans un français parlé correct ? La conversation contient-elle des stagnante ?
   (e) Le modèle semble-t-il inventer des choses de manière exagérée ? (On exclut les détails à vérifier dans une encyclopédie ; l’accent est mis sur les détails qui semblent inventés et perturbent la cohérence de la conversation.)

B Evaluation results

As explained in Section 5 and illustrated in Figure 2, Mistral (red) had a clear tendency to switch to English, code or another data format when prompted in French. This tendency disappeared for the Claire-Mistral model (orange). The results in Figure 2 are from our second evaluation campaign and indicate a score out of 80, as 80 was the number of surveys conducted.

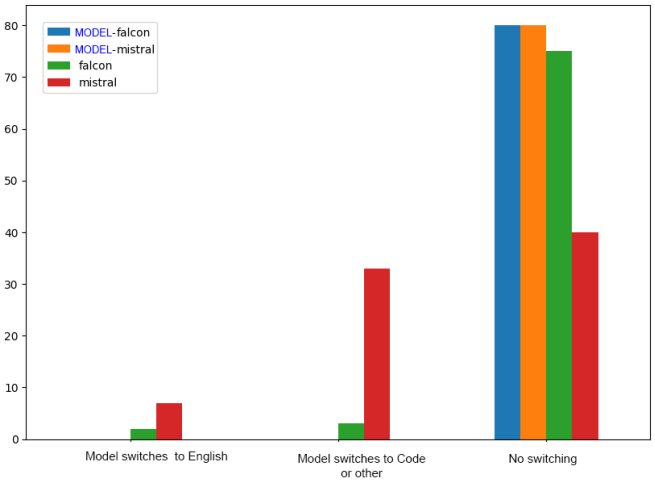


FIGURE 2 – The tendency of each model to stick to French when prompted in French

Results from both campaigns showed a clear overall preference for Claire-Falcon over Claire-Mistral, with Mistral coming it at fourth place. See Figure 3.

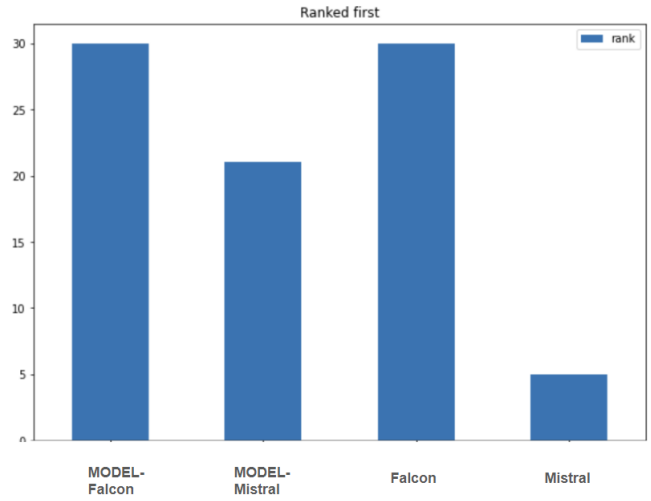


FIGURE 3 – Overall rankings of the four models in the second evaluation campaign

Figure 4 shows judgments regarding whether the model seemed human. These results were the most strongly correlated with overall preference. In this evaluation, Claire-Falcon (blue) was very slightly preferred over Falcon (green).

Finally, if we focus on "culture"-style prompts, as shown in Figure 5, we see that Claire-Falcon is clearly preferred over Falcon, a tendency that was observed in the first round of evaluations as well. Falcon takes the lead for "casual" prompts, however.

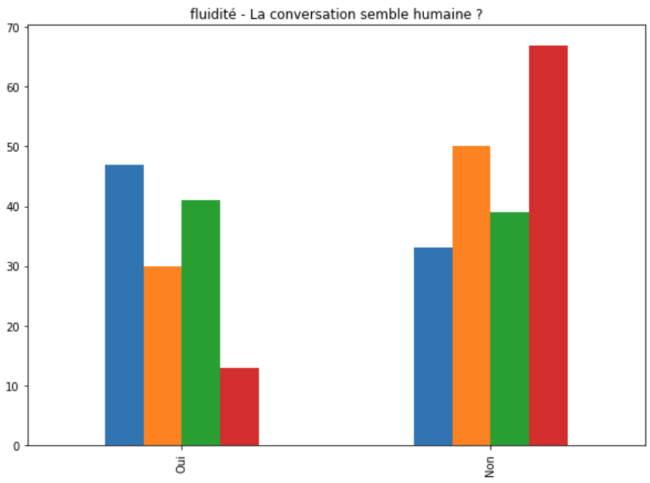


FIGURE 4 – Judgments concerning the number of times the response of a model was judged as human

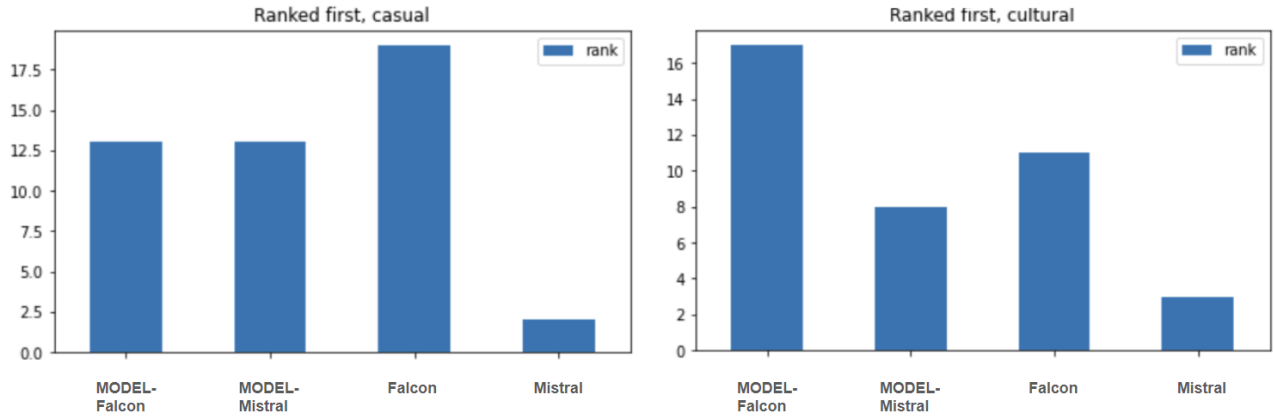


FIGURE 5 – Results showing that Claire-Falcon is clearly preferred for "culture" prompts. (A result reflected also in the first evaluation, which contained only culture prompts.)

C Sources for the original datasets used to train the principal Claire models

ACSYNT : Cognition, Langue, Langages, Ergonomie (CLLE) (2013)
Assemblée Nationale : Assemblée nationale (2023)
Orféo-CEFC : Debaisieux *et al.* (2016); Carruthers (2008); Tutin & Grossman (2014)
— **C-ORAL-ROM** : Cresti *et al.* (2004)
— **CRFP** : Équipe Delic (2004)
— **FLEURON** : no citation found
— **Valibel** : Dister *et al.* (2007)
Orféo : (other)
— **CFPB** : Dister & Labeau (2017)
— **Réunions De Travail** : (Work Meetings)
CFPP : Branca-Rosoff *et al.* (2013); CLESTHIA (2018)
CID : Bertrand *et al.* (2008); Blache *et al.* (2010); LPL (2021)

CLAPI : Baldauf-Quilliatre *et al.* (2016)
ESLO : Eshkol-taravella *et al.* (2011)
FREDSum : Rennard *et al.* (2023)
LinTO : Gravellier *et al.* (2021)
OFROM : Avanzi *et al.* (2012 2023)
Parole Publique :
— **Accueil UBS** : Nicolas *et al.* (2002); Antoine *et al.* (2009)
— **OTG** : Nicolas *et al.* (2002); Antoine *et al.* (2002)
Paris Stories : Kahane *et al.* (2021)
PFC : Durand *et al.* (2009); MoDyCo & RUG (2017)
Rhapsodie : Rhapsodie (2015); Branca-Rosoff *et al.* (2013); Avanzi (2012); Lacheret (2003); Mertens (1987); Avanzi *et al.* (2010); Eshkol-taravella *et al.* (2011); Durand *et al.* (2009)
SUMM-RE : Yamasaki *et al.* (2023)
TCOF : ATILF (2020)
Théâtre Classique : Théâtre classique (2022); Milling *et al.* (2021); Fischer *et al.* (2019)
Théâtre Gratuit : Théâtre Gratuit (2023)

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Modéliser la facilité d'écoute en FLE : vaut-il mieux lire la transcription ou écouter le signal vocal ?

Minami Ozawa[1, 2] Rodrigo Wilkens[1] Kaori Sugiyama[2] Thomas François[1]
(1) Université catholique de Louvain, IL&C, CENTAL, Louvain-la-Neuve, Belgique
(2) Université Seinan Gakuin, Fukuoka, Japon
minami.ozawa@uclouvain.be, rodrigo.wilkens@uclouvain.be,
sugiyama@seinan-gakuin.jp, thomas.francois@uclouvain.be

RÉSUMÉ

Le principal objectif de cette étude est de proposer un modèle capable de prédire automatiquement le niveau de facilité d'écoute de documents audios en français. Les données d'entrainement sont constituées d'enregistrements audios accompagnés de leurs transcriptions et sont issues de manuels de FLE dont le niveau est évalué sur l'échelle du Cadre européen commun de référence (CECR). Nous comparons trois approches différentes : machines à vecteurs de support (SVM) combinant des variables de lisibilité et de fluidité, wav2vec et CamemBERT. Pour identifier le meilleur modèle, nous évaluons l'impact des caractéristiques linguistiques et prosodiques ainsi que du style de parole (dialogue ou monologue) sur les performances. Nos expériences montrent que les variables de fluidité améliorent la précision du modèle et que cette précision est différente par style de parole. Enfin, les performances de tous les modèles varient selon les niveaux du CECR.

ABSTRACT

Modelling listenability for FFL : is it better to read the transcript or listen to the speech signal ?

Our main goal is to design a model able to automatically predict the level of listenability of audio documents in French as a foreign language (FFL). The training data consists of audio recordings accompanied by their transcriptions and are extracted from FFL textbooks whose level is assigned accordingly to the Common European Framework of Reference (CEFR) scale. We compare three different approaches : support vector machines combining readability and fluency variables, wav2vec, and CamemBERT. To identify the best model, we evaluate the impact of linguistic and fluency features as well as of speech style (dialogue or monologue) on performance. Our experiments show that fluency variables improve model accuracy, and that this accuracy differs by speech style. Finally, the performance of all models varies according to CEFR scales.

MOTS-CLÉS : facilité d'écoute, lisibilité, FLE, wav2vec.

KEYWORDS: listenability, readability, FFL, wav2vec.

1 Introduction

La compréhension orale est une activité langagière dynamique activement impliquée dans l'interprétation. Le développement de cette compétence est directement lié à l'acquisition de la compétence langagière en général : si la compréhension orale ne progresse pas, cela peut nuire à la bonne communication dans des situations authentiques (Kumai, 1992). Pour soutenir son développement

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 549–566.

chez les apprenants de langue étrangère, il est courant de faire écouter divers documents audios dans le cadre d'exercices de compréhension à l'audition. L'efficacité de documents audio-visuels authentiques est bien établie (Yoshimi, 2019), mais leur utilisation en contexte de classe reste difficile, car il est ardu de juger de l'adéquation de ce type de documents audios au niveau de compétence linguistique des apprenants. Une solution pour faire face à cette difficulté pratique pourrait consister à se reposer sur des modèles d'intelligence artificielle capables d'évaluer automatiquement le niveau de difficulté de documents oraux, sur le modèle de ce qui est fait en évaluation automatisé de la lisibilité des documents écrits. Si les travaux en lisibilité se sont imposés comme une des thématiques du TAL (François, 2011; Vajjala, 2021), ceux sur la facilité d'écoute (ou *listenability*, en anglais) sont nettement moins nombreux, en particulier en français.

Les objectifs de cette étude sont doubles. Tout d'abord, le principal objectif est de développer un modèle capable de prédire automatiquement le niveau de facilité d'écoute de documents audios en français. À notre connaissance, un tel modèle n'existant actuellement pas en français, il s'agit déjà d'une contribution significative. Dans cette optique, deux questions émergent. D'une part, quelle est la meilleure façon d'encoder les caractéristiques stylistiques des documents oraux ? Pour y répondre, nous comparerons l'apport de variables linguistiques définies par des experts et combinées à l'aide de machines à vecteur de support (SVM) à l'utilisation de l'architecture transformer et CamemBERT (Martin *et al.*, 2020) qui représente le contenu des documents audios à l'aide de plongements de mots. D'autre part, quel est l'apport de variables prosodiques ? En effet, la grande majorité des travaux en facilité d'écoute se limitent à prédire la difficulté sur la base de la transcription et de variables calculées sur le texte, sans prendre en compte les caractéristiques spécifiques à l'oralité. Nous comparerons donc ces deux approches au sein de nos modèles, encodant les caractéristiques prosodiques soit à l'aide de variables incluses dans les SVM, soit en tirant profit de l'architecture wav2vec (Baevski *et al.*, 2020). Le second objectif de cet article vise à évaluer l'impact du style de parole sur les performances des modèles précités. Nous disposons d'une variété de documents oraux que nous regroupons simplement selon deux styles de parole : le dialogue et le monologue. Nous examinons si les performances des modèles diffèrent en fonction de la présence ou de l'absence d'un interlocuteur dans le discours.

Cet article propose d'abord une vue d'ensemble de la notion centrale de la facilité d'écoute dans la section 2, suivie de la description de la méthodologie de recherche dans la section 3. Il présente et compare ensuite les résultats des modèles SVM, de wav2vec et de CamemBERT dans la section 4, avant de conclure.

2 Le domaine de la facilité d'écoute

Plusieurs définitions de la facilité d'écoute ont été proposées jusqu'à présent (Harwood, 1950; Harwood & Cartier, 1952; Cartier, 1952; Rubin, 2012). En tenant compte de ces définitions, dans cet article, le terme *facilité d'écoute* est défini comme la facilité ou la difficulté qu'un auditeur donné – avec son expérience spécifique – éprouve à comprendre un discours oral dans une situation de communication particulière, laquelle est fonction de l'effet des caractéristiques stylistiques de ce discours (ex. lexicale, syntaxique, prosodique, etc.) sur les processus cognitifs de l'auditeur.

En tant que domaine de recherche, la facilité d'écoute s'est principalement développée en langue anglaise. Dans les années 1940, ce champ a été initié par l'application des modèles issus des études de lisibilité à l'analyse des transcriptions de la langue orale (Flesch, 1943; Chall & Dial, 1948). Par

la suite, les travaux continuent à évaluer la facilité d'écoute sur la base de transcriptions et avec des formules de lisibilité développées pour l'écrit (Cartier, 1955; O'Keefe, 1971).

Toutefois, peu à peu, des formules spécifiques à l'oral apparaissent. Des recherches (Rogers, 1962; Fang, 1967; Kiyokawa, 1990) conçoivent des modèles directement sur des données orales, mais ils ne prennent toujours en compte que des variables liées au contenu (longueur des phrases, des mots, proportion de mots polylexicaux, etc.) et les observent sur des transcriptions. Au niveau de la facilité d'écoute dans le contexte spécifique des apprenants, il y a une recherche qui utilisent la régression multiple pour combiner une série de variables de contenu afin de prédire la difficulté subjective de compréhension orale des phrases anglaises, évaluée par 90 apprenants japonais de l'anglais (Ueda *et al.*, 2013). Le coefficient de corrélation multiple R de l'équation atteint seulement 0,54.

C'est avec les travaux de Kotani *et al.* (2014) que, en complément des variables de contenu, des caractéristiques phonologiques sont enfin considérées, à savoir le débit de parole, le taux d'élision [1], de réduction [2], de contraction [3], de liaison [4] et de déduction [5] dans une phrase. La présence de coefficients de régression négatifs pour le taux de contraction et de déduction suggère que des variations phonologiques peuvent accroître la facilité d'écoute pour les apprenants. Parallèlement, il y a une recherche qui examine dans quelle mesure le débit de parole, la complexité linguistique et l'explicitation (le degré d'expression explicite des idées) du texte influent sur la compréhension orale en L2 (Révész & Brunfaut, 2013). Les auteurs utilisent des analyses de Rasch et de régression pour estimer la difficulté de 18 tâches et sa relation avec les caractéristiques du texte. Les résultats démontrent que des indices de complexité lexicale et discursive expliquent une part significative de la difficulté des tâches de compréhension orale.

Enfin, plus récemment, les arbres de décision (Kotani & Yoshimi, 2017) ou les machines à vecteurs de support (SVM) (Yoshimi & Kotani, 2020) ont été utilisés. Néanmoins, toutes les recherches susmentionnées ont été menées en anglais. En ce qui concerne la langue française, les investigations sont très restreintes aux travaux de Ruggia (2019, 2020, 2021). De plus, un outil est aussi disponible pour l'évaluation automatique des textes oraux, DeepFLE [6], mais il reste encore limité à l'analyse des transcriptions et n'intègre pas de variables prosodiques. Enfin, l'utilisation de représentations latentes (ex. plongements de mots) et de wav2vec plutôt que des variables ne semble pas avoir encore été envisagé dans ce domaine.

3 Méthodologie

Dans cet article, nous cherchons à évaluer l'apport de variables linguistiques et prosodiques à un modèle de facilité d'écoute pour le français, mais posons également la question de savoir si les représentations latentes du contenu ne sont pas préférables à des variables définies par des experts, comme cela a déjà été établi en lisibilité (Martinc *et al.*, 2021; Yancey *et al.*, 2021). Cette section commence par décrire les données utilisées pour entraîner nos modèles, puis décrit les différentes variables « expert » considérées, avant de se clôturer avec une présentation des différents modèles évalués.

1. L'élision est l'élimination des phonèmes.
2. La réduction est l'affaiblissement du son en transformant une voyelle en schwa.
3. La contraction est la combinaison de deux mots.
4. La liaison consiste à relier le son final d'un mot au son initial du mot suivant.
5. La déduction est l'élimination des sons entre les mots.
6. http ://deeptext.unice.fr/FLE/.

3.1 Données

Les données sont constituées d'enregistrements audios accompagnés de leurs transcriptions que nous avons extraits de 25 manuels de FLE. Chaque enregistrement s'est vu attribué un niveau de compétence sur l'échelle du Cadre européen commun de référence (CECR), à savoir A1, A2, B1, B2 ou C (C1 et C2 sont regroupés pour notre étude). Le niveau attribué à un enregistrement donné est tout simplement celui du manuel dont il a été tiré.

Plusieurs prétraitements ont été effectués. Tout d'abord, le contenu d'une seule piste audio est en principe traité comme une seule donnée. Toutefois, si une piste contient, par exemple, plusieurs dialogues dans différents contextes, l'audio (et la transcription) est alors divisé manuellement. Ensuite, nous avons éliminé les éléments de soutien pédagogique (tels que les exercices de grammaire et les listes de mots), ainsi que les données provenant des premières et dernières unités de chaque manuel (qui sont trop proches du niveau précédent ou suivant). Dans notre corpus, les enregistrements comprennent des conversations quotidiennes, des annonces au public, des interviews médias et des monologues. Toutes ces données ont été classées manuellement en deux grands styles de parole : les dialogues et les monologues. Enfin, les données dont le contenu informatif est extrêmement élevé ou faible dans une certaine catégorie de niveau/style de parole peuvent réduire l'homogénéité lors de l'examen des caractéristiques de ce niveau/style de parole. Pour cette raison, les cinq données présentant des valeurs aberrantes extrêmes en termes de syllabes par minute sans les pauses ou/et de mots par minute sans les pauses ont été exclues. Ces cinq données se composaient d'un document B2, d'un document C1 et de trois documents C2. Les documents B1, C1 et un des C2 ne comportaient qu'une ou deux phrases. La petite taille des données par rapport au niveau/style pourrait être considérée comme une indication d'hétérogénéité. Les deux autres documents C2 étaient des données avec beaucoup de bruit et des phrases en langue étrangère non mentionnés dans les transcriptions. Ils ont été traitées comme des pauses pour des raisons de commodité et, par conséquent, les valeurs aberrantes ont montré que ces données étaient hétérogènes et ne convenaient pas comme données pour l'analyse. Le nombre total final de paires de données (audio et transcription) est de 1 323 documents (Table 1). Le nombre de mots ainsi que la durée d'enregistrement pour les niveaux C1 et C2 sont largement supérieurs à ceux des autres niveaux. Cependant, les variables que l'on utilise sont normalisées lorsque c'est nécessaire afin d'éviter les biais.

3.2 Variables

Dans cette étude, la difficulté de compréhension (comprenant nos cinq niveaux issus du CECR comme modalités) a été utilisée comme variable dépendante tandis que 68 variables capturant les caractéristiques prosodiques et le contenu textuel sont utilisées comme variables indépendantes. Les variables indépendantes sont divisées en deux catégories principales : les variables de fluidité (7 variables) et les variables de lisibilité (61 variables). Dans cet article, les variables liées à la prosodie se limitent à celles relatives au débit de parole, à la durée de parole et aux pauses.

La seule variable prosodique qui a été identifiée dans la littérature comme étant corrélée à la facilité d'écoute est le nombre de mots par minute (Harwood, 1955; Kotani *et al.*, 2014; Kotani & Yoshimi, 2017). Cependant, étant donné le nombre réduit de recherches sur cette question, nous avons décidé d'enrichir la liste des variables prises en compte en nous référant au concept de fluidité, qui est souvent utilisé dans les recherches sur la production orale. La fluidité se réfère à la capacité d'utiliser la langue en temps réel, de mettre l'accent sur le sens et d'utiliser les systèmes lexicaux (Skehan

	Nb de données (%)		Nb de mots (%)		Durée d'enregistrement (minutes) (%)	
	Dialogue	Monologue	Dialogue	Monologue	Dialogue	Monologue
A1	290 (100 %)		18 950 (100 %)		160 (100 %)	
	205 (71 %)	85 (29 %)	14 576 (77 %)	4 374 (23 %)	120 (75 %)	40 (25 %)
A2	309 (100 %)		33 755 (100 %)		230 (100 %)	
	180 (58 %)	129 (42 %)	25 665 (76 %)	8 090 (24 %)	180 (78 %)	50 (22 %)
B1	240 (100 %)		32 818 (100 %)		190 (100 %)	
	146 (61 %)	94 (39 %)	25 289 (77 %)	7 529 (23 %)	150 (79 %)	40 (21 %)
B2	241 (100 %)		51 920 (100 %)		290 (100 %)	
	131 (54 %)	110 (46 %)	30 825 (59 %)	21 095 (41 %)	170 (59 %)	120 (41 %)
C1/C2	243 (100 %)		127 396 (100 %)		710 (100 %)	
	121 (50 %)	122 (50 %)	95 934 (75 %)	31 462 (25 %)	520 (73 %)	190 (27 %)
Total	1 323 (100 %)		264 839 (100 %)		1 580 (100 %)	
	783 (59 %)	540 (41 %)	192 289 (73 %)	72 550 (27 %)	1 140 (72 %)	440 (28 %)

TABLE 1 – Description du jeu de données, qui précise le nombre de documents, de mots par document et la durée en minute, de façon globale et par style de parole.

& Foster, 1999). Étant donné que la compréhension orale, tout comme la production orale, est une aptitude qui requiert de telles compétences, il est raisonnable d'appliquer ce concept. Nous nous référons à la méthode de calcul utilisée dans les recherches précédentes (Cucchiarini *et al.*, 2002; Ginther *et al.*, 2010; Préfontaine, 2010; Peltonen, 2017; Segalowitz *et al.*, 2017) pour définir les variables utilisées dans cet article, à savoir : (1) le nombre de syllabes par minute, (2) le nombre de syllabes par minute sans les pauses, (3) le nombre de mots par minute, (4) le nombre de mots par minute sans les pauses, (5) la durée totale, (6) la durée moyenne des pauses, et (7) le nombre de pauses par minute.

Pour les variables de lisibilité, 61 variables de FABRA (Wilkens *et al.*, 2022) sont utilisées. FABRA est un outil en ligne permettant de calculer un large éventail de variables prédictives de lisibilité pour le français. Le domaine de la facilité d'écoute s'est développé par l'application des formules de lisibilité à l'analyse des transcriptions, et l'existence signalée de variables valides pour la facilité d'écoute ne peut être ignorée. Cet article se focalise sur les variables de lisibilité qui ont été considérées comme pertinentes dans les études précédentes sur la facilité d'écoute. Ainsi, bien que FABRA puisse fournir des informations sur 509 variables liées à la lisibilité, certaines d'entre elles sont moins pertinentes dans le contexte de la facilité d'écoute et ne font pas partie de notre analyse. Les variables liées au nombre de ponctuations et de guillemets dans FABRA, par exemple, ne sont pas directement pertinentes pour cette analyse. Bien que certaines variables relatives aux erreurs dans FABRA existent, il n'est pas non plus nécessaire de les considérer, étant donné que l'étude ne se concentre pas sur la production des apprenants. En outre, cet article se concentre sur les variables liées aux mots de contenu, nécessaires pour comprendre le sens du discours, plutôt qu'aux mots fonctionnels. De plus, certaines variables de FABRA sont relativement redondantes, mesurant le même phénomène à l'aide de variantes. Dans ce cas, seule l'une d'entre elles est prise en considération (par exemple, la proportion de mots A1 et A2 selon la ressource FLELex (François *et al.*, 2014) [7] est fort redondante. Dès lors, seule la proportion de mots A1 dans le texte selon la ressource a été prise en considération).

7. FLELex est un lexique pour le FLE qui donne les fréquences normalisées des lemmes à chaque niveau du CECRL (http://cental.uclouvain.be/flelex/).

Au terme de cette première sélection de variables, 61 variables de lisibilité ont été retenues parmi les variables de FABRA. Cette procédure de sélection est conforme à la première étape de sélection des variables, qui consiste à construire un meilleur ensemble de caractéristiques « ad hoc » à partir de la connaissance du domaine de la facilité d'écoute, selon Guyon & Elisseeff (2003). En résumé, les variables comprennent (1) 2 variables basées sur la longueur (ex. nombre de syllabes par mot), (2) 17 variables lexicales (ex. fréquence moyenne des lemmes dans FLELex pour les noms, les noms propres, les verbes, les adjectifs ou les adverbes, (3) 41 variables syntaxiques basées sur le formalisme UD et l'analyseur Stanza (Qi *et al.*, 2020) (ex. proportion d'indicatifs présent, mots identifiés comme adjectifs, etc.) et (4) le score de la formule de lisibilité de Kandel et Moles (Kandel & Moles, 1958) [8] [9].

3.3 Modèles

Nous comparons trois approches différentes pour entraîner notre modèle de facilité d'écoute. La première s'ancre dans les travaux de lisibilité computationnelle, combinant les 68 variables décrites à la section 3.2 à l'aide de SVM. Elle vise à reproduire une méthodologie fiable, qui a fait ses preuves en lisibilité et est relativement facile à entraîner. La seconde approche consiste simplement à affiner une architecture CamemBERT (Martin *et al.*, 2020).

Plus original, l'emploi du modèle wav2vec 2.0 (Baevski *et al.*, 2020) en facilité d'écoute vise à capturer au mieux la richesse du signal vocal. Le modèle traite la forme d'onde brute du signal vocal avec un réseau neuronal convolutionnel multicouches pour obtenir des représentations audio latentes. Ces représentations sont ensuite introduites dans un quantificateur et un transformateur. Le quantificateur choisit une unité vocale pour la représentation audio latente à partir d'un inventaire d'unités apprises. Environ la moitié des représentations audio sont masquées avant d'être introduites dans le transformateur. Le transformateur ajoute des informations provenant de l'ensemble de la séquence audio. Enfin, la sortie du transformateur est utilisée pour résoudre une tâche contrastive. Cette tâche exige que le modèle identifie les unités de parole quantifiées correctes pour les positions masquées. Pour la version française, nous utilisons le modèle *facebook/wav2vec2-large-xlsr-53-french* [10], qui est un modèle wav2vec 2.0 entraîné sur des données audio non annotées de 12 langues provenant du benchmark Common Voice. [11]

Nous avons également ajouté une tête de classification à wav2vec. Cette tête est composée d'un pooling de moyennes 1D sur le dernier état caché de wav2vec. Ensuite, les informations regroupées passent par un MLP composé de deux couches linéaires denses (la première avec une activation tanh et la seconde avec une activation softmax) chacune précédée d'une couche de Dropout avec des probabilités de 0,5 et 0,1.

Afin de tester nos deux hypothèses principales : l'effet des variables prosodiques et du style de parole, plusieurs modèles ont été entraînés sur la base des 3 architectures ci-dessus. En ce qui concerne l'effet des variables prosodiques, trois jeux de variables différents ont été utilisés pour les SVMs : un jeu ne comprenant que des variables de fluidité (SVM-F), un second jeu ne comprenant que des variables de lisibilité (SVM-L) et un dernier jeu incluant à la fois les variables de fluidité et celles de

8. Voir https://cental.uclouvain.be/fabra/docs.html pour le détail des variables.
9. Les détails de toutes les variables utilisées dans cette étude sont donnés en annexe.
10. https://huggingface.co/facebook/wav2vec2-large-xlsr-53-french
11. Pour plus d'informations, voir https://ai.meta.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/.

lisibilité (SVM-FL). Au niveau des représentations latentes, nous avons CamemBERT et wav2vec. En comparant CamemBERT avec SVM-L et wav2vec avec SVM-F, il est possible d'examiner s'il y a une contribution des modèles neuronaux et des représentations distribuées. En outre, en comparant SVM-FL, SVM-L et SVM-F, il est possible d'examiner si les variables de fluidité ont un impact sur la construction d'un modèle de facilité d'écoute. Par ailleurs, ces 5 modèles ont été déclinés sur trois ensembles de données différents : l'ensemble des données, les données correspondant au style de parole monologal et celles associées aux dialogues.

Enfin, en ce qui concerne la recherche des hyperparamètres, nous les avons explorés à l'aide d'une recherche par grille. Les performances de tous nos modèles ont été évaluées au moyen d'une procédure de validation croisée à 10 blocs et à l'aide des mesures d'évaluation de précision, de rappel et de F1. Dans la validation croisée à 10 blocs, 80 % des données servent à la phase d'entraînement, 10 % pour le développement et 10 % pour le test. Les moyennes des 10 plis obtenus pour l'exactitude et la F1 score sont rapportées.

4 Analyse des variables et des modèles

Dans cette section, nous présentons d'une part une analyse corrélationnelle des 68 variables à la section 4.1, puis les résultats des différents modèles (SVM-F, SVM-L, SVM-FL, CamemBERT et wav2vec) à la section 4.2.

4.1 Analyse corrélationnelle des variables

Tout d'abord, pour l'ensemble des données et chaque style de parole (dialogue, monologue), le coefficient de corrélation de Spearman a été utilisé pour examiner la corrélation entre chacune des 68 variables et le niveau. Seules les variables dont la valeur absolue du coefficient de corrélation est supérieure à 0,2 ($p < 0{,}05$) sont conservées. Lorsque plusieurs variables sont incluses dans une même catégorie, pour chaque catégorie, seules celles qui ont la plus forte corrélation avec le niveau sont conservées tandis que les autres sont exclues, ce qui permet de minimiser la colinéarité entre les variables lors de la construction du modèle. En suivant la procédure ci-dessus pour chaque ensemble de données (données entières, dialogue, monologue), les variables finales à incorporer dans nos modèles ont été déterminées.

En comparant les trois ensembles de variables, les différences dans les coefficients de corrélation sont particulièrement notables, surtout pour les variables liées à la fluidité (nombre de syllabes par minute, durée moyenne des pauses et nombre de pauses par minute). Toutes les variables sélectionnées sur l'ensemble des données se retrouvent dans les jeux de variables spécifiques au dialogue et au monologue, l'inverse n'étant pas vrai.

En confrontant les résultats obtenus sur les dialogues et sur les monologues, on constate que 17 variables sont communes aux deux ensembles. Cela représente 56 % des variables finales dans le dialogue et 80 % des variables finales dans le monologue, ce qui suggère que le niveau tend à être déterminé par davantage de variables dans le dialogue que dans le monologue. En particulier, dans le dialogue, les variables liées aux temps des verbes sont plus fréquemment mentionnées comme variables fortement corrélées avec le niveau que dans le monologue. Ainsi, malgré le fait que les variables finalement traitées ont été considérées dans les mêmes conditions dans les deux styles

de parole, il y a des différences dans le nombre de variables et les variables ne sont pas tout à fait identiques. En d'autres termes, cela montre déjà l'utilité de construire le modèle séparément pour chaque style de parole.

4.2 Comparaison des modèles

Il est intéressant d'analyser les performances des 5 modèles sur l'ensemble des données (dialogue et monologue) dont les résultats sont repris à la Table 2a, mais aussi leurs performances sur les deux styles de parole de façon isolée (Tables 2b et 2c).

En comparant SVM-L et CamemBERT pour examiner l'apport des variables linguistiques, CamemBERT apparait à première vue comme le meilleur modèle, aussi bien sur l'ensemble des données que sur les monologues. Pour chaque modèle, la F1 de CamemBERT se situe entre 0,5 et 0,7. Un tel résultat n'est pas surprenant, car les corpus issus de manuels de FLE se caractérisent généralement par une forte hétérogénéité au niveau des annotations de la difficulté entre manuels (François, 2014). Cependant, un problème important de CamemBERT est que son écart-type est particulièrement élevé, ce qui implique une instabilité dans le modèle et nous conduit à regarder avec prudence ces résultats.

Ensuite, SVM-F et wav2vec sont comparés afin d'examiner l'apport des variables prosodiques. Les résultats de wav2vec montrent que, pour chaque style de parole, la F1 est d'environ 0,2, ce qui est inférieur à celles de SVM-F. Cela semble indiquer que l'information audio identifiée par wav2vec ne constitue pas un signal assez fort pour identifier les niveaux de difficulté. Les performances de SVM-F, assez mauvaises également, confirment que les variables de fluidité envisagées dans cette étude ont une contribution limitée à la prédiction de la facilité d'écoute. Néanmoins, ces deux résultats de F1 montrent de manière générale des performances assez élevé sur le niveau A1 (0,44 à 0,53 pour wav2vec, 0,55 à 0,65 pour SVM-F) et, dans une moindre mesure, sur le niveau A2 (0,31 à 0,38 pour wav2vec, 0,28 à 0,31 pour SVM-F), pour tous les styles de parole. Or, ces niveaux sont notamment caractérisés par un débit plus lent : le nombre de syllabes par minute en A1 (181 syll./min.) est bien moindre que au niveau A2 (226 syll./min.)) et aux niveaux B1 (246 syll./min.) et B2 (255 syll./min.). Cela peut indiquer une tendance du modèle à utiliser des informations relatives au débit de parole pour évaluer la difficulté des discours. Toutefois, sur les niveaux supérieurs, le contenu devient plus critique et il semblerait que wav2vec ne capture pas suffisamment la teneur de celui-ci.

En ce qui concerne SVM-FL, il apparaît que, pour chaque style de parole et de façon générale, la F1 obtenue (0,49 à 0,54) est supérieure à celle de SVM-L. Cela signifie que les variables de fluidité apportent tout de même des informations utiles au modèle et différentes de celles des variables de lisibilité. Cependant, ces F1 sont inférieures à la F1 de CamemBERT, pour l'ensemble de données et pour les monologues. En d'autres termes, dans ces styles de parole, CamemBERT, qui est un modèle neuronal sans informations prosodiques, est plus performant.

Enfin, en examinant les résultats spécifiquement sur les dialogues et les monologues, ces deux styles de parole ne produisent pas systématiquement les mêmes résultats pour chaque modèle. Néanmoins, on peut observer une certaine régularité dans les résultats : la F1 du modèle de tous les SVMs et de wav2vec montre que ceux-ci sont presque toujours meilleurs sur les dialogues. Au contraire, CamemBERT se comporte toujours mieux sur les monologues. Cependant, comme mentionné ci-dessus, en raison des problèmes de variance dans les résultats de CamemBERT, les performances de ce modèle doivent être considérées avec précaution. Par conséquent, en se basant sur la tendance claire des résultats des SVMs et de wav2vec, on peut considérer que prédire la facilité d'écoute sur

	CamemBERT	wav2vec	SVM-FL	SVM-L	SVM-F
Exactitude (écart-type)	0,64 (0,36)	0,33 (0,03)	0,51 (0,05)	0,49 (0,02)	0,35 (0,03)
F1 (écart-type)	0,59 (0,42)	0,24 (0,04)	0,49 (0,05)	0,47 (0,02)	0,28 (0,04)
F1-A1 (écart-type)	0,50 (0,50)	0,53 (0,12)	0,72 (0,07)	0,69 (0,05)	0,59 (0,07)
F1-A2 (écart-type)	0,63 (0,40)	0,34 (0,05)	0,52 (0,09)	0,50 (0,04)	0,28 (0,09)
F1-B1 (écart-type)	0,59 (0,44)	0,11 (0,10)	0,30 (0,12)	0,24 (0,09)	0,06 (0,06)
F1-B2 (écart-type)	0,58 (0,44)	0,17 (0,09)	0,32 (0,09)	0,36 (0,11)	0,04 (0,05)
F1-C (écart-type)	0,63 (0,43)	0,05 (0,06)	0,60 (0,04)	0,58 (0,04)	0,44 (0,05)

(a) Résultat pour la tâche de prédiction du niveau de facilité d'écoute sur l'ensemble des données

	CamemBERT	wav2vec	SVM-FL	SVM-L	SVM-F
Exactitude (écart-type)	0,58 (0,36)	0,35 (0,04)	0,56 (0,04)	0,52 (0,04)	0,37 (0,04)
F1 (écart-type)	0,50 (0,43)	0,23 (0,07)	0,54 (0,04)	0,50 (0,06)	0,33 (0,03)
F1-A1 (écart-type)	0,52 (0,45)	0,53 (0,05)	0,74 (0,08)	0,70 (0,08)	0,65 (0,10)
F1-A2 (écart-type)	0,55 (0,43)	0,31 (0,08)	0,54 (0,11)	0,52 (0,10)	0,29 (0,07)
F1-B1 (écart-type)	0,48 (0,45)	0,11 (0,12)	0,39 (0,09)	0,25 (0,10)	0,27 (0,07)
F1-B2 (écart-type)	0,42 (0,48)	0,16 (0,14)	0,43 (0,17)	0,41 (0,15)	0,18 (0,12)
F1-C (écart-type)	0,53 (0,48)	0,06 (0,13)	0,60 (0,07)	0,65 (0,09)	0,26 (0,15)

(b) Résultat pour la tâche de prédiction du niveau de facilité d'écoute sur les données de style dialogue

	CamemBERT	wav2vec	SVM-FL	SVM-L	SVM-F
Exactitude (écart-type)	0,72 (0,36)	0,31 (0,07)	0,52 (0,10)	0,49 (0,08)	0,37 (0,04)
F1 (écart-type)	0,67 (0,43)	0,22 (0,07)	0,51 (0,09)	0,47 (0,08)	0,31 (0,03)
F1-A1 (écart-type)	0,63 (0,48)	0,44 (0,18)	0,68 (0,13)	0,56 (0,16)	0,55 (0,18)
F1-A2 (écart-type)	0,66 (0,45)	0,38 (0,07)	0,55 (0,12)	0,51 (0,11)	0,31 (0,13)
F1-B1 (écart-type)	0,65 (0,46)	0,00 (0,00)	0,36 (0,15)	0,27 (0,15)	0,09 (0,12)
F1-B2 (écart-type)	0,67 (0,44)	0,09 (0,10)	0,36 (0,22)	0,40 (0,19)	0,09 (0,11)
F1-C (écart-type)	0,75 (0,36)	0,18 (0,16)	0,59 (0,09)	0,59 (0,12)	0,50 (0,06)

(c) Résultat pour la tâche de prédiction du niveau de facilité d'écoute sur les données de style monologue

TABLE 2 – Résultat pour la tâche de prédiction du niveau de facilité d'écoute

les dialogues est plus aisée que sur les monologues. En ce qui concerne la nature des prédicteurs, les performances de CamemBERT et même des SVMs montrent que la transcription et les variables qui en sont dérivées contribuent plus aux performances que les informations prosodiques. Néanmoins, nous avons identifié que les caractéristiques du signal vocal ajoutent bien de l'information complémentaire pertinente pour les modèles.

5 Conclusion et perspectives

Cette étude a comparé divers modèles en vue de prédire automatiquement le niveau de facilité d'écoute de documents audios en français. Tout d'abord, aussi bien le SVM-L utilisant uniquement les variables de lisibilité que le CamemBERT affiné obtiennent une F1 relativement élevée. Sur l'ensemble de données et sur les monologues, la F1 de CamemBERT est légèrement supérieure à celle du SVM-L. Cependant, dans cette étude, CamemBERT souffre de problèmes d'échantillonnage

causant de grandes disparités d'une session d'entraînement à l'autre, ce que révèle l'écart-type élevé des métriques d'évaluation estimées pour CamemBERT. Ensuite, l'analyse des résultats de SVM-F, utilisant uniquement les variables de fluidité, et de wav2vec a révélé que prédire à partir des seules caractéristiques prosodiques ne permettait pas d'atteindre des résultats très élevé, surtout pour les niveaux plus élevés. Enfin, nous avons constaté que les variables de fluidité apportaient tout de même des informations utiles au modèle lorsqu'elles sont combinées avec les variables de lisibilité (SVM-FL). Il a également été constaté que les différents styles de parole produisent des résultats différents en termes de performance du modèle, mais que la plupart des modèles se comportent mieux sur les dialogues.

Enfin, il convient de noter que la F1 par niveau est très différente d'un niveau à l'autre, pour tous les modèles analysés dans cette étude. Dans l'ensemble, on constate que la F1 est faible pour les niveaux B1 et B2, qui sont des niveaux intermédiaires du CECR. L'examen minutieux des variables les plus aptes à discriminer les niveaux B est l'une des tâches futures. Une autre perspective serait de combiner wav2vec, qui a atteint une F1 d'environ 0,3 à 0,5 aux niveaux A1 et A2, et CamemBERT affiné, qui a eu une bonne performance de prédiction en général. L'analyse de ce modèle combiné, reposant sur des plongements, constitue également une perspective intéressante qui permettrait de déterminer dans quelle mesure ces deux modèles sont complémentaires en matière d'information. Dans le même ordre d'idée, il serait possible d'explorer des architectures hybrides combinant des plongements de mots avec des variables de lisibilité.

Références

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, **33**, 12449–12460.

CARTIER F. A. (1952). The social context of listenability research. *Journal of Communication*, **2 (1)**, 44–47. DOI : 10.1111/j.1460-2466.1952.tb00177.x.

CARTIER F. A. (1955). Ii. listenability and "human interest". *Communications Monographs*, **22 (1)**, 53–57. DOI : 10.1080/03637755509375134.

CHALL J. S. & DIAL H. E. (1948). Predicting listener understanding and interest in newscasts. *Educational Research Bulletin*, **27 (6)**, 141–168.

CUCCHIARINI C., STRIK H. & BOVES L. (2002). Quantitative assessment of second language learners' fluency : Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, **111(6)**, 2862–2873.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

FANG I. E. (1967). The "easy listening formula". *Journal of Broadcasting & Electronic Media*, **11 (1)**, 63–68.

FLESCH R. (1943). Marks of readable style, a study in adult education. *Teachers College Contributions to Education*.

FRANÇOIS T. (2014). An analysis of a french as a foreign language corpus for readability assessment. In *Proceedings of the 3rd workshop on NLP for Computer-assisted Language Learning, NEALT Proceedings Series Vol. 22, Linköping Electronic Conference Proceedings 107*, p. 13–32.

FRANÇOIS T. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère*. Thèse de doctorat, Université Catholique de Louvain.

FRANÇOIS T., GALA N., WATRIN P. & FAIRON C. (2014). Flelex : a graded lexical resource for french foreign learners. *International conference on Language Resources and Evaluation*.

GINTHER A., DIMOVA S. & YANG R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral english proficiency with implications for automated scoring. *Language Testing*, **27**, 379–399.

GUYON I. & ELISSEEFF A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, **3**, 1157–1182.

HARWOOD K. A. (1950). A concept of listenability. *Western Speech*, **14 (2)**, 10–12. DOI : 10.1080/10570315009373409.

HARWOOD K. A. (1955). Iii. listenability and rate of presentation. *Communications Monographs*, **22(1)**, 57–59. DOI : 10.1080/03637755509375135.

HARWOOD K. A. & CARTIER F. (1952). On definition of listenability. *Southern Journal of Communication*, **18 (1)**, 20–23. DOI : 10.1080/10417945209371245.

KANDEL L. & MOLES A. (1958). Application de l'indice de Flesch à la langue française. *Cahiers Études de Radio-Télévision*, **19**, 253–274.

KIYOKAWA H. (1990). A formula for predicting listenability : The listenability of english language materials 2. *Wayo Women's University Language and Literature*, **24**, 57–74.

KOTANI K., UEDA S., YOSHIMI T. & NANJO H. (2014). A listenability measuring method for an adaptive computer-assisted language learningand teaching system. *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, p. 387–394.

KOTANI K. & YOSHIMI T. (2017). Effectiveness of linguistic and learner features for listenability measurement using a decision tree classifier. *The Journal of Information and Systems in Education*, **16 (1)**, 7–11.

KUMAI N. (1992). Teaching listening comprehension : What and how (in japanese). *English and American Studies*, **27**, 21–30.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MARTINC M., POLLAK S. & ROBNIK-ŠIKONJA M. (2021). Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, **47**(1), 141–179.

O'KEEFE T. (1971). The comparative listenability of shortwave broadcasts. *Journalism Quarterly*, **48 (4)**, 744–748.

PELTONEN P. (2017). Temporal fluency and problem-solving in interaction : An exploratory study of fluency resources in l2 dialogue. *System*, **70(C)**, 1–13.

PRÉFONTAINE Y. (2010). Differences in perceived fluency and utterance fluency across speech elicitation tasks : a pilot study. *Papers from the Lancaster University Postgraduate Conference in Linguistics Language Teaching*, p. 134–154.

QI P., ZHANG Y., ZHANG Y., BOLTON J. & MANNING C. D. (2020). Stanza : A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 101–108.

ROGERS J. R. (1962). A formula for predicting the comprehension level of material to be presented orally. *The journal of educational research*, **56 (4)**, 218–220. DOI : 10.1080/00220671.1962.10882926.

RUBIN D. L. (2012). Listenability as a tool for advancing health literacy. *Journal of health communication*, **17 (3)**, 176–190. DOI : 10.1080/10810730.2012.712622.

RUGGIA S. (2019). Le deep learning : un outil pour la didactique du fle ? *Dialettica pedagogica*, **1**, 79–106.

RUGGIA S. (2020). Caractériser un texte en français : les passages-clés des niveaux a1 et a2 du cecrl. *Actes des 15èmes Journées internationales d'Analyse statistique des Données Textuelles*, p. 1–11.

RUGGIA S. (2021). Deepfle : l'intelligence artificielle pour décrire et prédire le(s) niveau(x) du cecrl d'un texte. *Les cahiers de l'AREFLE*, **2 (1)**, 103–109.

RÉVÉSZ A. & BRUNFAUT T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, **35**, 31–65.

SEGALOWITZ N., FRENCH L. & GUAY J.-D. (2017). What features best characterize adult second language utterance fluency and what do they reveal about fluency gains in short-term immersion ? *Revue canadienne de linguistique appliquée*, **20(2)**, 90–116.

SKEHAN P. & FOSTER P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, **49**, 93–120.

UEDA S., NANJO H., YOSHIMI T. & KOTANI K. (2013). A listenability formula considering listening proficiency level information of learners of english as a foreign language (in japanese). *The Association for Natural Language Processing, NLP2013*, p. 410–413.

VAJJALA S. (2021). Trends, limitations and open challenges in automatic readability assessment research. *arXiv preprint arXiv :2105.00973*.

WILKENS R., ALFTER D., WANG X., PINTARD A., TACK A., YANCEY K. & FRANÇOIS T. (2022). Fabra : French aggregator-based readability assessment toolkit. *Proceedings of the 13th Language Resources and Evaluation Conference*, p. 1217–1233.

YANCEY K., PINTARD A. & FRANCOIS T. (2021). Investigating readability of french as a foreign language with deep learning and cognitive and pedagogical features. *Lingue e Linguaggio*, **2021**(2), 229–258.

YOSHIMI K. (2019). The effectiveness and prospects of incorporating current topics into the learning of listening skills : A practical example of listening comprehension 1/2 in the school of contemporary international studies (in japanese). *Bulletin of Nagoya University of Foreign Studies*, **5**, 105–119.

YOSHIMI T. & KOTANI K. (2020). Non-linear regression analysis of the combined listening ease and accuracy index appropriate for learners' proficiency (in japanese). *Transactions of Japanese Society for Information and Systems in Education*, **37 (1)**, 44–49.

Annexes

Manuel	Éditeur	Année	Nb de données
A1			
Alter Ego	Hachette	2006	26
écho	CLE International	2011	63
Texto	Hachette	2016	19
Cosmopolite	Hachette	2017	104
Atelier	Didier	2019	78
A2			
Entre Nous	Maison des langues	2016	70
Tendances	CLE International	2016	81
Texto	Hachette	2016	16
Cosmopolite	Hachette	2017	72
Défi	Maison des langues	2018	35
Atelier	Didier	2019	35
B1			
Entre Nous	Maison des langues	2016	72
Tendances	CLE International	2016	90
Cosmopolite	Hachette	2018	51
Défi	Maison des langues	2019	27
B2			
Édito	Didier	2006	31
LE DELF B2	Didier	2016	60
Entre Nous	Maison des langues	2017	20
Tendances	CLE International	2017	50
Cosmopolite	Hachette	2019	80
C1/C2			
Réussir le DALF	Didier	2007	31
abc DALF	CLE International	2014	57
Le DALF 100 % réussite	Didier	2017	75
Édito	Didier	2018	70
Tendances	CLE International	2019	10

TABLE 3 – Liste des manuels

Variable	Méthode de calcul
Nombre de syllabes par minute	Nombre total de syllabes / durée totale d'enregistrement [minute].
Nombre de syllabes par minute sans les pauses	Nombre total de syllabes / (durée totale d'enregistrement [minute] - durée totale de pauses [minute]).
Nombre de mots par minute	Nombre total de mots / durée totale d'enregistrement [minute].
Nombre de mots par minute sans les pauses	Nombre total de mots / (durée totale d'enregistrement [minute] - durée totale de pauses [minute]).
Durée totale	Durée totale d'un enregistrement.

Durée moyenne des pauses	Durée totale de pauses [minute] / nombre total de pauses.
Nombre de pauses par minute	Nombre total de pauses / durée totale d'enregistrement [minute].

TABLE 4 – Liste des variables de fluidité

Famille	Variable	Description
Basé sur la longueur		
Longueur du mot	LENwrdSYL	Nombre de syllabes par mot.
Longueur de la phrase	LENsntWRD	Nombre de tokens par phrase, à l'exclusion de la ponctuation.
Variables lexicales		
Chevauchement de contenu	LEXcovLGAL	Tout lemme est partagé dans toutes les phrases.
Diversité lexicale	LEXdvrFLC	CTTR de tous les types de lemmes de noms, de noms propres, de verbes, d'adjectifs et d'adverbes dans le texte, en tenant compte de tous les tokens.
Fréquence lexicale	LEXfrqFCL	Fréquence de la forme du lemme de tous les noms, noms propres, verbes, adjectifs et adverbes en fonction de leur occurrence dans le corpus FLELex.
	LEXfrqLCL	Fréquence de la forme du lemme de tous les noms, noms propres, verbes, adjectifs et adverbes sur la base de l'occurrence dans le corpus Lexique3.
Lexiques gradués	LEXgrdBA1	Proportion de mots dans les descripteurs de niveau de référence du français de Beacco pour le niveau du CECR A1.
	LEXgrdFA1	Fréquence des mots dans la ressource FLELex pour le niveau du CECR A1.
	LEXgrdFFOA1	Proportion de lemmes de niveau A1 selon la méthode de la première occurrence (ni-veau du CECR où un mot est rencontré pour la première fois).
	LEXgrdFMLA1	Proportion de lemmes de niveau A1 selon la méthode d'apprentissage automatique de Pintard et François, 2020 (https ://aclanthology.org/2020.readi-1.13.pdf) entraînée sur les descripteurs de niveau de référence Beacco.
	LEXgrdFSOOUA1	Proportion de lemmes de niveau A1 selon la méthode Significant Onset of Use (Alfter et al., 2016 ; https ://aclanthology.org/W16-6501.pdf)
Voisins orthographiques	LEXnghPHO	Distance phonologique moyenne de Le-venstein, calculée sur Lexique3.
Normes lexicales	LEXnrmCNCR	Niveau de concrétion des mots.
	LEXnrmFAM	La familiarité des mots, également appe-lée fréquence subjective.
	LEXnrmIMG	Imageabilité de chaque mot.

Sophistication lexicale	LEXsopFK1	Nombre de lemmes dans les premières bandes de fréquences de 1000 mots de FLELex.
	LEXsopGK1	Nombre de mots dans les premières bandes de fréquence de 1000 mots de la liste de vocabulaire de Gougenheim.
	LEXsopLLK1	Nombre de lemmes dans les premières bandes de fréquence de 1000 mots de Lexique3.
Caractéristiques graduées	LEXgrdBLA1	Expressions lexicales de niveau A1 dans le chapitre 5 de Beacco.
Variables syntaxiques		
Développement du langage	SYNdevNPHRS	Nombre de constituants/phrases.
	SYNdevTUL	Longueur des T unités.
	SYNdevVG1	Nombre de verbes du 1er groupe français dans le texte.
Caractéristiques de la clause	SYNclsLEN	Longueur de la clause.
Caractéristiques des temps verbaux	SYNtnsfINDP	Indicatif présent.
	SYNtnsfINDI	Indicatif imparfait.
	SYNtnsfINDPS	Indicatif passé simple.
	SYNtnsfINDPC	Indicatif passé composé.
	SYNtnsfINDPQP	Indicatif plus-que-parfait.
	SYNtnsfINDPA	Indicatif passé antérieur.
	SYNtnsfINDFS	Indicatif futur simple.
	SYNtnsfINDFA	Indicatif futur antérieur.
	SYNtnsfCNDP	Conditionnel présent.
	SYNtnsfCNDPS	Conditionnel passé.
	SYNtnsfIMPP	Impératif présent.
	SYNtnsfIMPPS	Impératif passé.
	SYNtnsfSUBP	Subjonctif présent.
	SYNtnsfSUBPS	Subjonctif passé.
	SYNtnsfSUBI	Subjonctif imparfait.
	SYNtnsfSUBPQP	Subjonctif plus-que-parfait.
	SYNtnsfINF	Infinitif.
	SYNtnsfINFC	Infinitif composé.
	SYNtnsfPP	Participe présent.
	SYNtnsfPPS	Participe passé.
	SYNtnsfGERP	Gérondif.
	SYNtnsfGERPS	Gérondif passé.
	SYNtnsfTAP	Temps passif.
POS Tag	SYNposADJ	Mots identifiés comme ADJ (adjectif), suivant les étiquettes POS universelles et l'analyseur Stanza.
	SYNposADP	Mots identifiés comme ADP (adposi-tion), suivant les étiquettes POS univer-selles et l'analyseur Stanza.
	SYNposADV	Mots identifiés comme ADV (adverbe), suivant les étiquettes POS universelles et l'analyseur Stanza.

	SYNposAUX	Mots identifiés comme AUX (auxiliaire), suivant les étiquettes POS universelles et l'analyseur Stanza.
	SYNposCCONJ	Mots identifiés comme CCONJ (conjonc-tion de coordination), suivant les éti-quettes POS universelles et l'analyseur Stanza.
	SYNposINTJ	Mots identifiés comme INTJ (interjec-tion), suivant les étiquettes POS univer-selles et l'analyseur Stanza.
	SYNposNOUN	Mots identifiés comme NOUN (nom), suivant les étiquettes POS universelles et l'analyseur Stanza.
	SYNposNUM	Mots identifiés comme NUM (numéral), suivant les étiquettes POS universelles et l'analyseur Stanza.
	SYNposPART	Mots identifiés comme PART (particule), suivant les étiquettes POS universelles et l'analyseur Stanza.
	SYNposPRON	Mots identifiés comme PRON (pronom), suivant les étiquettes POS universelles et l'analyseur Stanza.
	SYNposPROPN	Mots identifiés comme PROPN (nom propre), suivant les étiquettes POS uni-verselles et l'analyseur Stanza.
	SYNposSCONJ	Mots identifiés comme SCONJ (conjonc-tion de subordination), suivant les éti-quettes POS universelles et l'analyseur Stanza.
	SYNposVERB	Mots identifiés comme VERB (verbe), suivant les étiquettes POS universelles et l'analyseur Stanza.
	SYNposX	Mots identifiés comme X (autre), suivant les étiquettes POS universelles et l'analyseur Stanza.
Variables de lisibilité		
Service de formule de lisibilité	REAfrmKM	Formule de lisibilité pour le français (Kandel-Moles).

TABLE 5 – Liste des variables de lisibilité

Variables	Corrélation Spearman	Variables	Corrélation Spearman
LENsntWRD	0,63	LEXgrdFSOOUA1	-0,61
syllables_per_minute	0,57	REAfrmKM	-0,56
LEXdvrFLC	0,47	LEXgrdBA1	-0,45
SYNtnsfPPS	0,41	pauses_per_minute	-0,35
SYNtnsfPP	0,37	length_pauses	-0,34
SYNtnsfTAP	0,36	SYNtnsfINDP	-0,30
SYNposADP	0,36	SYNposINTJ	-0,25
SYNposSCONJ	0,31	LEXnrmIMG	-0,24
SYNtnsfSUBP	0,30		
SYNposNOUN	0,29		
LEXnghPHO	0,27		
SYNtnsfINDI	0,26		
SYNposADJ	0,25		
SYNtnsfINDPQP	0,25		
SYNtnsfINDFS	0,23		
SYNtnsfINF	0,22		

TABLE 6 – 24 variables finales et leurs coefficients de corrélation de Spearman avec le niveau du CECR dans l'ensemble des données (p<,01)

Variables	Corrélation Spearman	Variables	Corrélation Spearman
LENsntWRD	0,65	LEXgrdFSOOUA1	-0,61
syllables_per_minute	0,64	REAfrmKM	-0,57
LEXdvrFLC	0,50	LEXgrdBA1	-0,45
SYNposADP	0,40	pauses_per_minute	-0,43
SYNposSCONJ	0,40	SYNtnsfINDP	-0,38
SYNtnsfTAP	0,40	length_pauses	-0,37
SYNtnsfPP	0,39	SYNposINTJ	-0,35
SYNtnsfPPS	0,37	LEXnrmIMG	-0,34
SYNtnsfSUBP	0,36		
SYNtnsfINDI	0,35		
SYNtnsfINDPQP	0,32		
SYNposNOUN	0,27		
SYNtnsfINDFS	0,26		
SYNposX	0,24		
SYNposADJ	0,23		
SYNtnsfCNDPS	0,23		
LEXnghPHO	0,22		
SYNtnsfINDPC	0,22		
SYNtnsfINF	0,21		
SYNtnsfINFC	0,21		
SYNtnsfINDPS	0,20		

TABLE 7 – 29 variables finales et leurs coefficients de corrélation de Spearman avec le niveau du CECR dans le dialogue (p<,01)

Variables	Corrélation Spearman	Variables	Corrélation Spearman
SYNdevTUL	0,57	LEXgrdFMLA1	-0,61
LEXdvrFLC	0,51	REAfrmKM	-0,53
SYNtnsfPPS	0,45	LEXgrdBA1	-0,42
syllables_per_minute	0,38	length_pauses	-0,26
SYNtnsfPP	0,36	SYNposPRON	-0,22
SYNtnsfTAP	0,31	pauses_per_minute	-0,22
LEXnghPHO	0,30		
SYNposNOUN	0,27		
SYNposADV	0,26		
SYNposADP	0,26		
SYNposSCONJ	0,24		
SYNtnsfSUBP	0,23		
SYNposADJ	0,22		
SYNtnsfINF	0,21		
SYNtnsfINDFS	0,21		

TABLE 8 – 21 variables finales et leurs coefficients de corrélation de Spearman avec le niveau du CECR dans le monologue (p<,01)

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Optimisation des performances d'un système de reconnaissance automatique de la parole pour les commentaires sportifs : fine-tuning de Whisper

Camille Lavigne Alex Stasica[1] Anna Kupsc[2]
(1) Utrecht University, Utrecht,3500-3585, Pays-Bas
(2) CLLE-ERSSàB-UMR-5263, Université Bordeaux Montaigne, 33600, Pessac, France
lavignecamille@yahoo.fr, a.stasica@uu.nl,
anna.kupsc@u-bordeaux-montaigne.fr

RÉSUMÉ

Malgré les performances élevées des systèmes automatiques de reconnaissance de la parole (Automatic Speech Recognition ; ASR) sur des corpus généraux, leur efficacité est considérablement réduite lorsqu'ils sont confrontés à des corpus spécialisés. Ces corpus peuvent notamment contenir du lexique propre à des domaines spécifiques, des accents ou du bruit de fond rendant la transcription ardue. Cette étude vise à évaluer les avantages de l'optimisation d'une transcription automatique, par opposition à manuelle, après *fine-tuning* d'un modèle d'ASR de dernière génération, Whisper (Radford *et al.*, 2023), sur un corpus spécialisé de commentaires sportifs de petite taille. Nos analyses quantitatives et qualitatives indiquent que Whisper est capable d'apprendre les particularités d'un corpus de spécialité, atteignant des performances égales où supérieures aux transcripteurs humains, avec une quantité de données limitée. Cette recherche met en lumière le rôle que l'intelligence artificielle, notamment les larges modèles de langage, peut jouer pour faciliter la création de corpus spécialisés.

ABSTRACT

Performance optimization of an automatic speech recognition system for sport commentaries : Whisper fine-tuning

Despite the great performance of automatic speech recognition (ASR) systems on general corpora, their performance is greatly impacted when confronted with specialized corpora. These corpora can include specialized lexicon, accents or background noise in the audio making the transcription harder. This study aims to evaluate the potential benefits of optimizing automatic transcription, as opposed to a manual one, by *fine-tuning* a state-of-the-art ASR model, namely Whisper (Radford *et al.*, 2023), on a small-size specialized corpus of sport commentaries. Our quantative and qualitative analyses indicate that Whisper is capable of learning the features of our specialized corpus, reaching equal or higher to human transcribers performance, with a limited quantity of data. This research highlights the role that artificial intelligence, particularly large language models, can play in facilitating the creation of specialized corpora.

MOTS-CLÉS : Whisper ; large modèle de langage ; fine-tuning ; corpus de spécialité ; reconnaissance automatique de la parole.

KEYWORDS: Whisper ; large language model ; fine-tuning, specialized corpus ; automatic speech recognition.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 567–581.

1 Introduction

La recherche sur les corpus oraux en France a connu un développement plus lent que celle des corpus écrits. Malgré la présence de quelques corpus oraux dès les années 70-80, comme ESLO Baude & Dugua (2016), leur exploitation a été freinée par des difficultés de transcription, d'annotation et d'automatisation de l'analyse, en raison des particularités de la langue parlée. Les corpus de langue parlée, bien qu'ils ne soient pas nécessairement des corpus de référence, soulignent la diversité des situations de parole, des locuteurs et des dialectes d'une langue. Les exigences nécessaires à l'exploitation des corpus oraux ont freiné leur développement. Par conséquent, bien que ces corpus permettent l'analyse de phénomènes linguistiques particuliers, leur utilisation pour l'étude de genres de discours spécifiques reste souvent limitée en raison de la difficulté à constituer des corpus de grande taille. De nos jours, malgré la présence de nombreux corpus oraux, les corpus de discours de spécialité restent rares.

Cet article se concentre sur le commentaire sportif en direct, un genre de discours médiatique et journalistique présentant des caractéristiques linguistiques et prosodiques particulières. Ce genre reste sous-étudié, souvent analysé sur de courts extraits et peu diversifiés en termes de sports (le plus souvent des sports d'équipes et de ballon de type football, basket, rugby). Pourtant, il présente des particularités linguistiques tout à fait intéressantes pour différents domaines de la linguistique (p.ex. syntaxe, lexicologie, prosodie ; voir Augendre *et al.* 2018; Fontagnol *et al.* 2023) et du TAL. En effet, la transcription automatique de ce type de production de parole pose un grand nombre de difficultés pour les modèles de reconnaissance automatique de la parole (ASR [1]) à cause du bruit de fond important, l'utilisation du lexique spécifique au sport ou un grand nombre d'entités nommées. Dans le même temps, les études qui s'intéressent à ce genre de discours, si elles sont fondées sur corpus, le sont très souvent sur des corpus de petite taille, car très longs à transcrire, et qui ne font pas l'objet d'une diffusion large auprès de la communauté. De ce fait, un moyen d'aligner et de transcrire automatiquement les données orales ou multimodales fournies par ce genre de corpus permettrait très certainement de développer les données prêtes pour l'analyse et de faciliter la description de ce discours de spécialité, tout en analysant un plus grand nombre de données. En effet, pour le moment la transcription de notre corpus est basé sur un système de *student sourcing* (Stasica *et al.*, 2023) : une cinquantaine d'étudiants transcrit chaque année quelques minutes d'un enregistrement de notre corpus et au vu du temps de transcription, de correction et de vérification par les chercheurs, seulement un match maximum est transcrit par année universitaire.

Nous proposons dans cet article de discuter de l'apport que constituent les ASR pour le traitement de corpus oraux de spécialité. Plus spécifiquement, nous détaillons une première évaluation de la performance de la famille de modèles Whisper (Radford *et al.*, 2023) sur la transcription orthographique d'un corpus multimodal de petite taille (9h30 heures d'enregistrement) de commentaires télévisuels en direct de matchs de rugby. Nous étudions également l'apport de modèles *fine-tuned*(FM [2]) sur notre corpus, permettant de combler les limites des modèles pré-entraînés (PM [3]) sur une des particularités de notre corpus : le lexique de spécialité. Nous explorons ainsi la possibilité qu'après un *fine-tuning* sur un petit ensemble de données, ces modèles atteignent des performances comparables à celles des transcripteurs humains.

Cet article est structuré comme suit : dans la section 2 nous décrivons l'état de l'art des modèles

1. Automatic Speech Recognition
2. *fine-tuned* Models
3. *Pre-trained Models*

d'ASR. Dans la section 3, nous décrivons notre méthodologie et notre corpus. Dans la section 4, nous présentons les résultats de notre recherche et enfin dans la section 5, nous discutons des perspectives possibles pour de futures recherches.

2 Etat de l'art

Depuis l'introduction du premier ASR il y a plusieurs décennies (Davis *et al.*, 1952), de nombreuses méthodologies ont été élaborées pour améliorer la précision de ces modèles, convertissant les ondes sonores en impulsions électriques. Plusieurs revues de la littérature existantes ont été publiées pour examiner l'évolution des différentes approches utilisées dans le développement des ASR (Ghai & Singh, 2012; Karpagavalli & Chandra, 2016). Pendant longtemps, le modèle de Markov caché (HMM [4]) associé au modèle de mélange gaussien (GMM [5]) est resté dominant et le plus performant dans le domaine de la reconnaissance vocale. Cependant, l'avènement des techniques du *deep learning* a ouvert la voie à des modèles surpassant les performances des HMM-GMM, notamment les modèles HMM-*Deep Neural Network* (HMM-DNN) et les modèles *end-to-end* (Wang *et al.*, 2019).

Les récents modèles d'ASR ont abandonné l'utilisation des HMM, mais ont continué à exploiter les techniques de *deep learning*. Parmi les modèles ayant révolutionné la reconnaissance vocale, on trouve le modèle Wav2Vec (Schneider *et al.*, 2019). Il s'agit du premier modèle à adopter une approche de pré-entraînement non supervisé, apprenant à partir de données audio non transcrites. Ce pré-entraînement permet à Wav2Vec d'apprendre des représentations générales de la parole, exploitables pour améliorer les performances sur des tâches ultérieures où les données annotées sont limitées. C'est le cas pour des tâches telles que la reconnaissance vocale, où l'annotation de données est fastidieuse et rare.

L'architecture de Wav2Vec se compose d'un réseau neuronal convolutif (CNN) suivi d'un réseau de neurones basé sur des *transformers*. Le CNN encode les formes d'ondes audio brutes pour extraire des représentations de haut niveau, puis les transmet à l'encodeur du *transformer* pour un traitement ultérieur, permettant la capture d'informations contextuelles.

Après le pré-entraînement, Wav2Vec peut être *fine-tuned* sur des données annotées spécifiques à des tâches d'ASR. Pendant cette phase de *fine-tuning*, le modèle apprend de manière supervisée, adaptant les représentations pré-entraînées aux caractéristiques spécifiques des données cibles, ce qui améliore les performances sur les tâches de reconnaissance vocale. L'utilisation d'un pré-entraînement non supervisé permet l'exploitation d'une plus grande quantité de données, car il n'est pas nécessaire de compléter les données sources avec des transcriptions. Wav2Vec maintient actuellement des performances de pointe sur divers ensembles de tests ASR, même lorsqu'il est formé sur des données annotées limitées.

Cependant, les encodeurs pré-entraînés entièrement non supervisés, tels que Wav2vec, présentent des limites. Selon Radford *et al.* (2023), bien qu'ils apprennent des représentations vocales de haute qualité, ils souffrent d'un manque de décodeur aussi performant pour associer ces représentations à des sorties utilisables. Cette lacune nécessite une étape de *fine-tuning* pour des tâches telles que la reconnaissance vocale, ce qui peut être complexe et limiter leur utilité.

Les auteurs notent également que des études ont démontré que les ASR pré-entraînés sur plusieurs en-

4. Hidden Markov Model
5. Gaussian Mixture Model

sembles de données ou domaines sont plus robustes et généralisent plus efficacement. Par conséquent, des recherches ont mis en avant l'utilisation de données audio avec des transcriptions de référence où l'exigence de validation manuelle a été assouplie, permettant l'utilisation d'un volume accru de données (p.ex. de 10 à 30 000 heures d'audio) pour l'entraînement des ASR. Cela permet d'établir un équilibre entre qualité et quantité de données.

Ainsi, Radford *et al.* (2023) proposent les modèles Whisper pour pallier les problèmes cités ci-dessus, étendant la reconnaissance vocale faiblement supervisée à un vaste ensemble de données comprenant 680,000 heures de données audio transcrites. Ces modèles incluent un entraînement multilingue et multitâche, et produisent des résultats comparables aux modèles de pointe actuellement disponibles. Les auteurs choisissent une architecture *transformer* d'encodeur-décodeur (Vaswani *et al.*, 2017), car cette architecture a montré son efficacité pour la génération de texte.

Ainsi, Whisper se distingue des modèles précédents. Contrairement à ses prédécesseurs, en tant que modèle *end-to-end*, Whisper ne requiert pas de *fine-tuning* avant son application à un ensemble de données, ce qui lui confère une robustesse surpassant les autres modèles, même si, ses performances sont plus faibles sur des données moins représentées lors de son entraînement (p.ex. Jain *et al.* (2023a)). En effet, ayant été entraîné sur une grande variété de données audio et évalué dans un cadre *zero-shot*, c'est-à-dire sur des ensembles de données différents de ceux de l'entraînement, il est capable de généraliser à travers divers domaines et tâches. Les auteurs visent ainsi à développer un système de traitement vocal fonctionnel adaptable à différents domaines, tâches et langues, sans nécessiter de réglages supplémentaires, ce qui le différencie des modèles antérieurs, souvent limités dans leur capacité à faire des généralisations. En effet, même de légères divergences entre les ensembles d'entraînement et de test altèrent significativement les performances des modèles précédents (Radford *et al.*, 2023).

Malgré les atouts de Whisper par rapport aux autres modèles, il convient de souligner que, dans le cadre de notre recherche sur les corpus de spécialité, certaines fonctionnalités requises ne sont pas intégrées dans les modèles Whisper de base. En particulier, ces modèles ne sont pas capables de diariser, c'est-à-dire de reconnaître les différents locuteurs, et leur alignement au signal sonore est mal adapté, fonctionnant davantage comme un générateur de sous-titres, pas assez précis pour notre objectif de recherche. Malgré la robustesse du modèle, Radford *et al.* (2023) admettent la nécessité d'étudier le *fine-tuning* de Whisper pour améliorer ses performances dans des contextes spécifiques. Par exemple, des adaptations ont été nécessaires pour décrire la parole enfantine (Jain *et al.*, 2023a), pour des langues peu représentées dans les données d'entraînement (comme le roumain (Păis *et al.*, 2023) ou le turc (Oyucu, 2023)), ainsi que pour des langues minoritaires ou mixtes (Xie *et al.*, 2023).

3 Corpus et méthodologie

3.1 Données

3.1.1 Données brutes

Notre corpus est un corpus oral de spécialité consistant en plusieurs matchs de rugby de deux coupes du monde différentes : celle de 2007 (Lortal & Mathon, 2008) et celle de 2015. Pour la première nous avons 36 heures d'enregistrement audio pour certains matchs et vidéos pour d'autres et pour la seconde nous possédons 25 heures d'enregistrement vidéo. Ce corpus est considéré de spécialité

de par ces nombreuses particularités linguistiques, notamment lexicales (Fontagnol *et al.*, 2023) et syntaxiques (Augendre *et al.*, 2018) avec une présence moindre de structures verbales et une proéminence de structures nominales spécifiques (p.ex. entités nommées) et de lexique spécifique au rugby et au commentaire sportif.

L'omniprésence de ce lexique spécifique dans notre corpus pourrait potentiellement affecter les performances des modèles d'ASR. C'est pourquoi le *fine-tuning* serait nécessaire pour permettre au modèle d'acquérir ces spécificités lexicales et ainsi augmenter sa performance jusqu'à atteindre celle d'un transcripteur humain.

La production des transcriptions se fait en 2 phases. La première phase repose sur le *student sourcing* pour obtenir une première transcription de l'audio, alignement compris. Ces transcriptions sont ensuite validées par un transcripteur expert (ie. enseignant chercheur ou doctorant). Toutes les transcriptions sont réalisées sur Transcriber (Barras *et al.*, 2001), et l'annotation d'actions de jeu sur les images des vidéos de deux matchs sont réalisées avec Aegissub [6]. Néanmoins, ces annotations ne sont pas incluses dans la présente étude.

Tous les matchs ne sont pas totalement transcrits et annotés. Pour cette raison, nous n'utilisons comme corpus de travail pour cette recherche que six matchs pour lesquels nous possédons une transcription «gold», c'est-à-dire vérifiée par un transcripteur expert. Les six matchs cumulent 9h30 d'audio. Ces 9h30 d'audio transcrits contiennent 94514 mots soit 138751 tokens (mots et la ponctuation) après tokenization. Pour chaque match, la transcription est orthographique, alignée au signal sonore, notant les pauses silencieuses d'au moins 200 ms comme requis dans l'état de l'art (Candea, 2000) et distingue les différents locuteurs présents dans l'audio.

Le tableau 1 (Annexe A) présente les deux commentateurs principaux, un journaliste et un expert commentant chaque match, le nom de chaque match du corpus de travail ainsi que l'année où il a été joué.

3.1.2 Pré-traitement des données

Le corpus de travail est préalablement divisé en un ensemble d'entraînement/validation et un ensemble de test. Pour les matchs utilisés dans l'ensemble d'entraînement, un échantillon sur six a été réservé pour la validation soit 1h20 d'audio (14% du corpus de travail). Les échantillons restant, 6h30 d'audio (68% du corpus de travail) constituent l'ensemble d'entraînement. Pour s'assurer de la capacité de généralisation après *fine-tuning*, nous réservons un match entier à l'ensemble de test soit 1h40 d'audio (18% du corpus de travail). Ce match, «Japon-Fidji», est donc *out-of-sample*, c'est-à-dire qu'aucun passage de ce match n'est donné au modèle durant l'entraînement. De plus, les commentateurs de ce match ne sont pas représentés dans l'ensemble d'entraînement. Pour chaque match l'audio et la transcription correspondante ont été découpés en tranches de 30 secondes, soit la taille maximale de la fenêtre de contexte de Whisper. Le découpage est fait sur les unités inter-pausales, définies par Nguyen *et al.* (2022) comme une détection d'activité vocale continue par un locuteur, délimitée par un silence de plus de 200 ms des deux côtés.

Conformément à la méthode utilisée par les auteurs de Whisper pour leur propre corpus, tous les enregistrements audio ont été rééchantillonnés à 16 000 Hz, et une représentation du Mel-spectrogramme en magnitude logarithmique à 80 canaux a été calculée sur des fenêtres de 25 millisecondes avec un pas de 10 millisecondes (Radford *et al.*, 2023).

6. https://aegisub.org/

3.2 Fine-tuning de Whisper

Comme mentionné précédemment, en raison de la fréquence élevée d'entités nommées et du lexique spécifique au rugby, les PM présentent généralement des performances limitées sur notre corpus. Ainsi, un processus de *fine-tuning* sur nos propres données s'avère nécessaire pour améliorer la reconnaissance de ce lexique spécifique. De plus, Whisper, comme tout modèle d'ASR, est sensible au bruit. Il est probable que les puissants bruits de fond particuliers à l'environnement des grands évènements sportifs dégradent la qualité de la transcription, dégradation que pourrait compenser un fine-tuning du modèle. Différentes méthodes de *fine-tuning* ont été proposées ces dernières années, parmi lesquelles le *Low-Rank Adaptation* (LoRA) (Hu *et al.*, 2021), que nous avons retenu pour sa sobriété computationnelle et sa facilité d'implémentation. L'utilisation de cette méthode nous a permis d'effectuer le *fine-tuning* de la famille de modèles Whisper du Tiny (33 millions de paramètres) au Large (1600 millions de paramètres) avec une unique NVDIA RTX 6000 24GB VRAM [7]. La *quantization* (Li *et al.*, 2023) couplée à l'utilisation d'un float half-précision (float 16bit) dégradaient fortement les performances des modèles lors de l'inférence. Nous avons donc entraîné nos modèles en full précision (float 32bit).

Concernant les hyper-paramètres de LoRA, le rang r=1, le *dropout* des matrices de poids a été fixé à 0.3 et alpha=64. Ce paramétrage permet de réduire l'*overfitting*, très problématique sur les modèles les plus larges. Le *learning rate* est fixé à 1e-3 avec *warm-up*. Le nombre d'epochs par défaut est fixé à 15 avec *early stopping*. L'entraînement d'un modèle sur les 6h30 heures d'audio des données d'entraînement prend entre 1h et 3h selon la taille du modèle.

3.3 Évaluation

L'enjeu du *fine-tuning* revêt une double importance : améliorer la qualité globale de la transcription (p.ex. transcription plus précise malgré le bruit de fond, meilleure performance en français), et affiner la transcription des éléments spécifiques à notre corpus, notamment le lexique propre aux commentaires sportifs pour les matchs de rugby.

Pour évaluer les améliorations apportées par nos modèles *fine-tuned*, nous avons adopté une approche double, à la fois quantitative et qualitative. Tout d'abord, l'évaluation quantitative suit le même protocole que celui utilisé dans les publications majeures en ASR ces dernières années, à savoir le calcul de la *Word Error Rate* (WER) entre la référence et une prédiction. Cette métrique présente néanmoins un défaut majeur dans notre cas. Elle pénalise de manière disproportionnée les hallucinations de Whisper [8], qui sont rares, facilement détectables et corrigeables en remplaçant la génération *greedy search* par du *sampling* avec *beam search* et température *schedduling* (Radford *et al.*, 2023). Puisque les transcriptions ayant une WER supérieure à 100 sont systématiquement des hallucinations, pour chaque segment de transcription évalué, nous définissons une borne supérieure telle que : WER $\in$ [0, 100].

Ensuite, nous avons procédé à des évaluations qualitatives afin de déterminer (i) si les performances des FM augmentent par rapport aux PM sur les éléments spécifiques à notre corpus de spécialité, et (ii) pour établir un lien entre son éventuelle amélioration quantitative et une meilleure compréhension

7. Les expériences présentées dans cet article ont été réalisées par l'intermédiaire de Gilles Boyé et de Catherine Mathon en utilisant la plateforme OSIRIM qui est administrée par l'IRIT et soutenue par CNRS, la région Midi-Pyrénées, le gouvernement français, et le FEDER (voir http://osirim.irit.fr/site/fr).

8. Un exemple d'hallucination qui peut être trouvé dans notre corpus : "alors on du mal au au au au au au au au au"

et transcription des caractéristiques de notre corpus. Plusieurs éléments peuvent être représentatifs de notre corpus, tels que les noms propres (p.ex. noms des joueurs, de l'arbitre, des commentateurs), dans la mesure où nous manquons de moyens efficaces pour permettre au modèle d'apprendre des noms propres qu'il n'aurait pas rencontrés lors de l'entraînement, nous nous concentrons sur le lexique spécialisé du rugby dans notre analyse qualitative.

Pour l'évaluation qualitative de la reconnaissance du lexique spécialisé, nous avons utilisé AnaText [9] afin de sélectionner dans notre corpus d'entraînement 20 lemmes qui sont peu fréquents dans un corpus de référence standard, mais qui présentent une spécificité élevée dans notre corpus. Nous étudions également la transcription de la disfluence "euh", de l'interjection "hein" et de la répétition d'amorces, qui sont fréquentes dans tout discours oral mais non transcrits par Whisper. Nous avons exclu de notre analyse les emprunts à l'anglais (p.ex. «drop») afin d'éviter que le large pré-entraînement dont Whisper a bénéficié sur l'anglais, n'influe sur nos analyses : si ces lemmes sont rares en français et spécifiques à notre corpus, ils sont possiblement plus fréquents en anglais. Nous avons également exclu les lemmes spécifiques à certains matchs, tels que la nationalité des joueurs (p.ex. roumaine, argentine), ainsi que des termes plus généraux (p.ex. «introduction»). Bien que ces termes puissent être analysés dans une étude future afin de graduer les performances de FM sur différents types d'éléments lexicaux.

Le premier tableau (Tableau 2) de l'annexe 6.2 résume les lemmes utilisés, leur fréquence dans notre corpus, leur fréquence dans le corpus de référence, et leur spécificité (LogLike). Nous avons ensuite vérifié que tous ces lemmes apparaissent dans la transcription du match utilisée pour le test, et les lemmes ainsi que leur fréquence et leur spécificité sont présentés dans le deuxième tableau (Tableau 3) de l'annexe 6.2. Enfin, nous avons examiné, pour chaque taille de modèles (PM et FM), la différence d'occurrence de ces lemmes avec la transcription de référence.

4 Résultats

4.1 Quantitatif

Les valeurs de WER sont reportées pour chacun des modèles Whisper par ordre croissant de nombre de paramètres en Figure 1. Pour tous les modèles, quel que soit leur taille, le *fine-tuning* avec la méthode LoRA a permis de diminuer significativement la métrique WER sur l'ensemble de test.

On remarque que plus le modèle est large, plus l'écart de performance entre le modèle *Pre-trained* et sa version *Fine-tuned* est important. Les modèles les plus larges sont donc ceux qui ont le plus bénéficié du fine-tuning. Par exemple le modèle Tiny a vu sa WER diminuer de 14% tandis que cette diminution est de 54% pour whisper Large-v2. Pour évaluer la qualité de la transcription d'un ASR, il est également pertinent de la comparer avec les performances humaines sur cette même tâche. La transcription de notre corpus reposant en partie sur du *student sourcing*, il nous est possible de calculer la WER entre les transcriptions des étudiants et ces mêmes transcriptions après correction par un expert (étudiants vs gold sur la Figure 1). Il est notable que les modèles medium et large pré-entraînés obtiennent des performances similaires à celles des étudiants, confirmant une fois de plus que Whisper produit des transcriptions décentes en *zero-shot learning*. L'écart se creuse après *fine-tuning*, Whisper Small dépassant déjà largement les étudiants en terme de WER. D'un point de vue quantitatif, il est certain que le *fine-tuning* est un excellent choix pour qui dispose de la puissance

9. http://phraseotext.univ-grenoble-alpes.fr/anaText/index.php

de calcul nécessaire. Néanmoins on ne peut déduire de ce simple chiffre ce qui a été amélioré par le *fine-tuning* dans la transcription.

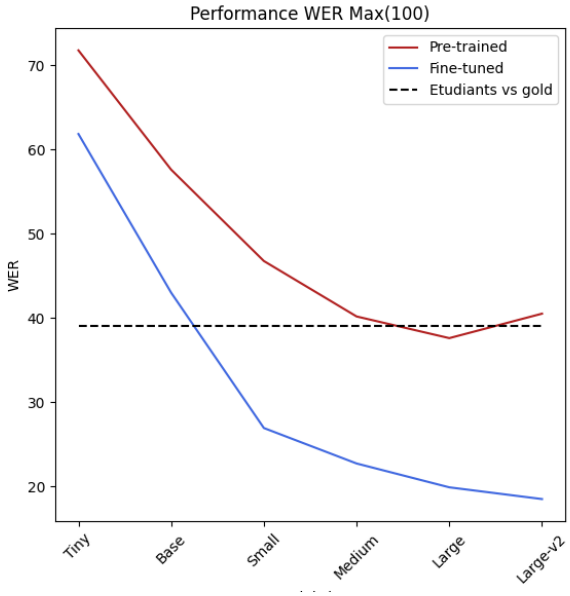


FIGURE 1 – Performance des modèles Whisper *Pre-trained* et *Fine-tuned* par ordre croissant de taille

4.2 Qualitatif

Les résultats de l'analyse qualitative sont présentés en Figure 2. Pour chaque PM et FM nous avons calculé l'erreur absolue moyenne (MAE [10]) du nombre total d'occurrences entre la transcription prédite et le gold pour 20 lemmes de notre ensemble de test appartenant au vocabulaire spécifique du sport. Les résultats montrent deux choses : (i) plus le modèle est large, plus la MAE est faible [11], (ii) les FM ont une MAE systématiquement plus faible que les PM, preuve que le modèle, avec seulement 6h30 d'audio et de transcriptions est capable d'apprendre du lexique spécifique.

Le détail des lemmes pour chaque modèle est exposé en annexe 6.3 (Tableau 4). Certains lemmes, tels que «pénalité», «essai» et «touche», sont déjà bien reconnus, même par les modèles de petites tailles PM et FM, suggérant leur présence fréquente dans l'ensemble d'apprentissage du pré-entraînement. Le Tiny-FM présente dans sa transcription une hallucination faisant apparaître ces lemmes plus de fois que dans la transcription de référence, comme observé avec «essai» qui apparaît 96 fois dans les transcriptions du Tiny-FM contre seulement 67 fois dans le gold.

D'autres lemmes, tels que «mêlée», voient leur reconnaissance s'améliorer avec l'augmentation de la taille du modèle et bénéficient d'une meilleure détection suite au *fine-tuning*. Les PM éprouvent des difficultés à détecter certains lemmes spécifiques au rugby, tels que «chandelle» et «rebond», détectés moitié de fois moins que par les FM en faisant la somme du total de détection pour ces lemmes pour toutes les tailles de PM comparés à la somme de leur détection par toutes les tailles de FM. Les deux lemmes comprenant un tiret («en-avant», «en-but») ne sont pas reconnus par les PM, exception faite du modèle Base-PM qui identifie presque toutes les occurrences de «en-but». Les tentatives de

10. Mean Absolute Error
11. Excepté pour le modèle large, faisant plus d'erreur que le modèle medium

recherche de ces lemmes sans tiret ou avec seulement une partie du lemme (avant, but) n'ont pas donné de résultats concluants non plus, les modèles ne parvenant pas à les détecter.

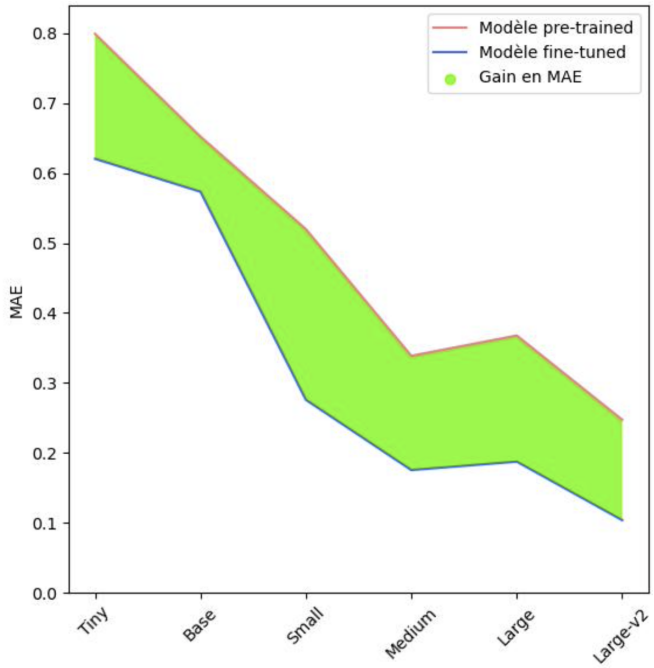


FIGURE 2 – MAE du nombre d'occurrences des PM et FM par rapport à la référence

Certaines disfluences et interjections apparaissent dans notre corpus. Elles présentent l'intérêt de ne pas être spécifiques à un corpus de spécialité, car présentes dans tout type de discours oral et plutôt absentes de notre type de corpus, contrairement au lexique, tout en étant absentes des transcriptions des modèles PM. On constate une détection de 0% sur toute la famille de modèles PM. Après *fine-tuning*, Whisper reconnaît entre 38% (base) à 92% (large-v2) la disfluence "euh" et l'interjection "hein", les plus courantes. D'autres types de disfluences, comme les amorces, ne sont jamais identifiés par Whisper. Cette lacune peut être attribuée à notre convention de transcription (ICOR, 2013), mais aussi au fait que Whisper n'a jamais été exposé à ces disfluences lors de son entraînement. En effet, Whisper est principalement basé sur du sous-titrage de films/séries, où la spontanéité du langage peut être discutée, et où l'annotation des disfluences n'est ni nécessaire ni souhaitée.

5 Conclusion et Perspectives

La présente étude vise à évaluer l'impact du *fine-tuning* des différentes tailles de modèles de Whisper sur la transcription de corpus spécialisés, en particulier dans le contexte des commentaires sportifs. Les résultats obtenus, tant qualitatifs que quantitatifs, indiquent une corrélation positive entre la performance du modèle et sa taille, soulignant ainsi l'efficacité accrue des modèles plus larges. Par ailleurs, plus les modèles sont larges plus l'écart entre leur performance et celle de leur équivalent PM est important. Cependant, en raison de la taille limitée de notre corpus et du nombre de modèles, il est difficile de tirer des conclusions définitives, et des recherches supplémentaires avec plus de données sont nécessaires pour confirmer ou infirmer cette corrélation taille du modèle /gain après *fine-tuning*. De plus, il serait intéressant de mettre en relation la taille des modèles avec la quantité de données annotées nécessaire pour atteindre un certain niveau de performance. Jain *et al.* (2023b) ont

amorcé une telle démarche mais de façon limitée.

Ensuite, alors même qu'une analyse lexicale a été effectuée dans cette étude, la spécificité de notre corpus est multidimensionnelle, justifiant ainsi la réalisation d'autres analyses pour évaluer l'impact, par exemple, des accents des commentateurs, de la syntaxe particulière des commentaires sportifs ou du bruit de fond sur la performance des ASR. Dans ce contexte, il convient d'explorer dans quelle mesure le *fine-tuning* de modèles tels que Whisper peut renforcer leur robustesse face à ces défis en les comparant aux PM.

En outre, pour mener des analyses grammaticales, l'utilisation d'analyseurs syntaxiques est courante, mais ces outils nécessitent de la ponctuation, que Whisper génère. Des études sont donc nécessaires pour évaluer la capacité de Whisper à insérer une ponctuation cohérente aux endroits appropriés.

Enfin, des recherches supplémentaires pourraient être entreprises pour mieux appréhender les types de périodes discursives les plus difficiles à traiter pour les PM et pour évaluer le potentiel des FM à surmonter ces difficultés. Dans le contexte des commentaires sportifs, qui se divisent généralement en deux types, à savoir les *colour commentaries* (Hartmann, 2013) et les narrations d'action en temps réel (*play-by-play*), une analyse séquentielle pourrait permettre de déterminer si les modèles performent mieux sur l'un ou l'autre type, et si le *fine-tuning* améliore leur performance dans ces deux catégories.

Cette étude s'est principalement concentrée sur l'amélioration de la transcription à partir d'un large modèle de langage affiné, visant à réduire le besoin de corrections humaines par rapport à une transcription manuelle ou avec un PM. Toutefois, pour optimiser la création de corpus spécialisés, notamment dans le cadre des études en linguistique de l'oral, et particulièrement dans notre contexte où des analyses prosodiques sont réalisées, il est essentiel de garantir la précision de l'alignement des unités inter-pausales. Malgré les capacités de Whisper dans cette tâche, son niveau de précision reste comparable à celui d'un générateur de sous-titres, ce qui n'est pas suffisant pour nos besoins de recherche. De plus, Whisper ne prend pas en charge la diarisation, c'est-à-dire la distinction entre les différents locuteurs.

Une précédente étude (Stasica *et al.*, 2023) a souligné la similarité en termes de temps nécessaire entre la transcription, l'alignement et la diarisation entièrement manuelles des commentaires sportifs, et le temps requis pour corriger la transcription de Whisper *pre-trained*, réaligner et diariser. Les résultats ont montré que la majeure partie du temps n'est pas consacrée à la correction de la transcription, mais à l'alignement et à la diarisation.

Actuellement, nos recherches se concentrent sur l'automatisation de l'alignement et de la diarisation afin d'optimiser la création de nos corpus. Whisper dispose déjà de *timestamps* sous forme de *token* spéciaux ajoutés durant la génération de texte. Néanmoins ceux-ci souffrent d'une précision instable, due à la nature de l'ensemble d'entraînement faiblement supervisé. Pour remédier à ce problème, des solutions ont été proposées offrant une précision à plus ou moins 100ms, suffisant pour notre usage, notamment le *Dynamic Time Warping* (Giorgino, 2009).

Aucune solution directe n'existe pour la diarisation avec Whisper. La solution la plus couramment utilisée est d'ajouter un second modèle indépendant de Whisper pour effectuer la diarisation puis fusionner diarisation / transcription / alignement, comme proposé par la librairie `pyannote` (Bredin *et al.*, 2020).

Au vue de la difficulté que pose ce triple problème, nous travaillons en parallèle à la création d'un modèle généraliste qui rendrait l'implémentation plus simple tout en limitant le risque de propagation d'erreurs entre les différentes étapes.

Références

AUGENDRE S., KUPŚĆ A., BOYÉ G. & MATHON C. (2018). Live TV sports commentaries : specific syntactic structures and general constraints. In D. LEGALLOIS, T. CHARNOIS & M. LARJAVAARA, Éds., *The Grammar of Genres and Styles : From Discrete to Non-Discrete Units*, p. 194–218. De Gruyter Mouton.

BARRAS C., GEOFFROIS E., WU Z. & LIBERMAN M. (2001). Transcriber : development and use of a tool for assisting speech corpora production. *Speech Communication*, **33**(1-2), 5–22.

BAUDE O. & DUGUA C. (2016). Les ESLO, du portrait sonore au paysage digital. *Corpus*, **15**.

BREDIN H., YIN R., CORIA J. M., GELLY G., KORSHUNOV P., LAVECHIN M., FUSTES D., TITEUX H., BOUAZIZ W. & GILL M.-P. (2020). Pyannote. audio : neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 7124–7128 : IEEE.

CANDEA M. (2000). Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané. *Etude sur un corpus de récits en classe de français*.

DAVIS K. H., BIDDULPH R. & BALASHEK S. (1952). Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, **24**(6), 637–642.

FONTAGNOL C., HANOTE S. & MATHON C. (2023). Sélection lexicale et réalisations prosodiques : impact des contraintes d'un genre discursif spécifique, le commentaire sportif télévisuel en direct. *Lexique et frontières de genres*, p. 97–116.

GHAI W. & SINGH N. (2012). Literature review on automatic speech recognition. *International Journal of Computer Applications*, **41**(8).

GIORGINO T. (2009). Computing and visualizing dynamic time warping alignments in R : the dtw package. *Journal of statistical Software*, **31**, 1–24.

HARTMANN C. (2013). *Pre-fabricated speech formulas as long-term memory solutions to working memory overload in routine language*. Thèse de doctorat, University of Zurich.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2021). Lora : Low-rank adaptation of large language models. *arXiv preprint arXiv :2106.09685*.

ICOR (2013). Convention ICOR. *Lyon : université de Lyon. URL : http ://icar. cnrs. fr/projets/corinte/documents/2013_Conv_ICOR_250313. pdf*.

JAIN R., BARCOVSCHI A., YIWERE M., CORCORAN P. & CUCU H. (2023a). Adaptation of Whisper models to child speech recognition. *arXiv preprint arXiv :2307.13008*.

JAIN S., KIRK R., LUBANA E. S., DICK R. P., TANAKA H., GREFENSTETTE E., ROCKTÄSCHEL T. & KRUEGER D. S. (2023b). Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv :2311.12786*.

KARPAGAVALLI S. & CHANDRA E. (2016). A review on automatic speech recognition architecture and approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, **9**(4), 393–404.

LI Y., YU Y., LIANG C., HE P., KARAMPATZIAKIS N., CHEN W. & ZHAO T. (2023). Loftq : Lora-fine-tuning-aware quantization for large language models. *arXiv preprint arXiv :2310.08659*.

LORTAL G. & MATHON C. (2008). Motion and Emotion or how to align emotional cues with game actions. In *WORKSHOP PROGRAMME &*, p.̃79.

NGUYEN T. A., KHARITONOV E., COPET J., ADI Y., HSU W.-N., ELKAHKY A., TOMASELLO P., ALGAYRES R., SAGOT B., MOHAMED A. & DUPOUX E. (2022). Generative spoken dialogue language modeling.

OYUCU S. (2023). Comparing the Fine-Tuning and Performance of Whisper Pre-Trained Models for Turkish Speech Recognition Task. In *2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, p. 1–4 : IEEE.

PĂIS V., MITITELU V. B., ION R. & IRIMIA E. (2023). Evaluating a Fine-Tuned Whisper Model on Underrepresented Romanian Speech. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, p. 141–145 : IEEE.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, p. 28492–28518 : PMLR.

SCHNEIDER S., BAEVSKI A., COLLOBERT R. & AULI M. (2019). wav2vec : Unsupervised pre-training for speech recognition. *arXiv preprint arXiv :1904.05862*.

STASICA A., BOYÉ G., KUPŚĆ A. & MATHON C. (2023). Chuchoter à l'oreille des corpus : Whisper, pour augmenter la production de corpus oraux de spécialité. COSEDI.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

WANG D., WANG X. & LV S. (2019). An overview of end-to-end automatic speech recognition. *Symmetry*, **11**(8), 1018.

XIE P., LIU X., CHEN Z., CHEN K. & WANG Y. (2023). Whisper-MCE : Whisper Model Finetuned for Better Performance with Mixed Languages. *arXiv preprint arXiv :2310.17953*.

6 Annexes

6.1 Annexe A

TABLE 1 – Noms des matchs, des experts et des journalistes et année des match

Match	Année	Journaliste	Expert
France/Argentine	2007	Thierry Gilardi	Thierry Lacroix
France/Canada	2015	Christian Jeanpierre	Bernard Laporte
France/Roumanie	2015	Christian Jeanpierre	Bernard Laporte
France/Nouvelle-Zélande	2015	Christian Jeanpierre	Bernard Laporte
Afrique du Sud/Argentine	2015	Christian Jeanpierre	Bernard Laporte
Japon/Fidji	2007	Nicolas Delage	Jérôme Papin

6.2 Annexe B : détail lexique spécifique

TABLE 2 – Fréquence et degré de spécificité de vingt lemmes dans la transcription gold des matchs servant à l'entraînement de Whisper classé par Anatext

Lemme	Fréquence	CorpusRef (par million)	LogLike (spécificité)
en-avant	126	0.035	2684.069
pénalité	132	1.09	1916.386
essai	208	20.895	1601.322
touche	184	16.28	1467.165
mêlée	121	4.875	1185.473
plaquage	45	0.4	646.637
talonneur	29	0.035	532.528
relance	29	0.595	341.643
pénaliser	26	0.47	319.679
buteur	20	0.15	294.184
pilier	44	10.305	260.996
hors-jeu	24	1.47	211.049
envoi	32	6.1	203.258
sélectionneur	12	0.035	199.177
rebond	18	0.91	166.349
plaqueur	10	0.035	162.334
en-but	11	0.12	153.557
chandelle	27	8.725	142.668
renvoi	20	4.36	121.545
ailier	14	1.065	116.263

TABLE 3 – Fréquence et degré de spécificité de vingt lemmes dans la transcription gold du match servant au test du *fine-tuning* de Whisper classé par Anatext

Lemme	**Fréquence**	**CorpusRef (par million)**	**LogLike (spécificité)**
pénalité	41	1.09	575.298
essai	67	20.895	564.366
touche	61	16.28	533.237
mêlée	30	4.875	293.029
en-avant	13	0.035	258.228
ailier	18	1.065	215.516
plaquage	12	0.4	160.688
hors-jeu	11	1.47	111.966
buteur	7	0.15	103.093
talonneur	5	0.035	89.761
en-but	5	0.12	71.732
pénaliser	6	0.47	68.017
pilier	9	10.305	52.310
plaqueur	3	0.035	50.791
renvoi	7	4.36	49.157
relance	4	0.595	39.808
sélectionneur	2	0.035	31.103
rebond	3	0.91	25.429
envoi	4	6.1	20.997
chandelle	4	8.725	18.216

6.3 Annexe C : Détails des résultats de l'analyse qualitative

TABLE 4 – Nombre d'occurrences des lemmes spécifiques pour chaque taille de modèle Whisper *Fine-tuned* (FM) et Pre-trained (PM)

	Gold	Tiny		Base		Small		Medium		Large		Largev2	
		PM	FM	PM	FM	PM	FM	PM	FM	PM	FM	PM	FM
pénalité	41	26	***27***	33	***36***	39	***39***	40	***36***	39	***41***	39	***40***
essai	67	14	***96***	38	***33***	42	***58***	55	***58***	57	***62***	57	***63***
touche	61	30	***75***	41	***69***	49	***50***	61	***48***	62	***62***	63	***63***
mêlée	30	0	***5***	5	***13***	7	***21***	12	***19***	15	***27***	20	***23***
en-avant	13	0	***0***	0	***3***	0	***6***	0	***4***	0	***7***	0	***8***
ailier	18	0	***0***	0	***0***	0	***0***	4	***16***	0	***9***	7	***13***
plaquage	12	0	***0***	0	***2***	0	***8***	8	***9***	5	***10***	7	***9***
hors-jeu	11	0	***0***	0	***3***	2	***0***	3	***9***	5	***5***	6	***7***
buteur	7	5	***5***	5	***4***	6	***7***	7	***7***	7	***6***	6	***7***
talonneur	5	1	***2***	1	***1***	2	***5***	5	***5***	4	***4***	5	***4***
en-but	5	0	***0***	4	***0***	0	***0***	0	***0***	0	***0***	0	***3***
pénaliser	6	3	***4***	6	***4***	5	***5***	4	***5***	6	***5***	6	***6***
pilier	9	4	***0***	6	***6***	9	***9***	9	***9***	8	***9***	7	***9***
plaqueur	3	0	***0***	0	***1***	0	***2***	2	***3***	3	***2***	3	***2***
renvoi	7	0	***0***	0	***4***	5	***7***	5	***7***	7	***7***	6	***7***
relance	4	3	***4***	5	***4***	4	***3***	4	***4***	4	***4***	4	***4***
sélectionneu	2	0	***2***	0	***0***	1	***2***	2	***2***	0	***2***	2	***2***
rebond	3	0	***0***	0	***0***	2	***3***	0	***3***	2	***3***	3	***3***
envoi	4	0	***0***	0	***3***	3	***4***	3	***3***	4	***4***	3	***5***
chandelle	4	0	***0***	0	***0***	0	***3***	4	***4***	0	***4***	3	***4***

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Optimiser le choix des exemples pour la traduction automatique augmentée par des mémoires de traduction

Maxime Bouthors[1, 2] Josep Crego[1] François Yvon[2]
(1) Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
(2) ChapsVision, 92150, Suresnes
bouthors@isir.upmc.fr, jcrego@chapsvision.com, yvon@isir.upmc.fr

RÉSUMÉ

La traduction neuronale à partir d'exemples s'appuie sur l'exploitation d'une mémoire de traduction contenant des exemples similaires aux phrases à traduire. Ces exemples sont utilisés pour conditionner les prédictions d'un décodeur neuronal. Nous nous intéressons à l'amélioration du système qui effectue l'étape de recherche des phrases similaires, l'architecture du décodeur neuronal étant fixée et reposant ici sur un modèle explicite d'édition, le Transformeur « multi-Levenshtein ». Le problème considéré consiste à trouver un ensemble optimal d'exemples similaires, c'est-à-dire qui couvre maximalement la phrase source. En nous appuyant sur la théorie des fonctions sous-modulaires, nous explorons de nouveaux algorithmes pour optimiser cette couverture et évaluons les améliorations de performances auxquelles ils mènent pour la tâche de traduction automatique.

ABSTRACT

Optimizing example selection for retrieval-augmented machine translation with translation memories

Retrieval-augmented machine translation leverages examples from a translation memory by retrieving similar instances. These examples are used to condition the predictions of a neural decoder. We aim to improve the upstream retrieval step and consider a fixed downstream edit-based model : the multi-Levenshtein Transformer. The task consists of finding a set of examples that maximizes the overall coverage of the source sentence. To this end, we rely on the theory of submodular functions and explore new algorithms to optimize this coverage. We evaluate the resulting performance gains for the machine translation task.

MOTS-CLÉS : Traduction Automatique, Recherche d'Information, Mémoires de Traduction, Fonctions Sous-Modulaires, Traduction à partir d'Exemples.

KEYWORDS: Machine Translation, Information Retrieval, Translation Memories, Submodularity, Example-based Translation.

1 Introduction

De nombreux travaux récents s'intéressent à la génération augmentée par des exemples (Li et al., 2022). En traduction, l'utilisation d'exemples remonte aux méthodes de traduction assistée par ordinateur par des traducteurs professionnels (Bowker, 2002) : éditer des segments très similaires à la phrase de référence permet d'accélérer la traduction. Cette idée est au fondement des méthodes

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 582–604.

basées sur des exemples (Nagao, 1984; Somers, 1999; Carl et al., 2004). Elle est adaptée aux systèmes statistiques par (Koehn & Senellart, 2010) et plus récemment aux méthodes neuronales.

Il existe en fait de nombreuses manières d'exploiter les exemples : la traduction conditionnelle qui introduit un système d'attention sur les exemples (Gu et al., 2018; Bulte & Tezcan, 2019; Hoang et al., 2022) ; l'affinage « léger » sur un ensemble d'exemples pour faire de la micro-adaptation (Farajian et al., 2017) ; les méthodes intégrant des exemples dans le contexte de grands modèles de langue (LLM) génératifs multilingues (Moslem et al. (2023), *inter alia*) ; l'édition directe du meilleur exemple similaire (Gu et al., 2019). Nous discutons ces études dans la section 2.

Ici, nous nous intéressons au transformeur « multi-Levenshtein » (Bouthors et al., 2023), un modèle d'édition qui combine $k(\geq 1)$ exemples pour calculer une traduction. Cette caractéristique le rend sensible à la qualité des exemples récupérés. En particulier, supposant fixé k le nombre de phrases à récupérer, nous cherchons à répondre à la question : comment identifier un ensemble optimal de k exemples ? Trouver exactement ce meilleur ensemble conditionnellement au modèle et à la phrase source est difficile, ce qui implique de considérer des heuristiques. Pour les construire, nous faisons l'hypothèse qu'un ensemble de phrases parallèles couvrant (en source) la phrase à traduire fournit des exemples couvrant (en cible) la traduction à produire. En dépit de ses limites, liées à des phénomènes linguistiques bien connus (variation lexicale, divergences morphologiques ou syntaxiques entre langues source et cible, etc.), cette hypothèse est adoptée dans les travaux de l'état-de-l'art.

Notre contribution principale est alors l'étude de plusieurs manières de définir la notion de couverture et de rechercher des k meilleurs exemples dans une mémoire de traduction. En tirant parti de la théorie des fonctions sous-modulaires, dont une sous-classe correspond à une notion très générique de couverture, nous analysons dans un cadre unifié les avantages comparés de ces différentes propositions, et évaluons leur impact dans une tâche de traduction multidomaines.

2 Travaux Connexes

De nombreux efforts pour intégrer des exemples dans la génération de textes ont été menés ces dernières années (Li et al., 2022). Au-delà des améliorations de performances, la possibilité de présenter aux utilisateurs améliore la transparence des décisions qui sont prises (Rudin, 2019). En traduction automatique, les exemples récupérés d'une mémoire de traduction sont fournis au modèle comme un contexte supplémentaire, par exemple en concaténant le côté cible des exemples au texte source (Bulte & Tezcan, 2019), ou en tirant parti à la fois de la source et de la cible (Pham et al., 2020). Gu et al. (2018); Xia et al. (2019); He et al. (2021b) considèrent des stratégies plus sophistiquées pour enrichir le contexte source.

Si beaucoup de travaux se limitent à considérer les $k \geq 1$ exemples les plus similaires, Cheng et al. (2022); Agrawal et al. (2023); Sia & Duh (2023) cherchent à trouver un ensemble d'exemples complémentaires entre eux. Le premier travail utilise l'algorithme de *Maximum Marginal Relevance* (MMR) (Goldstein & Carbonell, 1998), tandis que les deux autres proposent une forme de maximisation de couverture. Gupta et al. (2023) donne une formulation générale du problème de couverture, l'appliquant à l'apprentissage en contexte (*in context learning*) sur des tâches diverses.

Une autre extension de cette approche exploite des corpus monolingues, en recherchant les exemples directement dans la langue cible. Cai et al. (2021) proposent un modèle de recherche et de traduction unique entraîné de bout-en-bout dont la procédure de recherche translingue est optimisée pour

retourner des exemples utiles pour la Traduction Automatique (TA).

La plupart de ces travaux reposent sur des modèles de génération auto-régressifs (AR), impliquant que les exemples intégrés au contexte n'ont qu'un effet indirect sur la sortie. L'utilisation d'une mémoire de traduction avec un décodeur non auto-régressif (NAR) est étudiée par Niwa et al. (2022); Xu et al. (2023); Zheng et al. (2023) qui adaptent le transformeur de Levenshtein (Gu et al., 2019) pour éditer directement l'exemple le plus similaire en une traduction de la phrase source. Bouthors et al. (2023) étendent cette technique à plusieurs exemples.

Une autre approche consiste à utiliser des similarités au niveau des contextes de génération, c'est-à-dire d'états cachés du décodage des mémoires de traduction, plutôt qu'au niveau des phrases (Zhang et al., 2018). He et al. (2021a); Khandelwal et al. (2021), entre autres, utilisent des méthodes de plus proches voisins sur des contextes. La prédiction du token suivant est alors guidée par ceux trouvés dans des contextes proches. Diverses extensions sont apportées par Zheng et al. (2021); Meng et al. (2022); Martins et al. (2022).

Il est enfin difficile d'ignorer l'essor des grands modèles de langue multilingues (LLM) qui, amorcés par un contexte contenant une description de la tâche à accomplir et des exemples, peuvent générer des traductions de qualité. Cette approche a été testée sur de nombreux LLM dans le but de mettre en évidence leur capacité à traiter de multiples tâches. Pour ce qui concerne la TA, plusieurs travaux étudient l'impact du contexte d'entrée, en cherchant à optimiser le nombre d'exemples et leur sélection (Vilar et al., 2023; Zhang et al., 2023; Hendy et al., 2023; Bawden & Yvon, 2023). Voir également sur ces questions (Moslem et al., 2023; Mu et al., 2023; Agrawal et al., 2023; Sia & Duh, 2023; M et al., 2023).

3 Le modèle « multi-Levenshtein »

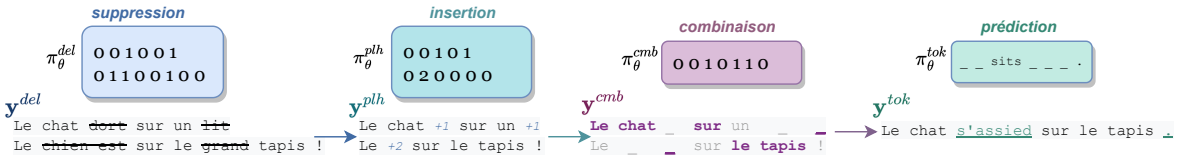


FIGURE 1 – Première étape de décodage de $\texttt{TM}^k\texttt{-LevT}$. Les deux exemples qui sont édités sont y_1 : « le chat dort sur un lit » et y_2 :« le chien est sur le grand tapis ». Les insertions prédites à l'étape 2 (insertion) sont représentées par des entiers, puis matérialisées par des '_' à l'étape 3 (combinaison).

Nous nous intéressons au modèle du transformeur multi-Levenshtein, $\texttt{TM}^k\texttt{-LevT}$, (Bouthors et al., 2023), qui repose sur un modèle explicite d'édition d'exemples. L'algorithme de décodage est représenté sur la Figure 1. Il s'agit d'un modèle d'édition type transformeur (Vaswani et al., 2017) qui prend en entrée k couples de phrases exemples cibles $\{y_1, \ldots, y_k\}$, et les édite conditionnellement à la source x en quatre étapes :

1. **Suppression** : pour chaque exemple cible y_i, délétion de tokens ;
2. **Insertion** : pour chaque exemple cible y_i, insertion de tokens vides PLH entre chaque token ;
3. **Combinaison** : combinaison des k exemples cibles en s'appuyant sur un multi-alignement qui permet de déterminer, pour chaque position, l'origine (parmi $y_1 \ldots y_k$) du token à conserver ;
4. **Prédiction** : pour chaque position comprenant un PLH, prédiction du mot à insérer.

Le calcul des associations entrées/sorties est réalisé par une architecture encodeur-décodeur non-autorégressive, dans lequel on remplace la couche de prédiction de mots habituelle par quatre couches linéaires (une pour chaque opération) qui projettent les représentations latentes sur l'ensemble des opérations possibles. Par exemple, pour l'insertion (étape 2), pour une source x et des exemples $y_1, \cdots, y_k$:

$$\texttt{insertion}^* = \arg\max \texttt{Insertion}(\texttt{Décodeur}(\texttt{Encodeur}(x), y_1, \cdots, y_k)) \quad (1)$$

Le modèle est entraîné par apprentissage par imitation (Daumé et al., 2009; Ross et al., 2011), en utilisant comme politique experte [1] la série d'opérations qui maximise la copie des tokens qui sont à la fois présents dans les exemples d'entrée et dans la référence. Autrement dit, cette politique optimale s'appuie sur une notion de couverture optimale. Pour déterminer cette politique, un algorithme d'alignement calcule la manière optimale de faire correspondre les exemples et la référence.

On se reportera à (Bouthors et al., 2023) pour une présentation détaillée de cette architecture de base et de diverses extensions (réalignement, pré-apprentissage) qui lui permettent de tirer efficacement parti de plusieurs exemples similaires. L'essentiel étant de noter que dans son principe même, cette architecture est particulièrement sensible aux exemples qui sont fournis en entrée du système.

4 Recherche d'Information dans une Mémoire de Traduction

4.1 Recherche de Phrases Similaires (RPS)

Supposons que l'on ait accès à une mémoire de traduction $\mathcal{D} = \{(x_1, y_1), \cdots (x_N, y_N)\}$ et soit x une phrase source que l'on cherche à traduire. Le cadre classique consiste à évaluer indépendamment chaque candidat selon un score de similarité $s(x_i, x)$ et récupérer les k exemples les plus similaires. s peut être un score lexical (similarité de Jaccard, TF-IDF, BM25, distance d'édition, rappel n-gramme, etc.) ou sémantique (similarité cosinus entre deux plongements). Les premiers (Jaccard, BM25) reposent sur des algorithmes simples et servent souvent à filtrer un premier ensemble T de candidats, que l'on peut ensuite évaluer avec des fonctions de comparaison plus sophistiquées.

Cependant, récupérer les k meilleurs candidats peut s'avérer sous-optimal, par exemple lorsque tous ces exemples sont très similaires entre eux. Pour évaluer globalement l'ensemble des candidats retournés par la RPS, des stratégies visant à introduire de la diversité (par exemple via l'algorithme de *Maximum Marginal Relevance*, MMR) ou des contraintes de couverture peuvent alors être déployées. Nous considérons la deuxième stratégie, la première étant documentée dans (Cheng et al., 2022).

4.2 Fonctions sous-modulaires et couverture

Par analogie aux formalisations développées dans un cadre de recherche d'information pour trouver des résultats variés (Lin & Bilmes, 2011; Krause & Golovin, 2014), nous nous appuyons sur la théorie des fonctions sous-modulaires pour formaliser la RPS basée sur la maximisation de la couverture

1. En apprentissage par imitation, la politique « experte » est celle que l'apprenti cherche à reproduire (Knyazeva et al., 2018) : elle indique ici les actions optimales qu'il faut effectuer pour éditer les exemples et générer la traduction de référence. Comme elle n'est pas observée dans les données d'apprentissage, il faut la calculer en s'appuyant sur diverses heuristiques.

de la phrase source. Nous commençons par rappeler quelques définitions, avant de présenter les algorithmes de sélection d'exemples.

4.2.1 Définitions

Définition 1 (Sous-modularité). *Soient Ω un ensemble et $f : 2^\Omega \to \mathbb{R}$, f est **sous-modulaire** si $\forall X, Y$ tels que $X \subset Y \subset \Omega, \forall z \in \Omega \setminus Y$:*

$$f(X \cup \{z\}) - f(X) \geq f(Y \cup \{z\}) - f(Y)$$

Intuitivement, cette définition exprime que le rendement marginal de f est décroissant : plus l'ensemble en entrée de f est grand, plus les incréments de f induits par l'ajout de nouveaux éléments sont faibles. Une classe de fonctions sous-modulaires bien documentée est la classe des fonctions de couverture pondérée (Krause & Golovin, 2014).

Définition 2 (Couverture pondérée). *Soit $N \in \mathbb{N}$ et $v^{(n)}(z)_{n\in[1,N]}$ une séquence de poids réels associée à $z \in \Omega$, correspondant à des aspects (de x) : $v^{(n)}(z)$ évalue à quel point z couvre l'aspects n. Une **fonction de couverture pondérée** associe à un sous-ensemble Z de Ω :*

$$f(Z) = \sum_{n=1}^{N} \max_{z\in Z} v^{(n)}(z)$$

Dans notre application, Ω est un ensemble d'exemples (source et références jointes) et N dénombre des aspects importants de la source x qu'il faut couvrir. Cet ensemble d'aspects peut être le sac-de-mots associé à $x = (x_1, \cdots, x_N)$, les indices de la séquence, l'ensemble des n-grammes ou des sous-arbres syntaxiques de taille bornée, etc. L'objectif est ensuite de trouver un ensemble Z de taille k qui maximise $f(Z)$, c.-à-d. qui garantit une couverture maximale pour chaque aspect.

Dans cette définition, le choix d'un opérateur $\max$ s'appliquant uniformément à tous les aspects est problématique et peut conduire à récupérer des phrases dans Z qui n'ont que peu de pertinence pour la TA, voir annexe B. En traduction, il est en effet plus important de couvrir certains aspects que d'autres (par exemple des lexèmes rares dans une représentation sac-de-mots). Pour pallier ce problème, nous introduisons une nouvelle fonction sous-modulaire.

Définition 3 (Couverture pondérée lissée). *Soit $N \in \mathbb{N}$, $Z = \{z_1, \cdots, z_{|Z|}\}$ et $(v_i^{(n)})_{n\in[1,N]}$ une séquence de poids réels pour $z_i \in Z$. Une **fonction de couverture pondérée lissée** de paramètre $\lambda \in [0, 1]$ est définie par :*

$$f(Z) = \sum_{n=1}^{N} \sum_{j=1}^{|Z|} \lambda^{j-1} v^{(n)}_{g^{(n)}(j)}, \tag{2}$$

avec $g^{(n)}$ une permutation telle que $i < j \Rightarrow v^{(n)}_{g^{(n)}(i)} \geq v^{(n)}_{g^{(n)}(j)}$, ordonnant les z_i selon les $v_i^{(n)}$.

La preuve de sa sous-modularité est en Annexe C.1. Avec cette nouvelle définition, un aspects n déjà couvert continue de contribuer au calcul de f, mais avec un coefficient qui diminue exponentiellement. Pour $\lambda = 0$, f correspond à la définition (2) et, pour $\lambda = 1$, la fonction devient modulaire et sa maximisation revient à choisir indépendamment les k exemples les plus couvrants.

4.2.2 Maximisation de la couverture

Maximiser f pour $\lambda < 1$ est NP-complet (Krause & Golovin, 2014). On ne connaît pas d'algorithme exact meilleur que l'énumération exhaustive. L'algorithme glouton 1 est la façon standard de maximiser une fonction sous-modulaire. Cependant, la combinaison linéaire figurant dans la définition de f à l'équation (2) devant être recalculée pour chaque candidat restant z, cet algorithme a une complexité temporelle $O(k^2N|T| + N|T|\log|T|)$ ainsi qu'un coût spatial lié aux permutations de tri g.

Pour améliorer la complexité, nous considérons l'algorithme 2 d'Agrawal et al. (2023), replacé ici dans le cadre de la théorie des fonctions sous-modulaires afin de garantir une borne inférieure.

Algorithme 1: Maximisation gloutonne d'une fonction sous-modulaire monotone

Données: source x, fonction sous-modulaire (de couverture de pondérée) f, ensemble de candidats T, nombre d'exemples souhaité k.
Résultat: Z
$Z \leftarrow \emptyset$;
tant que $|Z| < k$ **faire**
  $z^* \leftarrow \arg\max_{z \in T \setminus Z} f(Z \cup \{z\})$;
  $Z \leftarrow Z \cup \{z^*\}$;
fin
retourne Z ;

Algorithme 2: Maximisation gloutonne d'une fonction de couverture pondérée lissée

Données: source x, fonction sous-modulaire de couverture de x pondérée lissée f de facteur de sous-pondération λ, poids de couverture $v^{(n)}(z)$ pour $z \in T$ ensemble de candidats, nombre d'exemples souhaité k.
Résultat: Z
$Z \leftarrow \emptyset$;
$W \leftarrow (1, \ldots, 1)$; `/* |W| = N */`
tant que $|Z| < k$ **faire**
  $z^* \leftarrow \arg\max_{z \in T \setminus Z} W^T v(z)$; `/* séléctionner z* */`
  $Z \leftarrow Z \cup \{z^*\}$;
  $I^* \leftarrow \{n : v^{(n)}(z^*) > 0\}$; `/* aspects couverts par z* */`
  pour $n \in I^*$ **faire**
    $W_n \leftarrow \lambda W_n$; `/* sous-pondération des éléments couverts */`
  fin
  si $\lambda = 0$ ***et*** $W = (0, \ldots, 0)$ **alors**
    $W \leftarrow (1, \ldots, 1)$; `/* réinitialiser W */`
  fin
fin
retourne Z ;

L'algorithme 2 a une complexité temporelle $O(kN|T|)$. Il se rapproche de l'algorithme 1, à la différence près que les $v_i^{(n)}$ ne sont pas triés. Nous montrons en annexe C.2 que l'écart entre la solution retournée par l'algorithme 2 et la solution optimale est borné.

5 Cadre Expérimental

5.1 Données

Nous considérons des données anglais-français sur un panel de 6 corpus de domaine varié : ECB, EMEA, Europarl, JRC-Acquis, Ubuntu, Wikipedia [2]. Cela correspond à un sous-ensemble des données utilisées par Xu et al. (2023) dont nous reprenons la même partition entraînement/test. Une caractéristique des données de test est qu'elles ont été partitionnées en deux ensembles de 1000 phrases pour chaque domaine. Le premier ensemble (*test-0.4*) contient des phrases sources pour lesquelles le plus proche voisin dans la TM (au sens de la similarité de Levenshtein [3]) est à une distance entre 0,4 et 0,6. Pour le second (*test-0.6*), le plus proche voisin a un score d'au moins 0,6. Cela permet de différencier les comportements entre les échantillons avec des *« bonnes »* correspondances, et ceux avec des correspondances *« médiocres »*. Le tableau 1 résume les principales statistiques de ces corpus.

domaine	ECB	EME	Epp	JRC	Ubu	Wiki	**tout**
taille	195k	373k	2,0M	503k	9k	803k	3,9M
longueur moyenne	29,2	16,7	26,6	28,8	5,2	19,6	21,0

TABLE 1 – Nombre d'échantillons et longueur moyenne des phrases d'entraînement.

5.2 Scores de Couverture

Dans un premier temps, nous utilisons notre propre implémentation de BM25 (Robertson & Jones, 1976) pour récupérer les $T = 100$ meilleurs candidats. Ensuite, nous considérons plusieurs fonctions de couverture avec des scores et pondérations différents.

Couverture sac-de-mot (SDM) : Les aspects sont les termes sac-de-mot t_n de la source x et le poids $v^{(n)}(z)$ correspond au minimum du nombre d'occurrences de t_n dans le candidat z et dans x. Cette notion correspond au rappel modifié [4] : les termes ne peuvent pas être couverts plus que leur nombre d'occurrences dans la source.

Couverture 4-gramme ou moins (NGM) : Les aspects sont les 1-4-grammes t_n de x, et $v^{(n)}(z)$ est le minimum du nombre d'occurrences dans x et dans z.

Couverture par distance de Levenshtein (DL) : Les aspects sont les indices de x, et $v^{(n)}(z)$ vaut 1 (0 sinon) si et seulement si x_n appartient à une sous-chaîne copiée en calculant la distance de Levenshtein entre x et z. Puisqu'il peut exister plusieurs alignements optimaux, on peut soit calculer l'ensemble des sous-chaînes optimales et marginaliser pour chaque indice, soit échantillonner une solution et l'utiliser pour construire les $v^{(n)}(z)$.

Les $v^{(n)}(z)$ sont normalisés de deux manières différentes :

Normalisation par cardinalité : Pour SDM et NGM $v^{(n)}(z)$ est divisé par N le nombre d'aspects. Pour DL, on choisit de normaliser par le maximum de la taille de x et z afin de retrouver la formule

2. Les données sont en libre accès sur le site opus https://opus.nlpl.eu
3. Définie pour deux chaînes x, y par $1 - \frac{d(x,y)}{\max|x|,|y|}$, avec d la distance de Levenshtein.
4. "Modified recall", par analogie à la précision modifiée du score BLEU.

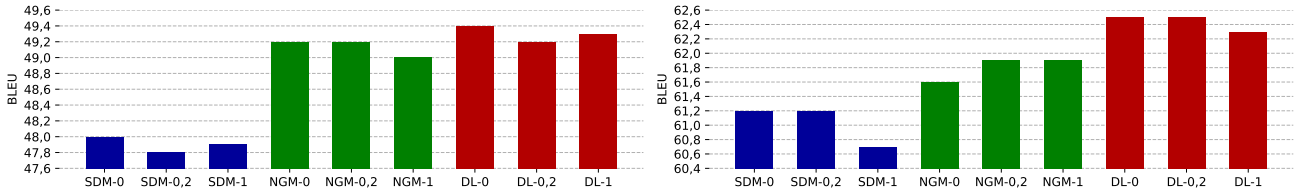


FIGURE 2 – Scores BLEU moyens selon le score de la similarité DL, pour $\lambda \in \{0; 0{,}2; 1\}$ sur *test-0.4* (gauche) et *test-0.6* (droite).

classique quand $\lambda = 1$.

Normalisation par rareté (IDF) : Chaque valeur $v^{(n)}(z)$ est normalisée avec des pondérations IDF, c.-à-d. que, pour SDM et DL, chaque mot w de x reçoit un poids $\mathrm{IDF}(w)/\sum_{w' \in x} \mathrm{IDF}(w)$. Pour NGM, l'indexation des n-grammes jusqu'à l'ordre 4 étant coûteuse, nous avons par simplification remplacé l'IDF de chaque n-gramme par la moyenne des IDF des termes qu'il contient.

Enfin, dans le cadre de l'introduction des fonctions de *couverture pondérée lissées*, nous étudions différentes valeurs de $\lambda \in \{0; 0{,}2; 0{,}5; 1\}$.

Recherche constrastive Nous comparons aussi nos résultats à MMR (Cheng et al., 2022) avec $\alpha = 0{,}3$ qui fait un compromis entre pertinence et diversité des exemples.

5.3 Métriques

Nous calculons les scores BLEU (Papineni et al., 2002) avec SacreBLEU (Post, 2018) [5]. Nous étudions également trois métriques calculées en comparant les phrases cibles récupérées à la référence : la couverture, la pertinence et la longueur des phrases. La couverture est calculée selon un rappel modifié, comme pour la fonction sous-modulaire unigramme avec $\lambda = 0$ sur les phrases tokénisées [6]. La pertinence est la précision sac-de-mot, c.-à-d. la proportion de termes utiles dans les exemples rapportée à la somme des longueurs des exemples. La longueur est calculée sur les phrases tokenisées.

6 Résultats

Rôle du score choisi : Nous effectuons une recherche pour les scores SDM, NGM et DL avec normalisation cardinale et $\lambda \in \{0; 0{,}2; 1\}$. Les histogrammes des scores BLEU de la figure 2 montrent : (1) La supériorité de la distance de Levenshtein (DL) sur les deux autres scores, même si NGM reste compétitif sur certains domaines (voir tableau 4 dans l'annexe D pour les résultats par domaine) ; (2) À l'exception de NGM-0 (*test-0.6*) maximiser la couverture avec $\lambda = 0$ est en moyenne préférable et conduit à un gain de +0,1 (resp. +0,2) pour *test-0.4* (resp. *test-0.6*) par rapport à $\lambda = 1$.

5. signature : `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0`;
6. Nous utilisons les scripts de Moses (https://github.com/moses-smt/)

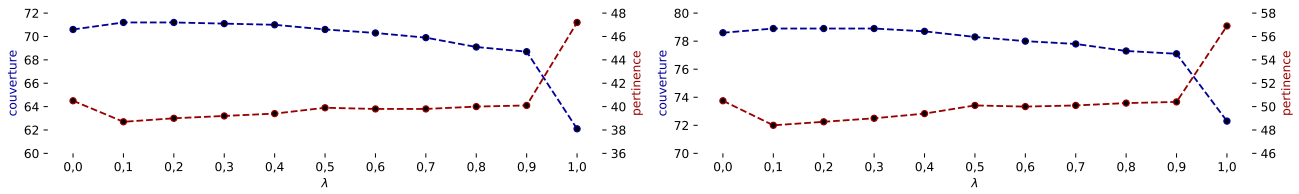


FIGURE 3 – Couverture et pertinence moyenne selon la valeur de λ pour DL sur *test-0.4* (gauche) et *test-0.6* (droite).

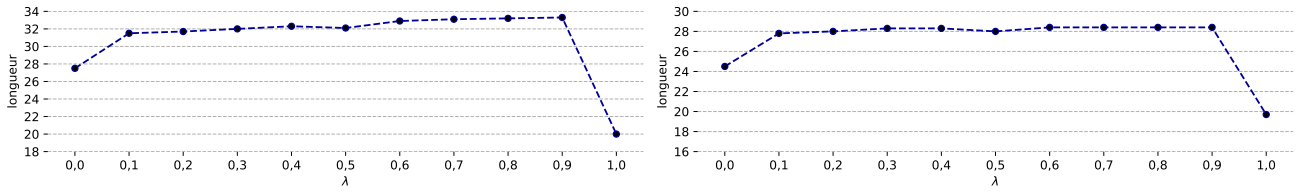


FIGURE 4 – Longueur moyenne selon la valeur de λ pour DL sur *test-0.4* (gauche) et *test-0.6* (droite).

Rôle de λ : Nous regardons l'impact de λ pour DL (Distance de Levenshtein) figures 3 et 4 en termes de couverture, pertinence et longueur. Nous trouvons un compromis entre pertinence et couverture. L'intervalle $[0{,}1; 0{,}9]$ est très stable. Nous observons une singularité à $\lambda = 0$. La couverture est légèrement plus faible que $\lambda = 0{,}1$. Le cas $\lambda = 1$ est singulier car ayant la couverture (resp. pertinence) la plus faible (resp. la plus élevée), ainsi que la plus faible longueur moyenne. Quant au score **BLEU** (voir figure 5), on observe un comportement différent entre les correspondances *médiocres* (*test-0.4*) et *bonnes* (*test-0.6*). $\lambda = 0$ semble en général mieux sur *test-0.4*, même si aucune tendance ne s'en dégage. Sur *test-0.6*, on observe une courbe en cloche avec une inflexion à $\lambda = 0$. La tendance semble indiquer que $\lambda = 0{,}5$ produit les meilleurs résultats, et cela même sur *test-0.4*. À noter qu'il y a des différences entre domaines (voir Annexe D).

Comparaison avec MMR Nous comparons les méthodes DL pour $\lambda \in \{0; 0{,}5; 1\}$ avec la méthode contrastive MMR. Sur la figure 6, nous observons de très légères différences entre les méthodes de recherche. MMR est légèrement mieux (+0,1 BLEU) sur *test-0.4*, mais DL-0,5 surpasse MMR de 0,1 sur *test-0.6*. Aucune des deux méthodes ne semble particulièrement supérieure.

Rôle de la normalisation : Par défaut, la normalisation se fait sur le nombre d'aspects à couvrir, sauf pour DL où il s'agit du maximum entre la longueur de la source et de l'exemple. Lorsqu'on

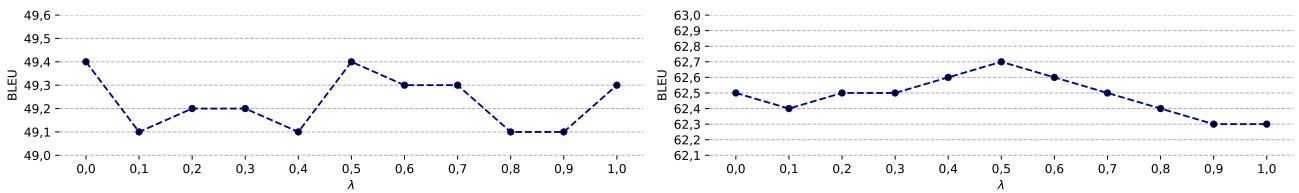


FIGURE 5 – Score BLEU moyen en fonction de λ pour DL sur *test-0.4* (gauche) et *test-0.6* (droite).

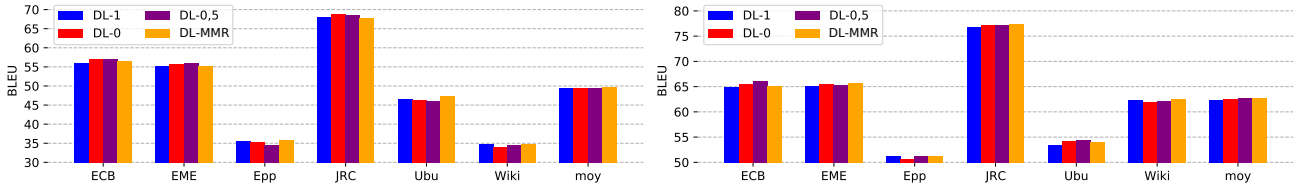


FIGURE 6 – Score BLEU avec $\lambda \in \{0; 0{,}5; 1\}$ et MMR sur *test-0.4* (gauche) et *test-0.6* (droite).

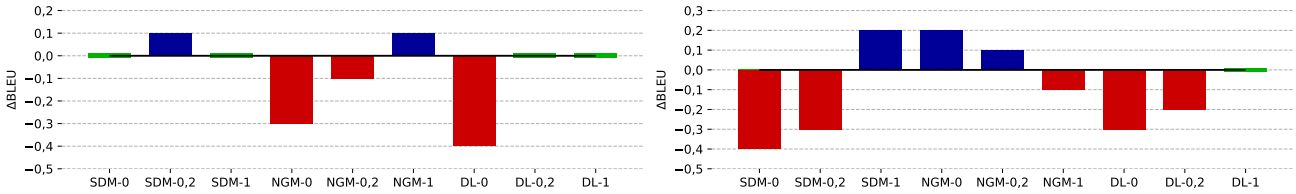


FIGURE 7 – Différence moyenne de BLEU entre normalisation IDF et cardinale, avec $\lambda \in \{0; 0{,}2; 1\}$ et SDM, NGM, DL sur *test-0.4* (gauche) et *test-0.6* (droite).

introduit une normalisation IDF, on observe un effet en moyenne négatif sur le score BLEU (voir figure 7). On peut expliquer ce comportement par le fait que le transformeur Multi-Levenshtein apprend à copier un maximum de tokens, même si ceux-ci sont fréquents. Autrement dit, il favorise la quantité à la qualité. Il vaut donc probablement mieux fournir des exemples plus couvrants, même s'il s'agit de termes communs.

7 Conclusion

Dans cet article, nous étudions comment optimiser la sélection d'un ensemble varié d'exemples dans une mémoire de traduction. Par analogie des travaux en recherche d'information, nous nous appuyons sur la théorie des fonctions sous-modulaires. Dans certaines configurations nous observons un léger gain avec l'utilisation de ce paradigme, avec de grandes variations selon le domaine. Une difficulté de cette approche est qu'elle sélectionne des exemples plus longs que la méthode de base, ce qui rend plus difficile leur édition conjointe et limite les améliorations des scores de traduction.

Dans le futur, nous souhaitons continuer à étudier la relation entre choix des exemples et modèle d'édition, par exemple en entraînant le modèle de recherche d'information conjointement avec le modèle de traduction. Une autre piste consiste à chercher à imposer une contrainte de longueur aux phrases couvrantes, afin qu'elles soient plus exploitables. Enfin, $\texttt{TM}^k\texttt{-LevT}$ apprend à maximiser la couverture de la cible sans se soucier de la rareté des mots couverts. Peut-être vaut-il mieux privilégier l'utilité des mots couverts plutôt que leur nombre, en incitant le modèle à conserver des termes qu'il aurait du mal à regénérer.

8 Remerciements

Ce projet a été partiellement financé par l'ANR dans le cadre du projet TraLaLam (ANR-23-IAS1-0006). Il a également bénéficié des resources HPC/AI de GENCI-IDRIS (2022-AD011013583 et 2023-AD010614012).

Références

AGRAWAL S., ZHOU C., LEWIS M., ZETTLEMOYER L. & GHAZVININEJAD M. (2023). In-context examples selection for machine translation. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., Findings of the Association for Computational Linguistics : ACL 2023, p. 8857–8873, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.564.

BAWDEN R. & YVON F. (2023). Investigating the translation performance of a large multilingual language model : the case of BLOOM. In M. NURMINEN, J. BRENNER, M. KOPONEN, S. LATOMAA, M. MIKHAILOV, F. SCHIERL, T. RANASINGHE, E. VANMASSENHOVE, S. A. VIDAL, N. ARANBERRI, M. NUNZIATINI, C. P. ESCARTÍN, M. FORCADA, M. POPOVIC, C. SCARTON & H. MONIZ, Éds., Proceedings of the 24th Annual Conference of the European Association for Machine Translation, p. 157–170, Tampere, Finland : European Association for Machine Translation.

BOUTHORS M., CREGO J. & YVON F. (2023). Towards example-based NMT with multi-Levenshtein transformers. In H. BOUAMOR, J. PINO & K. BALI, Éds., Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, p. 1830–1846, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.113.

BOWKER L. (2002). Computer-aided translation technology : A practical introduction. University of Ottawa Press.

BULTE B. & TEZCAN A. (2019). Neural fuzzy repair : Integrating fuzzy matches into neural machine translation. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, p. 1800–1809, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1175.

CAI D., WANG Y., LI H., LAM W. & LIU L. (2021). Neural machine translation with monolingual translation memory. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers), p. 7307–7318, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.567.

CARL M., WAY A. & DAELEMANS W. (2004). Recent advances in example-based machine translation. Computational Linguistics, **30**, 516–520. DOI : 10.1162/0891201042544866.

CHENG X., GAO S., LIU L., ZHAO D. & YAN R. (2022). Neural machine translation with contrastive translation memories. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, p. 3591–3601, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.235.

DAUMÉ H., LANGFORD J. & MARCU D. (2009). Search-based structured prediction. Machine Learning, **75**(3), 297–325. DOI : 10.1007/s10994-009-5106-x.

FARAJIAN M. A., TURCHI M., NEGRI M. & FEDERICO M. (2017). Multi-domain neural machine translation through unsupervised adaptation. In O. BOJAR, C. BUCK, R. CHATTERJEE, C.

FEDERMANN, Y. GRAHAM, B. HADDOW, M. HUCK, A. J. YEPES, P. KOEHN & J. KREUTZER, Éds., Proceedings of the Second Conference on Machine Translation, p. 127–137, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/W17-4713.

GOLDSTEIN J. & CARBONELL J. (1998). Summarization : (1) using MMR for diversity- based reranking and (2) evaluating summaries. In TIPSTER TEXT PROGRAM PHASE III : Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998, p. 181–195, Baltimore, Maryland, USA : Association for Computational Linguistics. DOI : 10.3115/1119089.1119120.

GU J., WANG C. & ZHAO J. (2019). Levenshtein transformer. In H. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. D'ALCHÉ-BUC, E. FOX & R. GARNETT, Éds., Advances in Neural Information Processing Systems, volume 32 : Curran Associates, Inc.

GU J., WANG Y., CHO K. & LI V. O. (2018). Search Engine Guided Neural Machine Translation. Proceedings of the AAAI Conference on Artificial Intelligence, **32**(1). DOI : 10.1609/aaai.v32i1.12013.

GUPTA S., GARDNER M. & SINGH S. (2023). Coverage-based example selection for in-context learning. In H. BOUAMOR, J. PINO & K. BALI, Éds., Findings of the Association for Computational Linguistics : EMNLP 2023, p. 13924–13950, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.930.

HE J., NEUBIG G. & BERG-KIRKPATRICK T. (2021a). Efficient nearest neighbor language models. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, p. 5703–5714, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.461.

HE Q., HUANG G., CUI Q., LI L. & LIU L. (2021b). Fast and accurate neural machine translation with translation memory. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers), p. 3170–3180, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.246.

HENDY A., ABDELREHIM M., SHARAF A., RAUNAK V., GABR M., MATSUSHITA H., KIM Y. J., AFIFY M. & AWADALLA H. H. (2023). How good are GPT models at machine translation ? a comprehensive evaluation. CoRR, **abs/2302.09210**. DOI : 10.48550/ARXIV.2302.09210.

HOANG C., SACHAN D., MATHUR P., THOMPSON B. & FEDERICO M. (2022). Improving Retrieval Augmented Neural Machine Translation by Controlling Source and Fuzzy-Match Interactions. arXiv :2210.05047 [cs], DOI : 10.48550/arXiv.2210.05047.

KHANDELWAL U., FAN A., JURAFSKY D., ZETTLEMOYER L. & LEWIS M. (2021). Nearest Neighbor Machine Translation. In Proceedings of the International Conference on Learning Representations.

KNYAZEVA E., WISNIEWSKI G. & YVON F. (2018). Les méthodes « apprendre à chercher » en traitement automatique des langues : un état de l'art [a survey of learning-to-search techniques in natural language processing]. Traitement Automatique des Langues, **59**(1), 39–63.

KOEHN P. & SENELLART J. (2010). Convergence of translation memory and statistical machine translation. In V. ZHECHEV, Éd., Proceedings of the Second Joint EM+/CNGL Workshop : Bringing MT to the User : Research on Integrating MT in the Translation Industry, p. 21–32, Denver, Colorado, USA : Association for Machine Translation in the Americas.

KRAUSE A. & GOLOVIN D. (2014). Submodular function maximization. In L. BORDEAUX, Y. HAMADI & P. KOHLI, Éds., Tractability : Practical Approaches to Hard Problems, p. 71–104. Cambridge University Press.

LI H., SU Y., CAI D., WANG Y. & LIU L. (2022). A survey on retrieval-augmented text generation. CoRR, **abs/2202.01110**.

LIN H. & BILMES J. (2011). A class of submodular functions for document summarization. In D. LIN, Y. MATSUMOTO & R. MIHALCEA, Éds., Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies, p. 510–520, Portland, Oregon, USA : Association for Computational Linguistics.

M A., PUDUPPULLY R., DABRE R. & KUNCHUKUTTAN A. (2023). CTQScorer : Combining multiple features for in-context example selection for machine translation. In H. BOUAMOR, J. PINO & K. BALI, Éds., Findings of the Association for Computational Linguistics : EMNLP 2023, p. 7736–7752, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.519.

MARTINS P. H., MARINHO Z. & MARTINS A. F. T. (2022). Chunk-based nearest neighbor machine translation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, p. 4228–4245, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : https ://aclanthology.org/emnlp-22/2022.emnlp-main.284.

MENG Y., LI X., ZHENG X., WU F., SUN X., ZHANG T. & LI J. (2022). Fast nearest neighbor machine translation. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., Findings of the Association for Computational Linguistics : ACL 2022, p. 555–565, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-acl.47.

MOSLEM Y., HAQUE R., KELLEHER J. D. & WAY A. (2023). Adaptive machine translation with large language models. In M. NURMINEN, J. BRENNER, M. KOPONEN, S. LATOMAA, M. MIKHAILOV, F. SCHIERL, T. RANASINGHE, E. VANMASSENHOVE, S. A. VIDAL, N. ARANBERRI, M. NUNZIATINI, C. P. ESCARTÍN, M. FORCADA, M. POPOVIC, C. SCARTON & H. MONIZ, Éds., Proceedings of the 24th Annual Conference of the European Association for Machine Translation, p. 227–237, Tampere, Finland : European Association for Machine Translation.

MU Y., REHEMAN A., CAO Z., FAN Y., LI B., LI Y., XIAO T., ZHANG C. & ZHU J. (2023). Augmenting large language model translators via translation memories. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., Findings of the Association for Computational Linguistics : ACL 2023, p. 10287–10299, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.653.

NAGAO M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. In A. ELITHORN & R. BANERJI, Éds., Artificial and human intelligence : Elsevier Science Publishers. B.V.

NIWA A., TAKASE S. & OKAZAKI N. (2022). Nearest neighbor non-autoregressive text generation. CoRR, **abs/2208.12496**. DOI : 10.48550/ARXIV.2208.12496.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In P. ISABELLE, E. CHARNIAK & D. LIN, Éds., Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, p. 311–318, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

PHAM M. Q., XU J., CREGO J., YVON F. & SENELLART J. (2020). Priming neural machine translation. In L. BARRAULT, O. BOJAR, F. BOUGARES, R. CHATTERJEE, M. R. COSTA-JUSSÀ, C. FEDERMANN, M. FISHEL, A. FRASER, Y. GRAHAM, P. GUZMAN, B. HADDOW, M. HUCK, A. J. YEPES, P. KOEHN, A. MARTINS, M. MORISHITA, C. MONZ, M. NAGATA, T. NAKAZAWA & M. NEGRI, Éds., Proceedings of the Fifth Conference on Machine Translation, p. 516–527, Online : Association for Computational Linguistics.

POST M. (2018). A call for clarity in reporting BLEU scores. In O. BOJAR, R. CHATTERJEE, C. FEDERMANN, M. FISHEL, Y. GRAHAM, B. HADDOW, M. HUCK, A. J. YEPES, P. KOEHN, C. MONZ, M. NEGRI, A. NÉVÉOL, M. NEVES, M. POST, L. SPECIA, M. TURCHI & K. VERSPOOR, Éds., Proceedings of the Third Conference on Machine Translation : Research Papers, p. 186–191, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6319.

ROBERTSON S. E. & JONES K. S. (1976). Relevance weighting of search terms. Journal of the American Society for Information Science, **27**(3), 129–146. DOI : https ://doi.org/10.1002/asi.4630270302.

ROSS S., GORDON G. & BAGNELL D. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In G. GORDON, D. DUNSON & M. DUDÍK, Éds., Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, volume 15 de Proceedings of Machine Learning Research, p. 627–635, Fort Lauderdale, FL, USA : PMLR.

RUDIN C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, **1**(5), 206–215. DOI : 10.1038/s42256-019-0048-x.

SIA S. & DUH K. (2023). In-context learning as maintaining coherency : A study of on-the-fly machine translation using large language models.

SOMERS H. (1999). Review article : Example-based machine translation. Machine Translation, **14**(2), 113–157. DOI : 10.1023/A :1008109312730.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, p. 6000–6010, Red Hook, NY, USA : Curran Associates Inc.

VILAR D., FREITAG M., CHERRY C., LUO J., RATNAKAR V. & FOSTER G. (2023). Prompting PaLM for translation : Assessing strategies and performance. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), p. 15406–15427, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.859.

XIA M., HUANG G., LIU L. & SHI S. (2019). Graph based translation memory for neural machine translation. Proceedings of the AAAI Conference on Artificial Intelligence, **33**(01), 7297–7304. DOI : 10.1609/aaai.v33i01.33017297.

XU J., CREGO J. & YVON F. (2023). Integrating translation memories into non-autoregressive machine translation. In A. VLACHOS & I. AUGENSTEIN, Éds., Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, p. 1326–1338, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.96.

ZHANG B., HADDOW B. & BIRCH A. (2023). Prompting large language model for machine translation : A case study. In Proceedings of the 40th International Conference on Machine Learning, ICML'23 : JMLR.org.

ZHANG J., UTIYAMA M., SUMITA E., NEUBIG G. & NAKAMURA S. (2018). Guiding neural machine translation with retrieved translation pieces. In M. WALKER, H. JI & A. STENT, Éds., Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers), p. 1325–1335, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1120.

ZHENG K., WANG L., WANG Z., CHEN B., ZHANG M. & TU Z. (2023). Towards a unified training for Levenshtein transformer. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p. 1–5. DOI : 10.1109/ICASSP49357.2023.10094646.

ZHENG X., ZHANG Z., GUO J., HUANG S., CHEN B., LUO W. & CHEN J. (2021). Adaptive nearest neighbor machine translation. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers), p. 368–374, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-short.47.

A Configuration du Modèle

Nous utilisons le transformeur Multi-Levenshtein [7]. La dimension de plongement est 512 ; la taille de la couche linéaire est de 2048 ; le nombre de têtes est 8 ; le nombre de couches de l'encodeur et de décodeur est 6 ; les plongements sont tous partagés ; le taux de dropout est 0,3.

L'entraînement est fait avec un optimisateur Adam $(\beta_1, \beta_2) = (0{,}9; 0{,}98)$; un planificateur racine carrée inversée ; un taux d'apprentissage de $5e^{-4}$; un lissage d'étiquette de 0,1 ; une mise à jour à 10 000 ; une précision flottante de 16. Le nombre d'itérations est fixé à 60k. La taille de lot et le nombre de GPU sont choisis pour avoir en moyenne 450 échantillons par itération. Le modèle est pré-entraîné sur des données synthétiques décrites par Bouthors et al. (2023). Nous utilisons un vocabulaire joint de taille 32k. Nous utilisons le réalignement et le réaffinage itératif avec une pénalité de 3 sur le fait d'insérer 0 à l'étape d'insertion, et un nombre maximum d'itérations de 10. (Gu et al., 2019).

Les données d'entraînement sont les 11 domaines en-fr de Bouthors et al. (2023), dont les 6 domaines que nous avons choisis. Les exemples sont construits avec comme dans la configuration DL, c-à-d avec un préfiltrage des 100 exemples avec les meilleurs scores BM25, puis les 3 meilleurs exemples de similarité DL. Cependant, les phrases avec un score $< 0{,}4$ sont retirées comme dans le papier original.

B Illustration

Le tableau 2 illustre le compromis entre **couverture** et **pertinence** (liée au score DL individuel). Ici, nous effectuons une RPS sur un micro corpus de 11 phrases plus ou moins similaires à la source. Le terme "vert" apparaissant uniquement dans une phrase très peu pertinente (dans le sens où peu de termes de la phrase sont dans la source), un λ trop faible la récupérera pour compléter le dernier terme couvrable. Au contraire, un λ trop élevé a tendance à récupérer des phrases similaires les unes aux autres, avec une haute pertinence, mais une faible couverture.

7. disponible à https://github.com/Maxwell1447/fairseq.

méthode	couv	score DL	**source :** Le chat est assis sur le tapis vert du salon .
DL-0 à DL-0,3	0,91	0,64	Le chat est assis sur le sol .
		0,27	J' ai acheté un nouveau tapis pour le salon .
		0,08	Après une longue journée de marche , au crépuscule , je décide enfin de me reposer à côté d' un grand rocher tout vert de mousse dans la forêt à côté du domaine de Courbetin .
DL-0,4 à DL-0,6	0,82	0,64	Le chat est assis sur le sol .
		0,27	J' ai acheté un nouveau tapis pour le salon .
		0,27	Regarde ce chat , il est assis sur le comptoir .
DL-0,7 à DL-0,9	0,64	0,64	Le chat est assis sur le sol .
		0,27	Regarde ce chat , il est assis sur le comptoir .
		0,45	Le chat est assis à l' entrée .
DL-1	0,64	0,64	Le chat est assis sur le sol .
		0,45	Le chat est assis à l' entrée .
		0,36	Le chat est dans une boîte en carton .
DL-MMR	0,82	0,64	Le chat est assis sur le sol .
		0,45	Le chat est assis à l' entrée .
		0,27	J' ai acheté un nouveau tapis pour le salon .

TABLE 2 – Illustration des effets des paramètres de RPS (λ et MMR) pour le score DL, avec un compromis entre couverture globale de la source (couv) et la proximité individuelle des phrases exemples. Les termes couvrants sont soulignés.

C Démonstrations

C.1 Sous-modularité de la *couverture pondérée lissée*

Lemme 1. *Soit $f : 2^{\Omega} \to \mathbb{R}$. La propriété suivante découle de la définition de la sous-modularité de f : f sous-modulaire si et seulement si $\forall X \subseteq \Omega$, $\forall x_1, x_2 \in \Omega \setminus X$ s.t. $x_1 \neq x_2$:*

$$f(X \cup \{x_1\}) + f(X \cup \{x_2\}) \geq f(X \cup \{x_1, x_2\}) + f(X)$$

Nous utilisons le lemme 1 pour montrer que la fonction de *couverture pondérée lissée* (2) est sous-modulaire.

Démonstration. Soit $f_n(Z)$ le terme dans l'équation (2) tel que $f(Z) = \sum_{n=1}^{N} f_n(Z)$:

$$f_n(Z) = v_0 + \lambda v_1 + \cdots + \lambda^{K-1} v_{K-1}, \tag{3}$$

où les v_i sont triés dans l'ordre décroissants, et $K = |Z|$. Soient v et v' la composante en n de deux nouveaux éléments de $\Omega \setminus Z$. On considère que $v \geq v'$. On nomme i et j les indices tels que : $v_i \geq v \geq v_{i+1}$ et $v_j \geq v' \geq v_{j+1}$.

- $f_n(Z \cup \{v\}) = v_0 + \cdots + \lambda^i v_i + \lambda^{i+1} v + \lambda^{i+2} v_{i+1} + \cdots + \lambda^K v_{K-1}$
- $f_n(Z \cup \{v'\}) = v_0 + \cdots + \lambda^j v_j + \lambda^{j+1} v' + \lambda^{j+2} v_{j+1} + \cdots + \lambda^K v_{K-1}$
- $f_n(Z \cup \{v, v'\}) = v_0 + \cdots + \lambda^i v_i + \lambda^{i+1} v + \lambda^{i+2} v_{i+1} + \cdots$
  $+ \lambda^{j+1} v_j + \lambda^{j+2} v' + \lambda^{j+3} v_{j+1} + \cdots + \lambda^{K+1} v_{K-1}$

On simplifie en nommant $L = \sum_{l \leq i} \lambda^l v_l$, $M = \sum_{i<l\leq j} \lambda^l v_l$ et $R = \sum_{j<l} \lambda^l v_l$, ce qui donne :

- $f_n(Z \cup \{v\}) = L + \lambda^i v + \lambda(M + R)$
- $f_n(Z \cup \{v'\}) = L + M + \lambda^j v + \lambda R$
- $f_n(Z \cup \{v, v'\}) = L + \lambda^i v + \lambda M + \lambda^{j+1} v' + \lambda^2 R$

Pour montrer que f_n est sous-modulaire, il suffit de montrer que ce terme est positif :

$$f_n(Z \cup \{v\}) + f_n(Z \cup \{v'\}) - f_n(Z \cup \{v, v'\}) - f_n(Z) = (1 - \lambda) \left[\lambda^j v' - (1 - \lambda) R\right]$$

Puisque $\lambda \in [0, 1]$, alors $1 - \lambda > 0$. D'autre part :

$$\begin{aligned} R \leq \lambda^{j+1} v_{j+1} &\Rightarrow -(1 - \lambda) R \geq -(1 - \lambda) \lambda^{j+1} v_{j+1} \\ &\Rightarrow \lambda^j v' - (1 - \lambda) R \geq \lambda^j v' - (1 - \lambda) \lambda^{j+1} v_{j+1} \end{aligned}$$

$\lambda^j > \lambda^{j+1}(1 - \lambda) \geq 0$ et $v' \geq v_{j+1} \geq 0$, donc $\lambda^j v' \geq (1 - \lambda) \lambda^{j+1} v_{j+1}$.

$f_n(Z \cup \{v\}) + f_n(Z \cup \{v'\}) - f_n(Z \cup \{v, v'\}) - f_n(Z) \geq 0$ donc f_n sous-modulaire. Puisque f est une somme de fonction sous-modulaire, alors f sous-modulaire. □

C.2 Majoration de l'erreur d'approximation

Théorème 1. *Soit f une fonction de couverture pondérée lissée de paramètre λ. Soit Z la solution de l'algorithme 2 et k le nombre d'exemples :*

$$f(Z) \geq \frac{1}{k} \left(\sum_{j=0}^{k-1} \lambda^j \right) \max_{\bar{Z} : |\bar{Z}| = k} f(Z) \tag{4}$$

Si $\lambda = 1$, la solution est optimale. Si $\lambda = 0$, elle est au pire $\frac{1}{k}$ de l'optimum. Dans le cas où les poids de couverture sont booléens (0 ou 1), et $\lambda = 0$, l'algorithme 2 est équivalent au 1 (hormis la partie de réinitialisation), ce qui nous permet de retomber sur la borne bien connue de $(1 - 1/e)$. Nous montrons également que dans des cas limites spécifiques, cette borne peut être atteinte.

Démonstration. L'équation (4) se démontre via les Lemmes (2) et (4). Nous adoptons la notation suivante : $\Delta(v|Z) := f(Z \cup \{v\}) - f(Z)$. Soit Z_k^* un maximiseur de f de taille k, et Z_k, la solution construite par l'algorithme 2, avec Z_i (pour $i < k$) les itérations successives depuis $Z_0 = \emptyset$. On note aussi, $Z_{i+1} = Z_i \cup \{v_{i+1}\}$.

Lemme 2.

$$\forall k > 0, f(Z_k^*) \leq \sum_{i=1}^{k} f(\{v_i^*\}),$$

où v_i^* sont les éléments de Z_k^* tels que $f(\{v_i^*\})$ est une suite décroissante.

Démonstration.

$$\begin{aligned} f(Z_k^*) &= \sum_{i=1}^{k} \Delta(v_i^* | \{v_1^*, \cdots, v_{i-1}^*\}) && \text{par téléscopage} \\ &\leq \sum_{i=1}^{k} \Delta(v_i^* | \emptyset) && \text{car } f \text{ sous-modulaire} \\ &= \sum_{i=1}^{k} f(\{v_i^*\}) \end{aligned}$$

□

Lemme 3.

$$\forall i, \forall v \notin Z_i, \Delta(v|Z_i) \geq w_i^T v,$$

Démonstration. Étant donné $n < N$, on nomme $z_i^{(n)}$ les éléments triés de $v_{[1:k]}^{(n)}$, tels que $z_1^{(n)} \geq \cdots \geq z_i^{(n)}$. f s'exprime comme :

$$f(Z_i) = \sum_{n=1}^{N} z_i^{(n)} \lambda^{i-1}$$

Pour calculer $f(Z_i \cup \{v\})$ pour n'importe quel v, pour chaque n, $v^{(n)}$ est inséré dans la séquence ordonnée des $z_i^{(n)}$. Soit $m(n)$ la position d'insertion. Autrement dit, $z_1^{(n)} \geq \cdots \geq z_{m(n)}^{(n)} \geq v^{(n)} \geq z_{m(n)+1}^{(n)} \geq \cdots \geq z_i^{(n)}$.

De même, on appelle $L^{(n)}$ et $R^{(n)}$ les parties gauche et droite de la somme :

$$L^{(n)} = \sum_{i \leq m(n)} z_i^{(n)} \lambda^{i-1}, \qquad R^{(n)} = \sum_{i > m(n)} z_i^{(n)} \lambda^{i-1}$$

Ainsi :

$$f(Z_i \cup \{v\}) = \sum_{n=1}^{N} L^{(n)} + \lambda^{m(n)} v^{(n)} + \lambda R^{(n)}.$$

$$\begin{aligned}\Delta(v|Z_i) &= f(Z_i \cup \{v\}) - f(Z_i) \\ &= \sum_{n=1}^{N} \left[L^{(n)} + \lambda^{m(n)} v^{(n)} + \lambda R^{(n)}\right] - \sum_{n=1}^{N} \left[L^{(n)} + R^{(n)}\right] \\ &= \sum_{n=1}^{N} \lambda^{m(n)} v^{(n)} - (1-\lambda) \sum_{n=1}^{N} R^{(n)}\end{aligned}$$

Maintenant, à propos de $w_{i-1}^T v$, on note $c(n)$ le nombre de $v_j^{(n)}$ strictement positifs ($1 \leq j < i$). Par construction, $w_{i-1}^{(n)} = \lambda^{c(n)}$, d'où :

$$w_{i-1}^T v = \sum_{n=1}^{N} \lambda^{c(n)} v^{(n)}.$$

$$\begin{aligned}&\Delta(v|Z_i) - w_{i-1}^T v \\ &= \sum_{n=1}^{N} \lambda^{m(n)} v^{(n)} - (1-\lambda) \sum_{n=1}^{N} R^{(n)} - \sum_{n=1}^{N} \lambda^{c(n)} v^{(n)} \\ &= \sum_{n=1}^{N} (\lambda^{m(n)} - \lambda^{c(n)}) v^{(n)} - (1-\lambda) \sum_{n=1}^{N} \sum_{i > m(n)} z_i^{(n)} \lambda^{i-1} \\ &\geq \sum_{n=1}^{N} (\lambda^{m(n)} - \lambda^{c(n)}) v^{(n)} - (1-\lambda) \sum_{n=1}^{N} \sum_{m(n) < i \leq c(n)} z_i^{(n)} \lambda^{i-1} \qquad \text{par construction de } c(n) \\ &\geq \sum_{n=1}^{N} (\lambda^{m(n)} - \lambda^{c(n)}) v^{(n)} - \sum_{n=1}^{N} v^{(n)} (1-\lambda) \sum_{m(n) < i \leq c(n)} \lambda^{i-1} \qquad \text{pour ces } i, z_i^{(n)} \leq v^{(n)} \\ &= \sum_{n=1}^{N} (\lambda^{m(n)} - \lambda^{c(n)}) v^{(n)} - \sum_{n=1}^{N} v^{(n)} (\lambda^{m(n)} - \lambda^{c(n)}) \qquad \text{somme téléscopique} \\ &= 0\end{aligned}$$

□

Lemme 4.

$$\forall k > 0, f(Z_k) \geq \sum_{i=1}^{k} f(\{v_i^*\}) \lambda^{i-1},$$

où v_i^* sont éléments de Z_k^* tels que $f(\{v_i^*\})$ est une suite décroissante.

Démonstration.

$$\begin{aligned} f(Z_k) &= \sum_{i=1}^{k} \Delta(v_i|Z_{i-1}) \\ &\geq \sum_{i=1}^{k} w_{i-1}^T v_i && \text{grâce au Lemme (3)} \end{aligned}$$

Soit $i \leq k$. Prouvons que $w_{i-1}^T v_i \geq \lambda^{i-1} f(\{v_i^*\}) = \lambda^{i-1} 1^T v_i^*$. Il y a deux cas :

— Si $v_i^* \notin Z_i$, alors par construction : $w_{i-1}^T v_i \geq w_{i-1}^T v_i^*$. Puisque chaque $w_{i-1}^{(n)} \geq \lambda^{i-1}$, alors $w_{i-1}^T v_i \geq \lambda^{i-1} 1^T v_i^*$.

— Si $v_i^* \in Z_i$, on sait que $\exists j < i$ tel que $v_j^* \notin Z_i$. Donc $w_{i-1}^T v_i \geq w_{i-1}^T v_j^* \geq \lambda^{i-1} 1^T v_j^*$. Puique $f(\{v_i^*\})$ suite décroissante, $1^T v_j^* \geq 1^T v_i^*$. Par conséquent, $w_{i-1}^T v_i \geq \lambda^{i-1} 1^T v_i^*$.

D'où, $f(Z_k) \geq \sum_{i=1}^{k} f(\{v_i^*\}) \lambda^{i-1}$. □

En combinant les Lemmes (2) et (4), on obtient :

$$\frac{f(Z_k)}{f(Z_k^*)} \geq \frac{\sum_{i=1}^{k} f(\{v_i^*\}) \lambda^{i-1}}{\sum_{i=1}^{k} f(\{v_i^*\})}.$$

Grâce à l'inégalité de Tchebychev pour les sommes :

$$\frac{f(Z_k)}{f(Z_k^*)} \geq \frac{\sum_{i=1}^{k} \lambda^{i-1}}{k} = \frac{1}{k} \frac{1-\lambda^k}{1-\lambda},$$

D'où la borne. □

Enfin, nous pouvons montrer que la borne est atteinte dans un scénario, signifiant qu'il s'agit bien de la meilleure borne.

Démonstration. Pour cela, nous supposons $N > k$.

On suppose que l'ensemble de candidats est $T = (z_1, \cdots, z_N, e_1, \cdots e_N)$, tel que

$$\begin{cases} \forall i, z_i = (\frac{1}{N}, \cdots, \frac{1}{N}) \\ \forall i, e_i^{(n)} = \mathbb{1}(i = n) \end{cases}$$

Si l'algorithme s'obstine à choisir seulement des z_j, alors $w_i = (\lambda^i, \cdots, \lambda^i)$. Ce faisant, $\forall z \in T, w_i^T z = \lambda^i$. Par conséquent, l'algorithme peut choisir de manière équivalente n'importe quel $z \in T$ pour la prochaine itération. Si il ne choisit que des z_j pour Z_k, alors :

$$f(Z_k) = 1 + \lambda + \cdots + \lambda^{k-1} = \frac{1-\lambda^k}{1-\lambda}$$

D'autre part, le meilleur ensemble Z_k^* de k éléments est constitué uniquement de e_j :

$$f(Z_k^*) = k$$

Le ratio est celui de l'équation (4). □

D Compléments

Pour avoir une vision complète des résultats, nous présentons les tableaux de couverture, pertinence et longueur (tableau 3), ainsi que des scores BLEU par domaine (tableau 4 et tableau 5).

Les domaines offrent une grande variabilité dont il est difficile d'extraire des tendances. Notons cependant que Europarl et Wikipedia ne bénéficient pas d'une plus grande couverture car ce sont les cas où $\lambda = 1$ qui donnent le meilleur BLEU. C'est le contraire pour ECB et JRC-Aquis.

	test-0.4			*test-0.6*		
recherche	couv.	perti.	long.	couv.	perti.	long.
SDM-0	70,8	36,7	35,6	77,4	46,0	31,3
SDM-0,2	71,0	36,5	36,1	77,5	45,7	31,7
SDM-1	66,7	38,1	37,8	75,2	47,7	32,1
SDM-IDF-0	70,4	36,3	34,8	77,2	45,5	31,4
SDM-IDF-0,2	70,8	36,4	35,0	77,4	45,5	31,5
SDM-IDF-1	68,5	37,7	37,6	75,3	46,9	32,4
NGM-0	70,3	38,5	30,5	77,4	47,2	28,3
NGM-0,2	70,3	39,1	31,0	77,4	48,1	28,6
NGM-1	66,6	40,4	32,4	74,6	49,7	29,1
NGM-IDF-0	70,4	38,3	30,3	77,5	46,9	28,1
NGM-IDF-0,2	70,6	38,9	30,8	77,5	47,7	28,4
NGM-IDF-1	66,9	40,2	32,3	74,8	49,1	29,3
DL-0	70,6	40,5	27,5	78,6	50,5	24,5
DL-0,2	71,2	39,0	31,7	78,9	48,7	28,0
DL-1	62,1	47,2	20,0	72,3	56,9	19,7
DL-IDF-0	70,6	39,5	28,1	78,6	49,5	25,1
DL-IDF-0,2	70,9	38,7	30,9	78,8	48,3	27,8
DL-IDF-1	62,1	47,2	20,0	72,3	56,9	19,7
DL-MMR	64,3	46,1	20,4	73,6	56,3	19,9

TABLE 3 – Couverture, pertinence et longueur moyennes selon le score de recherche choisi, avec ou sans normalisation IDF, et $\lambda \in \{0; 0{,}2; 1\}$.

	ECB	EME	Epp	JRC	Ubu	Wiki	moy
test-0.4							
SDM-0	55,0	54,1	33,8	66,8	45,7	32,5	48,0
SDM-0,2	54,7	54,4	33,2	66,8	45,4	32,2	47,8
SDM-1	55,1	53,1	34,3	66,9	45,0	33,1	47,9
SDM-IDF-0	55,1	54,5	33,5	67,1	45,9	32,0	48,0
SDM-IDF-0,2	55,0	54,4	33,6	67,0	45,4	32,3	47,9
SDM-IDF-1	55,5	53,4	33,1	67,3	44,7	33,2	47,9
NGM-0	56,1	56,6	34,7	68,2	46,0	33,6	49,2
NGM-0,2	56,2	56,2	34,9	68,3	45,8	33,9	49,2
NGM-1	56,0	55,7	35,3	67,8	45,3	33,8	49,0
NGM-IDF-0	55,7	56,4	34,4	68,0	45,5	33,5	48,9
NGM-IDF-0,2	56,0	56,5	34,7	68,1	45,4	33,8	49,1
NGM-IDF-1	55,8	55,6	34,9	67,8	45,7	34,7	49,1
DL-0	57,0	55,6	35,2	68,7	46,1	33,9	49,4
DL-0,2	56,6	55,6	35,0	68,4	45,7	33,8	49,2
DL-1	55,9	55,2	35,5	68,0	46,6	34,7	49,3
DL-IDF-0	56,1	55,5	34,5	68,1	46,1	33,5	49,0
DL-IDF-0,2	56,3	55,4	34,6	68,0	46,9	34,1	49,2
DL-IDF-1	55,9	55,2	35,4	68,0	46,7	34,8	49,3
test-0.6							
SDM-0	64,5	62,5	50,1	75,7	53,6	60,8	61,2
SDM-0,2	64,1	62,5	50,1	75,8	53,9	60,6	61,2
SDM-1	63,8	61,9	50,1	74,8	52,3	61,2	60,7
SDM-IDF-0	64,9	61,6	49,4	76,0	52,8	59,9	60,8
SDM-IDF-0,2	64,7	61,9	49,2	75,8	53,6	60,2	60,9
SDM-IDF-1	64,7	61,8	50,0	75,5	52,0	61,2	60,9
NGM-0	64,2	63,6	51,0	76,7	53,0	61,1	61,6
NGM-0,2	64,9	63,7	51,2	76,8	53,2	61,5	61,9
NGM-1	64,8	63,8	51,2	76,5	52,5	62,9	61,9
NGM-IDF-0	64,7	64,0	51,1	76,8	53,1	61,4	61,8
NGM-IDF-0,2	65,7	63,5	51,0	76,8	53,5	61,5	62,0
NGM-IDF-1	65,5	63,8	50,8	76,3	52,6	62,0	61,8
DL-0	65,4	65,5	50,6	77,1	54,2	62,0	62,5
DL-0,2	65,4	65,8	51,0	77,0	53,9	61,9	62,5
DL-1	64,9	65,1	51,2	76,8	53,3	62,4	62,3
DL-IDF-0	65,3	65,2	50,7	77,1	53,7	61,3	62,2
DL-IDF-0,2	65,7	64,9	51,0	76,9	53,9	61,2	62,3
DL-IDF-1	64,9	65,1	51,2	76,8	53,3	62,4	62,3

TABLE 4 – Score BLEU selon le score de recherche choisi, avec ou sans normalisation IDF, et $\lambda \in \{0; 0{,}2; 1\}$.

λ	ECB	EME	Epp	JRC	Ubu	Wiki	moy
test-0.4							
0	57,0	55,6	35,2	68,7	46,1	33,9	49,4
0,1	56,4	55,3	34,7	68,4	46,3	33,7	49,1
0,2	56,6	55,6	35,0	68,4	45,7	33,8	49,2
0,3	56,5	55,4	35,0	68,6	45,9	33,6	49,2
0,4	56,9	55,5	34,8	68,2	45,6	33,9	49,1
0,5	57,0	56,0	34,5	68,4	46,0	34,5	49,4
0,6	57,1	55,8	34,7	68,2	45,7	34,5	49,3
0,7	56,8	55,6	35,1	68,2	45,5	34,4	49,3
0,8	56,6	55,2	35,0	68,2	45,0	34,3	49,1
0,9	56,5	55,4	35,2	68,1	45,2	34,2	49,1
1	55,9	55,2	35,5	68,0	46,6	34,7	49,3
test-0.6							
0	65,4	65,5	50,6	77,1	54,2	62,0	62,5
0,1	65,5	65,4	50,8	77,0	53,8	61,8	62,4
0,2	65,4	65,8	51,0	77,0	53,9	61,9	62,5
0,3	65,6	65,5	51,2	77,1	53,8	62,0	62,5
0,4	65,9	65,3	51,2	77,1	54,2	62,1	62,6
0,5	66,0	65,2	51,2	77,2	54,4	62,1	62,7
0,6	66,0	65,0	51,1	77,1	54,2	62,4	62,6
0,7	65,7	64,7	51,1	76,9	54,1	62,3	62,5
0,8	65,6	65,1	50,8	76,8	53,7	62,4	62,4
0,9	65,5	65,2	50,8	76,8	53,5	62,4	62,3
1	64,9	65,1	51,2	76,8	53,3	62,4	62,3

TABLE 5 – Score BLEU selon la valeur de λ pour DL avec normalisation cardinale.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

ParaPLUIE - une mesure automatique d'évaluation de la qualité sémantique des systèmes de paraphrases

Quentin Lemesle[1] Jonathan Chevelu[1] Damien Lolive[1] Arnaud Delhay[1]
Philippe Martin[1]
(1) Univ Rennes, IRISA, CNRS, 22300 Lannion, France
{quentin.lemesle, jonathan.chevelu, damien.lolive,
arnaud.delhay, philippe.martin}@irisa.fr,

RÉSUMÉ

L'évaluation des systèmes de production automatique de paraphrases est une tâche difficile car elle implique, entre autre, d'évaluer la proximité sémantique entre deux phrases. Les mesures traditionnelles s'appuient sur des distances lexicales, ou au mieux des alignements de plongements sémantiques. Dans cet article nous étudions certaines de ces mesures sur des corpus de paraphrases et de non-paraphrases reconnus pour leurs qualités ou difficultés sur cette tâche. Nous proposons une nouvelle mesure, ParaPLUIE, s'appuyant sur l'utilisation d'un grand modèle de langue. D'après nos expériences, celui-ci est plus à même de trier les paires de phrases par proximité sémantique.

ABSTRACT

ParaPLUIE : ParaPhrase, Llm Used for Improved Evaluation

Evaluating automatic paraphrase production systems is a difficult task because it involves, among other things, assessing the semantic proximity between two sentences. Usual measures are based on lexical distances, or at least on semantic embedding alignments. In this article we study some of these measures on datasets of paraphrases and non-paraphrases known for their quality or difficulty on this task. We propose a new measure, ParaPLUIE, based on the use of a large language model. According to our experiments, this one is better to sort pairs of sentences by semantic proximity.

MOTS-CLÉS : paraphrase, évaluation sémantique, grand modèle de langue.

KEYWORDS: paraphrase, semantic evaluation, large language model.

1 Introduction

Dans le domaine de la production automatique de paraphrases, de nombreuses définitions d'une paraphrase ont été proposées (Mel'čuk, 1997; Barzilay & McKeown, 2001; Sekine, 2005; Zhao *et al.*, 2009; Fabre *et al.*, 2021). Toutes ces définitions, tout comme les travaux de linguistiques traitant de paraphrase (Leeman, 1973), contiennent une notion de conservation du sens, celle-ci étant par nature ambiguë.

Malgré cela, les systèmes ont besoin de mesures automatiques de proximité sémantique pour s'entraîner ou se comparer. Généralement, celles utilisées fonctionnent par comparaison de lexiques (Papineni *et al.*, 2002) ou de plongements (Zhang *et al.*, 2020). Par construction, les approches par lexiques

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 605–616.

ont des difficultés à rapprocher des transformations simples comme le remplacement d'un mot par son synonyme (Banerjee & Lavie, 2005). Elles auront aussi du mal à rejeter deux phrases proches syntaxiquement même si elles ont un sens opposé. D'un autre coté, les mesures utilisant des plongements sémantiques reposent sur une notion d'alignement sous-phrastique sans vision globale des phrases. Ces deux points ont été mis en évidence par Zhang *et al.* (2019) et ont conduit à construire le corpus PAWS.

L'architecture *Transformer* et l'émergence des grands modèles de langues (*LLM*) ont permis de nombreuses avancées dans le domaine du traitement automatique du langage (Vaswani *et al.*, 2017). En particulier, le mécanisme d'auto-attention permet de capturer, sur un grand contexte, des relations sémantiques. Nous proposons d'explorer l'utilisation d'un *LLM* pour la mise au point d'une nouvelle mesure de similarité sémantique nommée ParaPLUIE.

Nous commençons par présenter, dans la section 2, les mesures communément utilisées, puis les corpus d'évaluation en section 3. Dans la section 4, nous présentons une expérience préliminaire d'utilisation d'un *LLM* en tant que classifieur de paraphrases avant de définir ParaPLUIE en section 5. Nous comparons ensuite ParaPLUIE aux autres mesures, en terme de dynamique des scores en section 6, puis en terme de re-classement de paraphrases en section 7.

2 Mesures sémantiques automatiques pour les paraphrases

L'état de l'art des mesures d'évaluation automatique de conservation de sens entre deux phrases peut être séparé en deux groupes. Le premier groupe concerne les mesures estimant la distance lexicale qui ont pour but d'évaluer à quel point les structures de deux phrases sont similaires. Le second regroupe les mesures estimant la proximité sémantique entre deux phrases.

Dans le premier groupe, on peut inscrire la distance de Levenshtein (LEV.) (Levenshtein, 1965), WER (Woodard & Nelson, 1982), BLEU (Papineni *et al.*, 2002) et METEOR (Banerjee & Lavie, 2005).

LEV. donne une mesure de différence entre deux chaînes de caractères. Cette mesure repose sur la détermination du nombre minimal de suppressions, insertions et remplacements pour passer de la chaîne de caractères hypothèse à la chaîne référence. LEV. augmentant avec la taille des chaînes de caractères considérées, elle est généralement normalisée par la taille de la chaîne de caractères la plus grande parmi l'hypothèse et la référence.

WER est un dérivé de LEV., travaillant au niveau des mots et non des caractères. Cette mesure correspond au rapport entre le nombre de mots communs entre la phrase hypothèse et celle de référence, et le nombre de mots de la plus longue des deux phrases.

BLEU a été conçu pour être une mesure de qualité de traduction. Elle consiste en la comptabilisation de la présence de ngrammes d'une phrase hypothèse dans une ou plusieurs phrases de référence. Généralement, tous les ngrammes de longueur 1 à 4 mots sont considérés. BLEU est associé à un score de rappel des ngrammes de l'hypothèse dans la référence. Par la suite, dans cet article, la version de BLEU que nous utilisons est l'implémentation de *torchtext* [1] avec les paramètres par défaut.

METEOR reprend les principes de BLEU en calculant une moyenne harmonique, à partir de la précision et du rappel de l'apparition d'un ngramme hypothèse parmi les références. De plus, METEOR prend

1. https://pytorch.org/text/stable/data_metrics.html

en compte une correspondance synonymique lors du calcul du score. METEOR a montré une meilleure corrélation avec le jugement humain que BLEU.

On pourrait argumenter que si deux phrases ont une structure lexicale très proche alors elles sont plus probablement des paraphrases. La faiblesse de cette hypothèse est que deux phrases peuvent partager une structure commune sans véhiculer le même sens. Pour pallier ce problème, un effort de recherche a été employé à la création d'un second groupe de mesures qui reposent sur une distance sémantique. Ces métriques s'appuient sur les plongements de symboles représentant des mots au sein d'un *LLM*. Dans ce groupe, on peut considérer notamment $\text{BERT}_{\text{score}}$ (Zhang *et al.*, 2020) et ParaScore (Shen *et al.*, 2022).

$\text{BERT}_{\text{score}}$ est un score de similarité de chaque plongement de symbole, composant une phrase hypothèse, avec les plongements de symboles d'une phrase de référence. Sa définition repose sur l'hypothèse que, s'il existe un appariement entre deux phrases tel que, tous les plongements qui les composent sont proches, alors leur sens est proche. Par la suite, dans cet article, nous utilisons la version de $\text{BERT}_{\text{score}}$ provenant de *Hugging Face* [2]. Celle-ci utilise le modèle BERT (Devlin *et al.*, 2019), nous spécifions le type du modèle en tant que *"bert-base-uncased"*.

Shen *et al.* (2022) observent que lorsque les distances lexicales séparant deux paraphrases augmentent, les performances des mesures diminuent. Ils proposent donc ParaScore, une mesure qui étend $\text{BERT}_{\text{score}}$, en ajoutant au calcul de similarité une distance de Levenshtein normalisée.

Il est à noter que les mesures de similarité sémantiques considèrent un alignement mot-à-mot, sans prise en compte de relations sémantiques de plus hauts niveaux. Ainsi un risque quant à la qualité de la classification de paraphrase existe.

Dans la suite de ce document, nous proposons une évaluation de ces différentes métriques dans le cadre de la similarité sémantique, sur deux corpus de paraphrases.

3 Corpus

L'étude des mesures automatiques, pour mesurer la proximité sémantique de paires de phrases, implique l'utilisation d'un corpus annoté en paraphrase/non paraphrase. Idéalement, afin de mesurer la pertinence des mesures dans des cas difficiles, les couples étiquetés non-paraphrases doivent être proches lexicalement ou sémantiquement (sans toutefois être considérés comme paraphrases par des évaluateurs humains). Notre choix s'est donc porté sur deux corpus en langue anglaise : PAWS (Zhang *et al.*, 2019), construit pour tromper les mesures lexicales, et MRPC (Dolan & Brockett, 2005) contenant des exemples d'inférence sémantique (mais asymétrique).

Pour PAWS, nous utilisons ici le sous-ensemble *dev*. Celui-ci comprend 8 000 couples dont 3 539 sont des paraphrases, soit 44% du corpus. L'entièreté du corpus est formée de 108 463 couples et a été générée de façon semi-automatique par inversion de mots et par traduction inverse. Lors de l'annotation de ces données, pour chaque couple de phrases, cinq juges ont été interrogés pour déterminer de façon binaire si les deux phrases sont des paraphrases. PAWS a été conçu pour être un challenge pour les modèles automatiques de détection de paraphrases. En effet, la génération de phrases par inversion de mots génère souvent des non-paraphrases, tout en maintenant une forte similarité lexicale. Voici un exemple de non-paraphrase caractéristique de PAWS : « *flights from New*

2. https://huggingface.co/spaces/evaluate-metric/bertscore

York to Florida » et « *flights from Florida to New York* ».

Le corpus MRPC utilisé est disponible sur *HuggingFace* [3] et comprend 5 801 couples dont 3 900 paraphrases, soit 67% du corpus. Ce corpus a été créé de façon automatique, depuis un grand corpus d'articles de presse regroupés par thème. Lors de l'annotation de ces données, le protocole a été le suivant : pour chaque couple de phrases, deux juges ont été interrogés pour savoir si les deux phrases pouvaient être considérées comme sémantiquement équivalentes ; ils ne pouvaient répondre que de façon binaire, par oui ou par non, et en cas de désaccord entre les deux jugements, un troisième juge répondait à la même consigne. Voici un exemple de non-paraphrase caractéristique de MRPC : « *Last year, Bush appointed him to the Homeland Security Advisory Council.* » et « *He has also served on the president's Homeland Security Advisory Council.* ».

L'ensemble des deux corpus comporte 54% de paraphrases. Les distributions, des mesures présentées à la section 2 et appliquées sur chacune des classes de ces deux corpus, sont présentées dans la table 1. On constate effectivement que les paires de phrases de PAWS sont très proches contrairement à MRPC.

Au sein d'un même corpus, pour toutes les mesures considérées (sauf ParaScore sur PAWS), on constate que les moyennes des différentes classes sont bien cohérentes avec l'étiquette de référence. En revanche, les écarts-types laissent penser qu'une classification ou qu'un tri des phrases en fonction de leur score, ne permettraient pas de bien identifier les paraphrases. De plus, si on croise les résultats des corpus, on constate que le score moyen des non-paraphrases de PAWS est meilleur que le score moyen des paraphrases de MRPC, quelle que soit la mesure. Rappelons toutefois que nous traitons volontairement des corpus très difficiles pour des mesures de proximité sémantique.

Corpus	*Para.*	Taille	LEV. ↓	WER ↓	BLEU ↑	METEOR ↑	$\text{BERT}_{\text{score}}$ ↑	ParaScore ↑
MRPC	Oui	3900	0,38 ±0,16	0,51 ±0,20	0,40 ±0,21	0,69 ±0,14	0,82 ±0,07	0,83 ±0,07
	Non	1901	0,51 ±0,13	0,67 ±0,19	0,28 ±0,18	0,56 ±0,15	0,74 ±0,08	0,76 ±0,09
PAWS	Oui	3539	0,20 ±0,15	0,26 ±0,18	0,62 ±0,18	0,91 ±0,06	0,94 ±0,04	0,92 ±0,03
	Non	4461	0,32 ±0,15	0,37 ±0,18	0,49 ±0,19	0,88 ±0,07	0,91 ±0,04	0,92 ±0,04
Total	Oui	7439	0,30 ±0,18	0,39 ±0,23	0,51 ±0,22	0,80 ±0,16	0,88 ±0,08	0,88 ±0,07
	Non	6362	0,38 ±0,17	0,46 ±0,23	0,43 ±0,21	0,79 ±0,18	0,86 ±0,10	0,87 ±0,09

TABLE 1 – Moyenne des scores de chaque mesure sur les corpus MRPC et PAWS. Les corpus ont été découpés en sous-corpus séparant les paraphrases et non-paraphrases (colonne *Para.*). La taille des corpus dénote le nombre de couples de phrases. Le signe ↑ associé à une mesure indique que plus sa valeur est élevée, meilleur est son score, et ↓ signale l'inverse.

4 Expérience préliminaire : classer avec un *LLM*

Les mesures actuelles se focalisent sur la notion de proximité lexicale ou au mieux d'alignement entre plongements de mots. En conséquence, elles ne peuvent pas prendre en compte des relations complexes entre paraphrases. Récemment, les progrès des architectures de type *Transformer* ont montré qu'il était possible d'avoir une meilleure prise en compte de relations internes à un texte, grâce

3. https://huggingface.co/docs/datasets/v1.13.0/about_dataset_features.html?highlight=mrpc

aux mécanismes d'auto-attention (Vaswani *et al.*, 2017). Pour la création d'une mesure de proximité sémantique, notre intuition est que ces derniers seraient plus à même « d'aligner » les parties des phrases ayant un sens proche.

Afin de vérifier la capacité d'un *LLM* à détecter la relation de paraphrase, une expérience préliminaire consiste à l'utiliser en mode génératif sur le principe d'un agent conversationnel. Pour cette expérience, le modèle Mistral (Jiang *et al.*, 2023), un modèle de taille intermédiaire parmi les grands modèles de langue est utilisé. Précisément, nous utilisons la version *7B Instruct v0-2* [4] au format demi-précision. Ce modèle est basé sur l'architecture *Transformer* et utilise une fenêtre d'attention glissante, dans le but de réduire le coût de calcul. Comme son nom de version l'indique, il possède 7 milliards de paramètres. Le corpus utilisé lors de son entraînement n'est pas divulgué. Dans cette configuration, l'empreinte mémoire du modèle est de 15 gigaoctets. Les expériences présentées ici ont été réalisées sur une machine équipée d'une carte carte graphique Nvidia RTX 4090.

On pose la paraphrase hypothétique (H) et la phrase de référence (R). Le classifieur binaire considéré ici consiste à appliquer un patron sur cette paire (Patron(H, R)) et à utiliser le modèle pour produire les symboles suivants les plus probables. Puisque ce modèle fonctionne sur un principe d'agent conversationnel, le patron simule le début d'un échange entre un utilisateur ($user$), et assistant ($assistant$). Un premier patron testé consiste à présenter les deux phrases et à demander si elles ont le même sens. Si le premier symbole produit en anglais est « *yes* », alors les paires sont considérées comme paraphrases. Toute autre réponse (« *no* », « *they are similar* », ...) est associée à un label négatif. Ce patron noté **Patron**$_{\text{Direct}}$ est précisément construit comme ceci :

Patron$_{\text{Direct}}(H, R)$:
($user$) : « You will receive two sentences A and B. Do these two sentences mean the same thing? Answer with only one word "yes" or "no". »
($assistant$) : « Please provide the sentences for me to evaluate. »
($user$) : « A : "(R)" ; B : "(H)" »

Qiao *et al.* (2023) relèvent le fait que les performances d'un *LLM* sont améliorées si l'on utilise des étapes intermédiaires lors de la génération. Cela permet de simuler un cheminement de pensée (Wei *et al.*, 2022). De part la nature auto-régressive des *LLM*, ajouter des informations dans le patron aide à la résolution de la tâche. Nous proposons les deux patrons détaillés ci-dessous utilisant la génération d'une explication intermédiaire (E) avant la classification de relation entre deux phrases.

Patron$_{\text{Expliqué}}(H, R)$
($user$) : « You will receive two sentences A and B. Do these two sentences mean the same thing? »
($assistant$) : « Please provide the sentences for me to evaluate. »
($user$) : « A : "(R)" ; B : "(H)" »
($assistant$) : « (E) »
($user$) : « Summarize your answer with only one word "yes" or "no". »

De peur que le système ne favorise la catégorisation en paraphrase, nous proposons un patron demandant si les phrases ont exactement le même sens.

Patron$_{\text{Exact}}(H, R)$
($user$) : « You will receive two sentences A and B. Do these two sentences mean **exactly** the same

4. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

thing? »
($assistant$) : « Please provide the sentences for me to evaluate. »
($user$) : « A : "(R)" ; B : "(H)" »
($assistant$) : « (E) »
($user$) : « Summarize your answer with only one word "yes" or "no". »

Corpus	Patron	VP	VN	FP	FN	Exactitude	F1	Rappel	Précision
MRPC	Direct	3 550	956	945	350	0,78	0,84	0,91	0,79
MRPC	Expliqué	3 089	1 343	558	811	0,76	0,82	0,79	0,85
MRPC	Exact	1 965	1 661	240	1 935	0,62	0,64	0,50	0,89
PAWS	Direct	3 381	1 297	3 164	158	0,58	0,67	0,95	0,52
PAWS	Expliqué	3 177	1 919	2 542	362	0,64	0,68	0,90	0,55
PAWS	Exact	2 453	3 074	1 387	1 086	0,69	0,66	0,69	0,64
Total	Direct	6 931	2 253	4 109	508	0,66	0,75	0,93	0,63
Total	Expliqué	6 266	3 262	3 100	1 173	0,69	0,74	0,84	0,67
Total	Exact	4 418	4 735	1 627	3 021	0,66	0,65	0,59	0,73

TABLE 2 – Performance du classifieur par *LLM* en fonction du patron utilisé. Le comptage des résultats est donné en tant que Vrai Positif (VP), Vrai Négatif (VN), Faux Positif (FP) et Faux Négatif (FN).

Les résultats de classification de ces approches sur les corpus PAWS et MRPC, sont rapportés dans la table 2. L'utilisation du patron direct donne une exactitude (*accuracy*) supérieure au hasard (0,66 contre 0,54). Comme supposé, l'ajout de l'explication permet d'améliorer les performances de l'approche directe. En revanche, imposer une relation d'équivalence sémantique exacte dégrade les résultats. On notera qu'une variation, même minime de la formulation du patron, peut influencer grandement les résultats.

Cette expérience montre la capacité du *LLM* à discriminer les paraphrases des non-paraphrases, malgré un contexte difficile. En revanche, compte-tenu du caractère continu de la relation de paraphrase, il semble plus judicieux de travailler avec un degré de proximité sémantique plutôt qu'une classification binaire. L'usage d'un *LLM* semble donc pertinent pour construire une mesure plus performante que celles traditionnellement utilisées.

5 Proposition : ParaPLUIE

Pour rappel, les modèles de langues sont avant tout une modélisation des probabilités d'apparition d'un symbole textuel, sachant un historique. Il est donc possible de comparer deux séquences pour calculer un degré d'appartenance à une classe comme Chen *et al.* (2023). Ainsi, nous proposons ParaPLUIE (*ParaPhrase, Llm Used for Improved Evaluation*), une mesure de proximité sémantique reposant le modèle probabiliste d'un *LLM*. ParaPLUIE est définie comme le logarithme des rapports de vraisemblances, sur le fait que le patron appliqué à la paraphrase hypothétique (H) et à la référence (R) est suivi du symbole « yes » ou « no », c'est-à-dire :

$$\text{ParaPLUIE}(H, R) \quad = \quad \log \left(\frac{p\left(\text{yes}|\text{Patron}(H, R)\right)}{p\left(\text{no}|\text{Patron}(H, R)\right)} \right)$$

Si les patrons sont identiques et les mots « yes » et « no » ne sont codés que sur un unique symbole, alors ce ratio de probabilité est égale au ratio des perplexités (ppl), à une puissance près. La perplexité reflétant justement la « surprise » du modèle lors de l'apparition des symboles. De plus, généralement, les *LLM* sont justement appris en utilisant la perplexité comme fonction objectif ($loss$). Ainsi, le calcul de la mesure devient :

$$\begin{aligned} \text{ParaPLUIE}(H,R) &= \log\left(\frac{ppl\left(\text{Patron}(H,R)\circ\text{no}\right)^{T+1}}{ppl\left(\text{Patron}(H,R)\circ\text{yes}\right)^{T+1}}\right) \\ &= (T+1)\times\left[loss_{LLM}(\text{Patron}(H,R)\circ\text{no}) - loss_{LLM}(\text{Patron}(H,R)\circ\text{yes})\right] \end{aligned} \tag{1}$$

où T est le nombre de symboles dans le patron et « $\circ$ » l'opération de concaténation de deux textes.

La mesure proposée est donc à valeurs réelles. Plus le score est élevé et plus le système estime que les deux phrases sont vraisemblablement des paraphrases alors qu'un score inférieur à zéro indiquerait que le sens des deux phrases est différent. Notons que cette propriété aide à l'interprétation des résultats contrairement à d'autres scores.

Comme pour la section 4, nous utilisons le *LLM* Mistral pour ParaPLUIE. Puisque les résultats de la table 2 sont relativement proches sur la tâche de classification, et par soucis de simplicité, nous utilisons le $\textbf{Patron}_{\text{Direct}}$.

6 Dynamique des scores

Nous nous intéressons ici à la dynamique des scores ParaPLUIE au regard des étiquettes paraphrases/non-paraphrases fournis dans les corpus de la section 3.

La distribution des scores ParaPLUIE sur les deux corpus de référence est présentée table 3. Contrairement à ce qui a été observé dans la table 1, le score moyen des paraphrases est bien supérieur à celui des non-paraphrases en intra-corpus et en inter-corpus. On regrettera que le score moyen des non-paraphrases ne soit pas négatif. Encore une fois, cela peut s'expliquer par le caractère volontairement trompeur des corpus considérés dans cette expérience.

Corpus	*Para.*	ParaPLUIE ↑
MRPC	Oui	20,02 ±8,94
	Non	4,41 ±15,43
PAWS	Oui	22,04 ±6,64
	Non	12,80 ±13,46
Total	Oui	20,98 ±7,99
	Non	10,29 ±14,59

TABLE 3 – Moyenne et écart-type des scores ParaPLUIE sur les corpus MRPC et PAWS.

La figure 1 propose une comparaison graphique des distributions des scores en fonction de l'étiquette des phrases. Puisqu'aucune mesure ne classe parfaitement les corpus, on observe un chevauchement

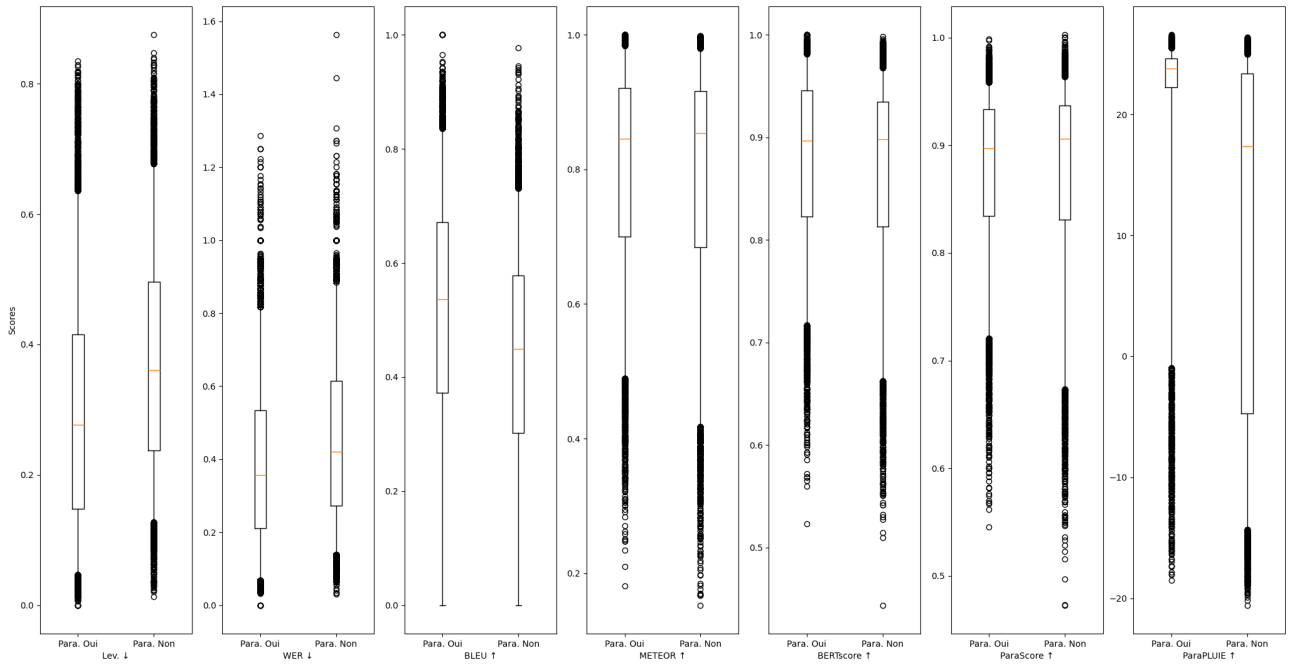


FIGURE 1 – Boîte à moustache des scores pour les différentes mesures. Les cercles correspondent aux 5% des valeurs les plus extrêmes.

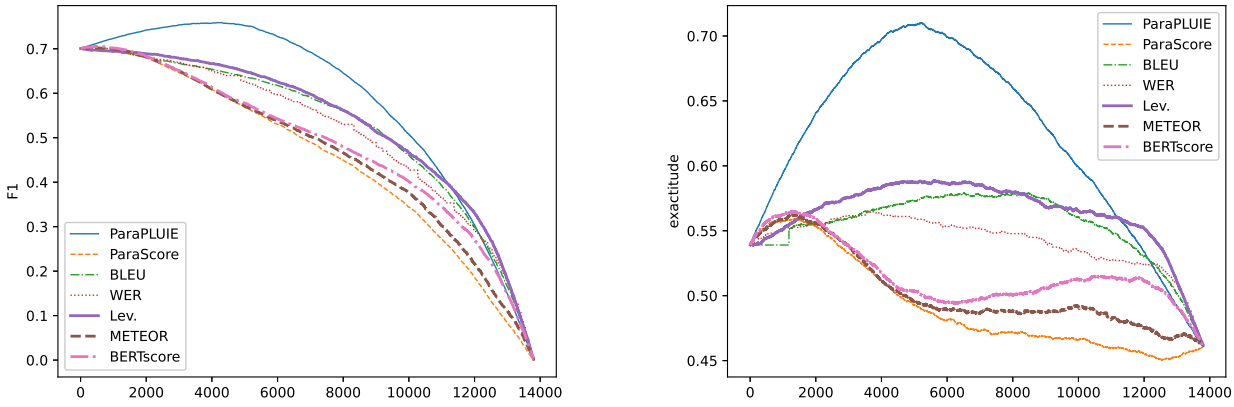
des scores. Mais pour ParaPLUIE, contrairement aux autres mesures, le score des non-paraphrases semble avoir une dynamique bien différente du score des paraphrases. Notons aussi que les mesures lexicale semble meilleurs sur ces corpus que les mesures sémantiques, hormis ParaPLUIE.

7 Re-classement de paraphrases

Comparons la capacité qu'ont les différentes mesures à ordonner les paraphrases et les non-paraphrases des corpus. Les paires de phrases sont triées par score décroissant pour LEV. et WER, et par score croissant pour les autres. Ainsi, plus une paire de phrases a un rang élevé, plus ces dernières sont considérées comme proches l'une de l'autre par la mesure. En faisant varier le rang à partir duquel on considère qu'une paire est effectivement paraphrase, et sachant l'étiquette d'une paire, il est possible de calculer la variation du score F1 (moyenne harmonique du rappel et de la précision), ainsi que l'exactitude de classification. Ces résultats sont présentés en figure 2.

Comme l'illustre la figure 2a, le score F1 de toutes les mesures chute rapidement, sauf celui de ParaPLUIE. On remarque figure 2b que l'exactitude reste basse, sauf pour ParaPLUIE, et qu'elle chute rapidement pour METEOR, $\text{BERT}_{\text{score}}$ et ParaScore. Sur ce type de corpus, étonnamment, les mesures lexicales exactes surpassent les mesures plus sémantiques autres que ParaPLUIE. Cette dernière est meilleure que les mesures lexicales, pour la très grande majorité des rangs, et surpasse toujours celles plus sémantiques.

Si l'on se concentre sur le maximum a posteriori de score F1, on constate qu'à l'exception de ParaPLUIE, l'optimal consiste approximativement à classifier toutes les paires comme paraphrases. Pour le maximum d'exactitude a posteriori, avec 0,71, ParaPLUIE est sensiblement meilleur que



(a) F1 en fonction du rang.

(b) Exactitude en fonction du rang.

FIGURE 2 – Évolution de l'exactitude et de la F1 des différentes mesures en fonction du rang.

les autres mesures, comprises entre 0,56 et 0,59. Ces résultats sont détaillés dans la table 4. Notons que, compte-tenu du léger déséquilibre des classes, étiqueter toutes les phrases comme paraphrases donnerait une exactitude de 0,54.

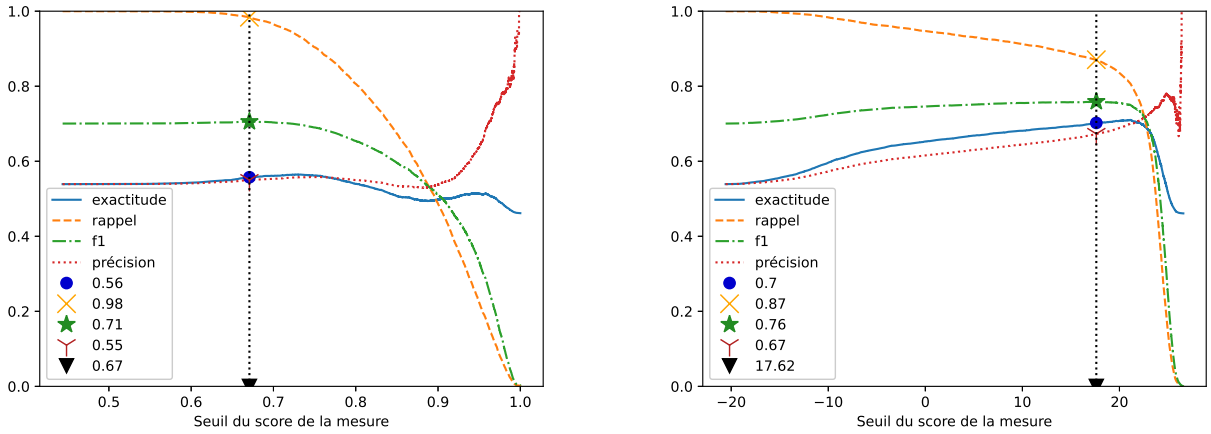
Score	LEV.	WER	BLEU	METEOR	$\text{BERT}_{\text{score}}$	ParaScore	ParaPLUIE
F1 max.	0,70	0,70	0,70	0,70	0,71	0,70	**0,76**
$\text{Rappel}_{\text{F1 max.}}$	**1,00**	**1,00**	**1,00**	0,99	0,98	0,99	0,87
$\text{Précision}_{\text{F1 max.}}$	0,54	0,54	0,54	0,55	0,55	0,55	**0,67**
Exact. max.	0,59	0,56	0,58	0,56	0,56	0,56	**0,71**

TABLE 4 – Récapitulatif des scores de rappel et de précision selon la meilleure F1 *(F1 max.)* ainsi que de l'exactitude la plus haute *(Exact. max.)* pour chaque mesure sur les deux corpus. Les valeurs les plus élevées sont en gras.

Les figures 3a et 3b permettent de comparer plus en détail les classements produits par $\text{BERT}_{\text{score}}$ et ParaPLUIE. On constate la chute précoce du rappel pour $\text{BERT}_{\text{score}}$ alors qu'elle intervient beaucoup plus tard pour ParaPLUIE. Pour la précision, celle de $\text{BERT}_{\text{score}}$ n'augmente que pour les valeurs maximales – un score très proche de 1 semble être un indicateur fiable de la relation de paraphrase – alors que celle de ParaPLUIE augmente beaucoup plus tôt. Malgré cela, on notera la présence de non-paraphrases dans les meilleurs scores de ParaPLUIE.

ParaPLUIE étant définie par une soustraction (voir l'équation 1), il existe un point de bascule entre les deux probabilités en 0. En fixant un seuil de classification sur cette valeur, l'exactitude est de 0,65. Cette performance reste supérieur à toutes les autres mesures, quelque soit le seuil choisi. Autrement dit, le seuil de classification fixé à priori pour ParaPLUIE est meilleur que toutes les autres mesures, même en fixant leurs seuils de classification à posteriori.

Ces expériences, sur des corpus complexes semblent indiquer que ParaPLUIE est une bonne mesure de proximité sémantique plus performante que l'état l'art.



(a) BERT$_{\text{score}}$ (b) ParaPLUIE

FIGURE 3 – Évolution de la qualité du tri des paraphrases/non-paraphrases en fonction de la mesure. Les valeurs mises en avant correspondent au maximum de score F1.

8 Conclusion et perspectives

ParaPLUIE est une nouvelle mesure de proximité sémantique utilisant un *LLM*. Nous avons évalué la qualité de notre mesure par rapport aux autres mesures communément utilisées. Cette évaluation a été effectuée sur deux corpus de paraphrases en anglais, annotés par des humains, reconnus pour leurs qualités ou difficultés. Notre analyse montre que ParaPLUIE obtient une meilleure exactitude que les autres mesures de l'état de l'art. Nous proposons des variantes de patrons et montrons que s'appuyer sur une étape d'explication intermédiaire générée par *LLM* améliore l'exactitude.

Cette étude a été menée sur une quantité limitée de données, dû au temps de calcul important nécessaire à l'utilisation d'un *LLM*. Dans ces expériences, nous n'avons pas réalisé de *few-shots prompting*, c'est-à-dire, ajouté un exemple de résolution de la tâche dans le patron utilisé. Nous sommes confiants dans le fait que cela améliorerait les résultats de ParaPLUIE mais nous avons souhaité éviter de créer un biais sur les corpus que nous avons étudiés.

Être capable d'estimer si deux phrases sont des paraphrases permet l'avancée sur d'autres terrains de recherche, comme la création de petit modèles de langue dédiés à la tâche de production de paraphrase, en coopération avec des méthodes d'apprentissage par distillation de connaissances (Hsieh *et al.*, 2023).

Nous souhaitons étendre cette étude sur des corpus de paraphrases de grande distance lexicale, là où cette étude s'est tournée sur un ensemble de couples avec une faible distance lexicale. Nous souhaitons également concevoir un petit modèle de langue dédié à l'identification de paraphrase.

Enfin nous appelons à ne pas utiliser ParaPLUIE avec un *LLM* de très grande taille. Intuitivement, l'utilisation d'un tel modèle améliorerait les résultats de ParaPLUIE ; néanmoins, rien ne le garantit et leur utilisation est particulièrement coûteuse. En revanche l'étude du patron utilisé semble être une piste prometteuse pour améliorer la mesure.

Remerciements

Recherche soutenue financièrement par le Ministère des Armées - Agence de l'Innovation de la Défense.

Références

BANERJEE S. & LAVIE A. (2005). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. In J. GOLDSTEIN, A. LAVIE, C.-Y. LIN & C. VOSS, Éds., *Proceedings of the Association for Computational Linguistics Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, p. 65–72, Ann Arbor, Michigan : Association for Computational Linguistics.

BARZILAY R. & MCKEOWN K. R. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, p. 50–57, Toulouse, France : Association for Computational Linguistics. DOI : 10.3115/1073012.1073020.

CHEN Y., WANG R., JIANG H., SHI S. & XU R. (2023). Exploring the use of large language models for reference-free text quality evaluation : An empirical study. In J. C. PARK, Y. ARASE, B. HU, W. LU, D. WIJAYA, A. PURWARIANTI & A. A. KRISNADHI, Éds., *Findings of the Association for Computational Linguistics : IJCNLP-AACL 2023 (Findings)*, p. 361–374, Nusa Dua, Bali : Association for Computational Linguistics.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DOLAN W. B. & BROCKETT C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

FABRE B., URVOY T., CHEVELU J. & LOLIVE D. (2021). Neural-driven search-based paraphrase generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 2100–2111, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.180.

HSIEH C.-Y., LI C.-L., YEH C.-K., NAKHOST H., FUJII Y., RATNER A., KRISHNA R., LEE C.-Y. & PFISTER T. (2023). Distilling step-by-step ! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics : Association for Computational Linguistics 2023*, p. 8003–8017, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.507.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., CASAS D. D. L., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L. *et al.* (2023). Mistral 7b. *arXiv preprint arXiv :2310.06825*.

LEEMAN D. (1973). La paraphrase. *Langages*, p. 43–54.

LEVENSHTEIN V. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Dokl. Akad. Nauk SSSR*, **163**, 845–848.

MEL'ČUK I. M. (1997). *Vers une linguistique sens-texte : leçon inaugurale faite le vendredi 10 janvier 1997*. Collège de France.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, p. 311–318.

QIAO S., OU Y., ZHANG N., CHEN X., YAO Y., DENG S., TAN C., HUANG F. & CHEN H. (2023). Reasoning with language model prompting : A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 5368–5393, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.294.

SEKINE S. (2005). Automatic paraphrase discovery based on context and keywords between NE pairs. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

SHEN L., LIU L., JIANG H. & SHI S. (2022). On the evaluation metrics for paraphrase generation. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 3178–3190, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.208.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

WEI J., WANG X., SCHUURMANS D., BOSMA M., XIA F., CHI E., LE Q. V., ZHOU D. *et al.* (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, **35**, 24824–24837.

WOODARD J. & NELSON J. (1982). An information theoretic measure of speech recognition performance. In *Workshop on standardisation for speech I/O technology, Naval Air Development Center, Warminster, PA*.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). Bertscore : Evaluating text generation with bert. In *International Conference on Learning Representations*.

ZHANG Y., BALDRIDGE J. & HE L. (2019). PAWS : Paraphrase Adversaries from Word Scrambling. In *Proceedings of North American Chapter of the Association for Computational Linguistics*.

ZHAO S., LAN X., LIU T. & LI S. (2009). Application-driven statistical paraphrase generation. In K.-Y. SU, J. SU, J. WIEBE & H. LI, Éds., *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, p. 834–842, Suntec, Singapore : Association for Computational Linguistics.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Prédiction de la complexité lexicale : Une étude comparative entre ChatGPT et un modèle dédié à cette tâche.

Abdelhak Kelious[1] Mathieu Constant[1] Christophe Coeur[2]
(1) Université de Lorraine et CNRS/ATILF, France
(2) Consultant, France
(1) (abdelhak.kelious,mathieu.constant)@univ-lorraine.fr,
christophe.coeur@gmail.com

RÉSUMÉ

Cette étude s'intéresse à la prédiction de la complexité lexicale. Nous explorons des méthodes d'apprentissage profond afin d'évaluer la complexité d'un mot en se basant sur son contexte. Plus spécifiquement, nous examinons comment utiliser des modèles de langue pré-entraînés pour encoder le mot cible et son contexte, en les combinant avec des caractéristiques supplémentaires basées sur la fréquence. Notre approche obtient de meilleurs résultats que les meilleurs systèmes de SemEval-2021 (Shardlow *et al.*, 2021) sur cette tâche. Enfin, nous menons une étude comparative avec ChatGPT afin d'évaluer son potentiel pour prédire la complexité lexicale en comparaison avec un modèle dédié à cette tâche.

ABSTRACT

Lexical Complexity Prediction : a Comparative Study Between ChatGPT and a Dedicated Model for this Task

This study focuses on lexical complexity prediction. We explore deep learning methods to assess the complexity of a word based on its context. Specifically, we investigate how to use pre-trained language models to encode both the sentence and the target word, and then fine-tune them by combining them with additional frequency-based features. Our approach outperforms the best systems in SemEval-2021 (Shardlow *et al.*, 2021). Finally, we conduct a comparative study with ChatGPT to assess its potential for predicting lexical complexity compared to a model dedicated to this task.

MOTS-CLÉS : Traitement du langage naturel, Prédiction de la complexité lexicale, Modèles linguistiques, ChatGPT.

KEYWORDS: Natural language processing, Lexical complexity prediction, Language models, ChatGPT.

1 Introduction

Comprendre le langage est un défi qui mobilise de nombreuses compétences linguistiques, dont la maîtrise de la grammaire, l'enrichissement du vocabulaire et la production de discours cohérents. La complexité lexicale, influencée par la fréquence et le contexte d'utilisation des mots, joue un rôle crucial dans l'apprentissage des langues, affectant la rapidité et l'aisance avec lesquelles les apprenants maîtrisent de nouvelles compétences. Les difficultés de compréhension surgissent notamment quand des mots apparaissent dans des contextes peu familiers, poussant le lecteur à mal interpréter, ignorer

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 617–629.

ou perdre le fil du texte. Le traitement automatique des langues offre des outils pour identifier ces mots complexes.

La prédiction de la complexité lexicale constitue un champ de recherche dédié à estimer la difficulté des mots dans un contexte spécifique. Cette difficulté est mesurée sur une échelle continue entre 0 et 1. Les études dans ce domaine ont exploité des caractéristiques lexicales comme la fréquence des mots ou la longueur des phrases (Zampieri *et al.*, 2016) et des modèles de langue avancés tels que BERT (Devlin *et al.*, 2019). Cette tâche a des applications pratiques notables, surtout en didactique des langues et en compréhension de texte (Alfter, 2021).

Notre recherche s'inscrit dans la continuité des efforts récents, notamment ceux de SemEval-2021, et propose une méthode associant des modèles de langue pré-entraînés comme DeBERTa (He *et al.*, 2023) à des analyses de fréquence textuelle pour prédire la complexité des mots. Par ailleurs, nous examinons la capacité des modèles de langue de grande taille (LLMs) génératifs, tels que ChatGPT, à réaliser cette tâche par rapport à un modèle spécifiquement appris pour la prédiction de la complexité lexicale.

Les principales contributions de cet article sont les suivantes :

— Un nouveau modèle de prédiction de complexité lexicale entraîné sur le jeu de données Complex 2.0 (Shardlow *et al.*, 2021), combinant des modèles de langue pré-entraînés avec des caractéristiques de fréquence. Ce modèle obtient des performances dépassant celles de la compétition SemEval 2021 avec ces même données.
— Une évaluation comparative de la capacité de ChatGPT pour notre tâche. Les résultats montrent qu'il n'est pas performant pour prédire un score de complexité, mais il présente une certaine capacité à classer les contextes selon leur complexité, en particulier lorsque les contextes sont clairement difficiles ou clairement faciles.

2 Travaux connexes

2.1 Identification des mots complexes : méthodes et ensembles de données

Il existe plusieurs raisons d'évaluer l'identification des mots complexes dans une phrase (contexte), et cela fait l'objet de recherches depuis plusieurs années. Par exemple, cela a été exploré dans le contexte de la simplification lexicale, qui consiste à remplacer automatiquement les mots complexes par des alternatives plus simples (Shardlow, 2013).

La tâche d'identification des mots complexes (CWI : Complex Word Identification) a été étudiée en tant que tâche d'annotation de séquences, prenant en compte le contexte (phrase) dans lequel le mot apparaît (Gooding & Kochmar, 2019). Cette tâche a particulièrement été mise en avant lors de la compétition SemEval en 2021 *Lexical Complexity Prediction* (Shardlow *et al.*, 2021), ce qui a permis de mettre en place un jeu de données dédié de référence ainsi que de standardiser des métriques d'évaluation. Cette compétition a mis en évidence l'impact de l'utilisation de modèles de langue pré-entraînés pour la CWI, ainsi que l'utilisation de diverses techniques d'affinage (fine-tuning). Par exemple, l'un des meilleurs systèmes, JUST BLUE (Bani Yaseen *et al.*, 2021), combine les modèles BERT et RoBERTa (Liu *et al.*, 2019). Les modèles ont été affinés séparément pour prédire les scores de complexité, le score final étant leur moyenne. Un autre exemple est le système DeepBlueAI (Pan *et al.*, 2021) qui intègre des modèles de langue pré-entraînés affinés avec des techniques telles que le

pseudo-étiquetage, l'augmentation de données, les modèles d'entraînement empilés et l'utilisation d'une couche de *dropout* multi-échantillon.

Il existe divers jeux de données pour l'identification des mots complexes. Tout d'abord, CWI-2016 (Paetzold & Specia, 2016) est un jeu de données pour l'analyse de la complexité des mots en anglais en contexte, annoté par des locuteurs non natifs. Ce jeu de données a deux versions : une première version donnant pour chaque instance (mot-cible dans un contexte donné) les étiquettes binaires des 20 annotateurs ou annotatrices (1 si elles ou ils la jugeaient complexe, 0 sinon) ; une deuxième version attribue une seule étiquette à chaque instance, 1 si au moins un des 20 annotateurs ou annotatrices la jugeait complexe, sinon 0. CWI-2018 (Yimam *et al.*, 2018) est un jeu de données pour l'analyse de la complexité des mots dans plusieurs langues (anglais, espagnol, allemand et français uniquement pour les tests pour ce dernier), annoté par des locuteurs natifs et non natifs. Le jeu de données fournit deux types d'étiquetage : une étiquette binaire (0 ou 1) si l'instance est considérée comme facile ou difficile par les annotateurs ; une valeur réelle entre 0 et 1 indiquant la proportion d'annotateurs trouvant l'instance difficile. Dans cet article, nous utilisons un jeu de données plus récent Complex 2.0 (Shardlow *et al.*, 2022). Ces données ont été annotées manuellement en anglais en utilisant une échelle de Likert à 5 points pour indiquer le degré de complexité. Chaque instance (mot-cible dans un contexte donné) est ainsi annotée par une valeur réelle étant la moyenne des scores likert donnés par les différents annotateurs ou annotatrices normalisée entre 0 et 1. Complex 2.0 représente une amélioration significative par rapport à la version précédente, Complex 1.0 (Shardlow *et al.*, 2020) avec, notamment, un nombre d'annotateurs supplémentaires. Ce jeu de données offre ainsi un réglage plus précis pour identifier la complexité des mots.

La tâche de la prédiction de la complexité lexicale s'apparente d'une certaine manière à la tâche de prédiction automatique de la lisibilité d'un texte qui utilise des techniques proches. Les textes peuvent être simplifiés en fonction des scores de lisibilité, rendant l'information plus accessible (Wilkens *et al.*, 2022) (Wilkens *et al.*, 2024). Des recherches récentes indiquent que l'emploi de modèles de langue avancés, tels que ceux basés sur des *transformers*, combiné à l'application de caractéristiques linguistiques spécifiquement choisies, permet d'affiner l'évaluation de la lisibilité (Lee *et al.*, 2021). Cependant, bien que ces caractéristiques linguistiques puissent améliorer les résultats sur des échantillons de taille réduite, leur impact reste variable sur les modèles d'apprentissage profond plus sophistiqués, ne se traduisant pas systématiquement par un avantage marqué (Deutsch *et al.*, 2020).

2.2 ChatGPT

Avec l'avènement des modèles génératifs, ChatGPT notamment, il est difficile d'approcher notre étude sans inclure une comparaison par rapport à ce type de modèle. Des études récentes ont démontré le potentiel prometteur de ChatGPT [1] pour diverses tâches d'annotation (Dai *et al.*, 2023; Kuzman *et al.*, 2023) et de classification de texte (Liu *et al.*, 2023; Amin *et al.*, 2023; Zhang *et al.*, 2022). Parmi les tâches qui ont suscité l'intérêt de la communauté scientifique, il y a l'augmentation de données, et l'une des techniques d'augmentation de données utilisées par ChatGPT est la paraphrase. En reformulant le texte d'entrée de diverses manières, le modèle peut générer un ensemble plus diversifié d'exemples pour l'entraînement. Cela aide le modèle à saisir le sens sous-jacent du texte plutôt que de simplement mémoriser des phrases ou des motifs spécifiques. Par exemple, AugGPT (Dai *et al.*, 2023) reformule les phrases d'entraînement en plusieurs échantillons similaires mais sémantiquement différents, améliorant ainsi les performances du modèle. L'étude montre que AugGPT

1. Malgré un potentiel intéressant, ChatGPT a de nombreuses limites (Ray, 2023) que nous évoquerons au fil de l'article.

améliore significativement les performances du modèle. Huang *et al.* (2023) comparent ChatGPT aux annotateurs humains, révélant que ChatGPT peut détecter efficacement les tweets haineux avec une précision de 80%. Les désaccords restants sont généralement sujets à débat, mais les résultats montrent que les évaluations humaines tendent à soutenir les classifications de ChatGPT. L'étude montre aussi que ChatGPT produit des explications en langage naturel comparables à celles des humains en termes d'informativité et de clarté.

3 Un modèle dédié à l'identification des mots complexes

3.1 Données

Nous avons utilisé le jeu de données "CompLex 2.0" (Shardlow *et al.*, 2022) lors des phases d'entraînement et de test. Le corpus comprend des évaluations humaines de la complexité lexicale pour un ensemble de textes anglais, utilisant une échelle de Likert à 5 points, provenant de sources telles que Wikipedia, des livres éducatifs et des articles de journaux, couvrant divers sujets. Les textes ont été annotés par des évaluateurs humains qui ont évalué la complexité lexicale du mot cible dans son contexte (phrase) à l'aide de l'échelle de Likert. Chaque instance a été annotée plusieurs fois, et la moyenne de ces annotations a été prise comme score pour chaque instance de données. Ce score est normalisé en une valeur continue entre 0 et 1. La taille des données d'entraînement est de 7662 et de 917 pour le test.

3.2 Modèle

Dans notre étude, nous avons développé un modèle de réseau neuronal pour prédire la complexité des mots en utilisant des plongements lexicaux des mots basés sur des *transformers (encodeur)* et des caractéristiques telles que la fréquence du mot cible et les mots qui composent la phrase. Pour affiner les prédictions, nous avons intégré la loi de Zipf via la bibliothèque *wordfreq* (Speer, 2022), exploitant la fréquence d'occurrence des mots dans plus de 40 langues. Nous avons défini plusieurs caractéristiques d'entrée : F1 (le score Zipf de la fréquence des mots), F2 (le score Zipf moyen dans une phrase), F3 (la différence entre le score Zipf du mot cible et le score moyen), F4 (le nombre de mots avec un score Zipf supérieur au mot cible) et F5 (une valeur binaire indiquant si le mot cible est considéré comme rare avec un score inférieur ou égal 3). Ces caractéristiques sont combinées à des couches cachées pour capter la non-linéarité des données.

La formule de prédiction de complexité s'exprime comme suit dans notre modèle :

$$\hat{y} = f(W_h \cdot \sigma(W_e \cdot E + W_f \cdot F + b_e) + b_h)$$

où :

- E représente les plongements lexicaux des mots basés sur des *transformers*,
- F est le vecteur des caractéristiques d'entrée $[F_1, F_2, F_3, F_4, F_5]$,
- W_e et W_f sont les poids appliqués respectivement aux plongements lexicaux et aux caractéristiques d'entrée.
- b_e et b_h sont les biais pour les couches d'entrée et cachées, respectivement.
- σ est une fonction d'activation non-linéaire (ReLU) , appliquée à la combinaison linéaire des plongements lexicaux et des caractéristiques d'entrée.

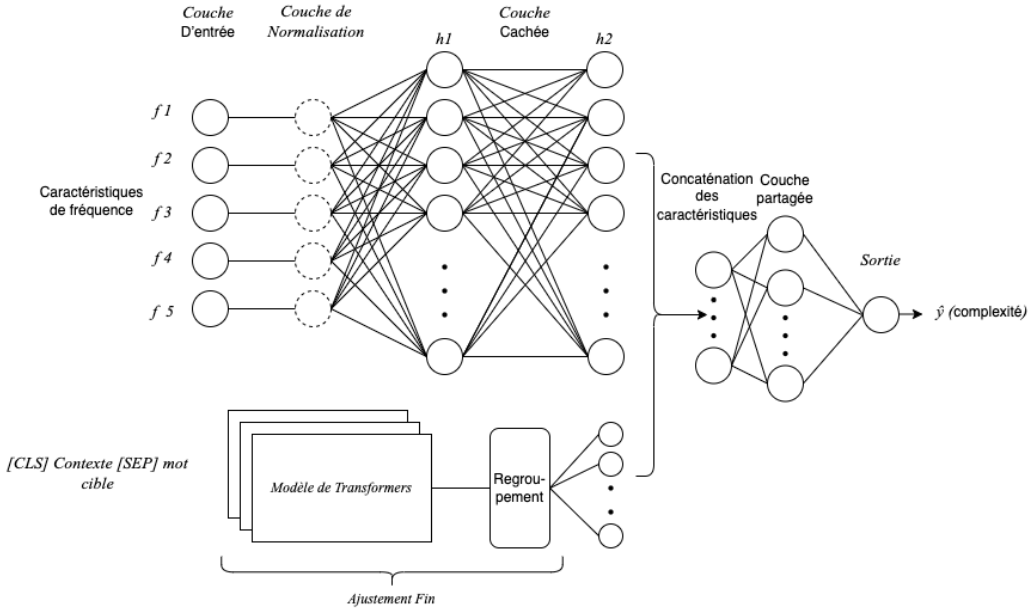


FIGURE 1 – L'architecture globale pour prédire les scores de complexité

— W_h sont les poids de la couche cachée.
— f est la fonction linéaire d'activation de la couche de sortie.
— $\hat{y}$ correspond à la valeur de prédiction (valeur réelle entre 0 et 1).

3.3 Évaluation

Comme le montre la figure 1, en combinant les différentes caractéristiques, nous avons utilisé le modèle pour faire des prédictions sur la complexité des mots. Nous avons mené nos expériences pour quatre modèles de plongements de mots : bert-base, DeBERTa-base-v3, ainsi que leurs versions large et multilingues. Pour évaluer les résultats, nous utilisons les mêmes métriques de corrélation que celles utilisées dans Shardlow *et al.* (2021) : Pearson, Spearman et le score R^2. Nous avons conservé les mêmes hyperparamètres pour chaque entraînement, incluant un taux d'apprentissage de 5e-5, une taille de lot de 4 et une longueur de séquence maximale de 300. Le tableau 1 montre les résultats de notre approche en utilisant les modèles de langue cités. Nous incluons également le meilleur score obtenu lors de la tâche partagée SemEval-2021 (Shardlow *et al.*, 2021). Le modèle de base fournie par les organisateurs de la tâche est reproduite en utilisant la log-fréquence du Google Web1T (Evert, 2010) et la régression linéaire. Nous constatons une amélioration significative des performances en utilisant le modèle de langue DeBERTa-large-v3, atteignant un score R^2 de **0.65**, supérieur aux performances obtenus dans la compétition de SemEval 2021. Il est à noter que la comparaison n'est pas totalement juste par rapport aux systèmes de SemEval 2021 puisque nous avons utilisé des modèles de langue plus récents (DeBERTa).

Afin de comprendre l'impact des modèles de langue pré-entraînés et des caractéristiques de fréquence dans chacun des modèles, nous avons mené une étude d'ablation pour mesurer l'importance de chaque composant. La figure 2 est une représentation graphique qui montre la performance des modèles pour prédire la complexité lexicale. Chaque colonne représente un modèle différent, avec différentes couleurs pour distinguer les types de modèles. La barre bleue représente un modèle

Modèles	Pearson	Spearman	R^2
Deberta-v3-large	**0,81**	**0,74**	**0,65**
Deberta-v3-base	0,79	0,74	0,62
Le score le plus élevé dans SemEval 2021 (Bani Yaseen et al., 2021)	*0,78*	-	*0,61*
mDeberta-v3-base	0,75	0,70	0,57
bert-base-cased	0,74	0,70	0,55
bert-base-multilingue	0,67	0,64	0,45
Baseline de fréquence fournie par (Shardlow *et al.*, 2021)	0,52	-	0,27

TABLE 1 – Résultats avec différents modèles de langues

qui utilise uniquement les caractéristiques de fréquence, la barre rouge représente un modèle basé uniquement sur des modèles de langue, tandis que la verte utilise à la fois les caractéristiques de fréquence et les modèles de langue. Les scores de corrélation de Pearson sont utilisés pour évaluer la performance des modèles en comparant leurs prédictions aux données de référence. Les résultats indiquent qu'ajouter des caractéristiques de fréquence aux modèles de langue améliore légèrement la prédiction de la complexité lexicale. En d'autres termes, les modèles qui utilisent à la fois des caractéristiques de fréquence et des modèles de langue sont meilleurs pour prédire la complexité lexicale que les modèles qui utilisent seulement l'un ou l'autre. On notera que notre modèle basé uniquement sur les caractéristiques de fréquence obtient de meilleurs résultats que la baseline dans le tableau 1 avec un score de corrélation de Pearson de 0,61.

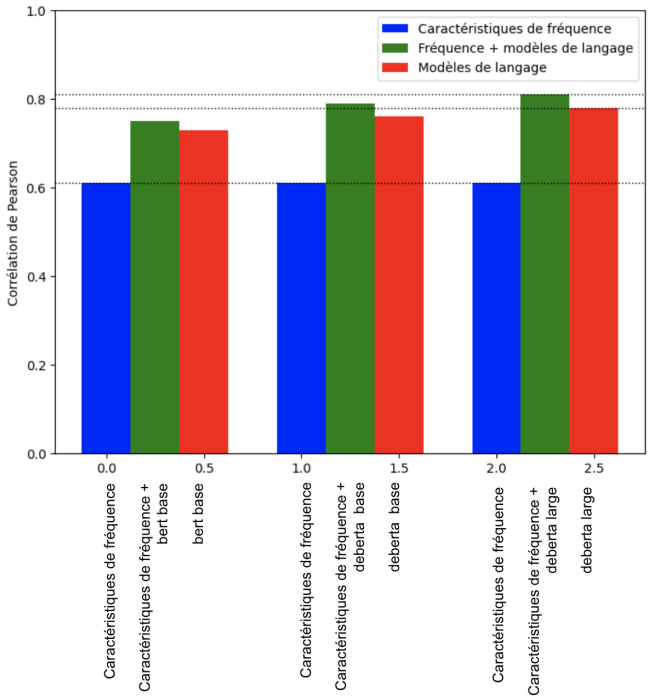


FIGURE 2 – Comparaison des valeurs de corrélation de Pearson pour les différents modèles.

4 Évaluation comparative avec ChatGPT

L'un des objectifs de cet article vise à comparer ChatGPT avec un modèle dédié pour mesurer la complexité lexicale, afin de voir si ChatGPT représente une alternative viable. Pour cette expérience, nous avons utilisé ChatGPT Turbo 3.5 (le 4 octobre 2023) via l'API fournie par OpenAI. Soulignons néanmoins une limite importante de notre expérience : ChatGPT est régulièrement mis à jour, ce qui pose des problèmes de reproductibilité (Chen *et al.*, 2023). Par ailleurs, il est difficile d'interpréter en profondeur les résultats obtenus puisque ChatGPT fonctionne comme une "boîte noire". Nous nous limiterons donc à des observations de surface des résultats.

4.1 Méthodologie de comparaison

Notre idée initiale était de fournir à ChatGPT une instance en entrée (un mot-cible et son contexte d'apparition) et de lui demander d'évaluer sa complexité, sur une échelle de 0 à 1. Cependant, nous avons rapidement constaté des résultats très médiocres, le score de corrélation de Pearson entre les évaluations humaines et ChatGPT étant de l'ordre de 0,034. Comme ChatGPT n'a pas été spécifiquement entraîné pour la tâche de prédiction de la complexité lexicale sur notre ensemble de données spécifique, une telle comparaison directe est quelque peu injuste. De plus, ChatGPT n'a pas accès à des informations complètes sur la façon dont les humains ont évalué ces données. Même avec une consigne sophistiquée, la tâche reste très complexe. Nous devons donc rendre l'évaluation plus équitable. Au lieu d'évaluer la capacité de ChatGPT à prédire un score de complexité pour une instance donnée (mot cible + contexte), nous évaluons sa capacité à comparer deux instances selon leur complexité, et ainsi à classer un ensemble d'instances selon celle-ci. En d'autres termes, nous éliminons la nécessité de prédire des scores entre 0 et 1, permettant à ChatGPT de se concentrer uniquement sur l'évaluation de l'ordre relatif de complexité parmi les instances. Pour ce faire, nous avons utilisé l'algorithme de tri à bulles pour trier une liste d'instances, où la comparaison entre deux instances est effectuée par ChatGPT [2]. Pour chaque paire d'instances à comparer, nous avons utilisé le *prompt* suivant :

> """
>
> I give you two sentences, evaluate the complexity of the target word in quotes based on its context, and return only the sentence or the target word that is simpler to understand. The output format should be as follows :
>
> {'simplest sentence' : sentence}
>
> The two sentences are :
>
> sentence 1
>
> sentence 2
>
> """

Ce prompt a été produit par essais-erreurs successifs [3]. La première difficulté a été de créer un prompt

2. Notons que pour un ensemble d'instances de taille importante, cet algorithme peut avoir un impact écologique non négligeable du fait d'un appel potentiel à ChatGPT pour chaque paire d'instances.

3. Le "prompt engineering" ne repose pas sur une méthodologie systématique, ce qui est une limite claire de cette approche.

qui produit de manière stable la sortie désirée. Le paramètre de température a été fixé à 0 pour éviter de favoriser la créativité pouvant entraîner des effets indésirables (Peng *et al.*, 2023).

Pour rendre la sortie pleinement comparable avec nos données de référence et notre approche décrite dans la section 1, la liste des instances est triée selon leurs scores de référence et leurs scores prédits. Pour l'évaluation, nous utilisons le *Tau de Kendall* (également connu sous le nom de *coefficient de corrélation de rang de Kendall*), qui est une mesure statistique utilisée pour quantifier la similarité entre deux classements. Il évalue la correspondance ou l'accord entre les classements du même ensemble d'éléments dans deux listes différentes. Le Tau de Kendall est fréquemment utilisé lorsqu'on travaille avec des données ordinales, où le classement ou l'ordre des éléments est pertinent. Nous emploierons cette métrique lors de la comparaison avec ChatGPT.

Nous divisons notre évaluation en deux expériences. Dans la première expérience, pour chaque mot cible, nous trions la liste des instances où le mot apparaît, l'objectif de cette expérience est d'évaluer la capacité de ChatGPT à prédire la complexité d'un même mot dans différents contextes. Dans la deuxième expérience, nous trions des listes d'instances choisies au hasard en variant la taille de l'échantillon, l'objectif cherche à déterminer si ChatGPT est capable de distinguer les niveaux de difficulté lorsque des mots différents sont utilisés dans des contextes variés.

4.1.1 Tri des contextes par mot cible

Dans la première expérience, pour chaque mot cible, la liste des instances où le mot apparaît est triée. Sur 917 instances, nous ne conservons que celles dont le mot-cible est le mot-cible d'au moins deux instances, réduisant l'ensemble à 685 instances. Les mots-cibles avec un seul contexte sont exclus pour éviter un biais dans la comparaison, car elles entraîneraient un taux de correspondance de 100%.

L'évaluation pour chaque mot cible de notre modèle et de ChatGPT repose sur le calcul du tau de Kendall, qui compare les classements obtenus à partir de notre modèle ou de ChatGPT avec ceux basés sur des annotations humaines. Le score global d'un système est dérivé de la moyenne des scores de tau de Kendall pour l'ensemble des mots cibles du jeu de données. Cette méthode révèle que, même sans entraînement spécifique sur le jeu de données, ChatGPT montre des performances supérieures à notre approche (cf. Table 2). Notons néanmoins que les performances de ChatGPT et de notre approche ont tendance à se rapprocher quand le nombre d'instances par mot-cible augmente.

Modèles	Score de Tau de Kendall
ChatGPT	**0.61**
Notre approche	0.52

TABLE 2 – Résultats des scores basés sur les classements.

4.1.2 Tri des instances échantillonnées

Dans cette partie, nous évaluons le classement des instances échantillonnées aléatoirement dans le jeu de données de test. Cela signifie que les listes considérées peuvent inclure des contextes avec divers mots cibles. L'idée est d'évaluer la capacité des différents systèmes à gérer la complexité des mots en général, indépendamment d'un mot cible donné. Dans notre expérience, nous varions la taille de l'échantillon pour évaluer son impact sur les performances de classement. Nous allons réaliser les

expériences pour 5 tailles d'échantillon (n=4, 5, 8, 10, 20). Pour chaque taille d'échantillon, nous effectuons 10 tirages aléatoires séparés pour obtenir des mesures plus robustes et améliorer notre évaluation. Nous calculons la moyenne et l'écart type de ces dix tirages pour obtenir une estimation plus précise.

n	Moyenne		Écart Type	
	ChatGPT	Notre modèle	ChatGPT	Notre modèle
4	0.145	**0.491**	0.566	0.391
5	0.011	**0.152**	0.391	0.288
8	0.062	**0.376**	0.275	0.281
10	-0.078	**0.153**	0.236	0.324
20	-0.079	**0.415**	0.065	0.095

TABLE 3 – Moyennes et écarts-types pour différentes tailles d'échantillons (100% aléatoire)

La Table 3 montre que notre modèle obtient systématiquement de meilleures performances par rapport à ChatGPT quelle que soit la taille de l'échantillon n. Par exemple, avec $n = 20$, notre modèle atteint un score de Kendall moyen de 0,415, tandis que ChatGPT obtient -0,079, ce qui indique une différence significative. De plus, à mesure que la taille de l'échantillon augmente, l'écart-type diminue principalement en raison de la réduction de la variabilité résultant d'un plus grand nombre d'instances, ce qui rend la moyenne une estimation plus précise de la tendance centrale.

Pourquoi cette tâche semble-t-elle être difficile pour ChatGPT ? L'une des hypothèses que nous voulons vérifier est que ChatGPT pourrait avoir des difficultés à distinguer les degrés de complexité lorsqu'ils sont proches. Pour tester cette hypothèse, nous continuerons à échantillonner aléatoirement des exemples, mais 50% des exemples auront un niveau de complexité facile (degré <0,25), et les 50% restants auront un niveau de complexité difficile (degré > 0,5). Cela créera une distinction claire entre les exemples en fonction de leur complexité.

n	Moyenne		Écart Type	
	ChatGPT	Notre modèle	ChatGPT	Notre modèle
4	**0.655**	0.425	0.261	0.321
5	**0.6**	0.567	0.249	0.3
8	**0.483**	0.412	0.154	0.142
10	**0.504**	0.429	0.116	0.099
20	**0.495**	0.432	0.077	0.091

TABLE 4 – Moyennes et écarts-types pour différentes tailles d'échantillons (aléatoire mais 50% contextes facile, 50% très difficile)

La Table 4 et la figure 3 montrent que les scores sont meilleurs avec ChatGPT dans ce scénario. ChatGPT se comporte de manière plus satisfaisante lorsque l'écart entre les niveaux de complexité des contextes est plus élevé, ce qui indique qu'il peut plus facilement différencier entre les contextes ayant des niveaux de complexité clairement distincts. Les valeurs d'écart-type deviennent également plus petites lorsqu'on les compare avec les résultats de la Table 3, ce qui indique que l'échantillon est plus représentatif et que les valeurs sont plus proches de la moyenne par rapport à lorsque les données sont échantillonnées de manière aléatoire à 100%.

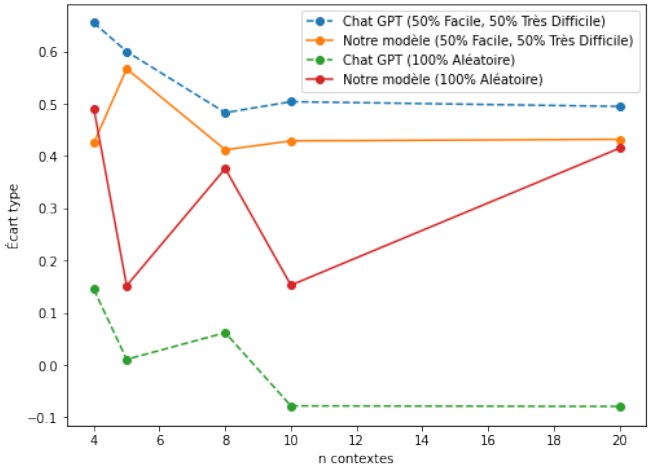


FIGURE 3 – Scores moyens en fonction de la taille de l'échantillon.

5 Conclusion

Dans cet article, nous avons proposé une méthode pour prédire la complexité lexicale. Cette méthode repose sur l'utilisation de modèles de langue pré-entraînés et de caractéristiques de fréquence. Nous avons montré que ChatGPT a une certaine capacité à comparer et classer des instances ayant le même mot-cible en fonction de leur complexité lexicale, notamment lorsque le nombre d'instances est limité. Dans de tels cas, il obtient de meilleures performances que notre modèle. Cependant, cette tâche devient plus difficile pour ChatGPT à mesure que le nombre d'instances comparées augmente, en particulier lors de l'introduction de différents mots-cibles dans des contextes variés. En effet, ChatGPT rencontre des difficultés lorsque la complexité entre les instances est très similaire. Dans de telles situations, il semble préférable d'utiliser un modèle spécifiquement entraîné pour cette tâche. De plus, un tel modèle peut produire un degré de complexité plus précis.

Pour les prochaines étapes de recherche, nous envisageons d'explorer davantage le potentiel du modèle dans des environnements multilingues. Cette exploration impliquera l'utilisation de méthodes d'apprentissage par transfert pour étendre la capacité de notre approche à prédire la complexité lexicale dans différentes langues. Nous envisageons également d'explorer l'utilisation d'autres modèles open source, en plus de ChatGPT, pour des questions de reproductibilité. Par ailleurs, nous prévoyons également d'explorer des techniques d'annotation et d'augmentation des données pour améliorer les performances de prédiction de la complexité lexicale. L'annotation supplémentaire des données pourrait aider à mieux capturer les nuances de la complexité lexicale dans différents contextes, tout en aidant à accroître la diversité des exemples disponibles pour l'entraînement du modèle, améliorant ainsi sa capacité à généraliser à de nouveaux contextes.

Références

ALFTER D. (2021). *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*.

AMIN M. M., CAMBRIA E. & SCHULLER B. W. (2023). Will affective computing emerge from foundation models and general ai ? a first evaluation on chatgpt. *arXiv preprint arXiv :2303.03186*.

BANI YASEEN T., ISMAIL Q., AL-OMARI S., AL-SOBH E. & ABDULLAH M. (2021). JUST-BLUE at SemEval-2021 task 1 : Predicting lexical complexity using BERT and RoBERTa pre-trained language models. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, p. 661–666, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.semeval-1.85.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

CHEN L., ZAHARIA M. & ZOU J. (2023). How is chatgpt's behavior changing over time ? *arXiv preprint arXiv :2307.09009*.

DAI H., LIU Z., LIAO W., HUANG X., CAO Y., WU Z., ZHAO L., XU S., LIU W., LIU N., LI S., ZHU D., CAI H., SUN L., LI Q., SHEN D., LIU T. & LI X. (2023). Auggpt : Leveraging chatgpt for text data augmentation.

DEUTSCH T., JASBI M. & SHIEBER S. (2020). Linguistic features for readability assessment. *arXiv preprint arXiv :2006.00377*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

EVERT S. (2010). Google web 1t 5-grams made easy (but not for the computer). In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, p. 32–40.

GOODING S. & KOCHMAR E. (2019). Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1148–1153, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1109.

HE P., GAO J. & CHEN W. (2023). Debertav3 : Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

HUANG F., KWAK H. & AN J. (2023). Is ChatGPT better than human annotators ? potential and limitations of ChatGPT in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023* : ACM. DOI : 10.1145/3543873.3587368.

KUZMAN T., LJUBEŠIĆ N. & MOZETIČ I. (2023). Chatgpt : beginning of an end of manual annotation ? use case of automatic genre identification. *arXiv preprint arXiv :2303.03953*.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LEE B. W., JANG Y. S. & LEE J. H.-J. (2021). Pushing on text readability assessment : A transformer meets handcrafted linguistic features. *arXiv preprint arXiv :2109.12258*.

LIU Y., HAN T., MA S., ZHANG J., YANG Y., TIAN J., HE H., LI A., HE M., LIU Z. *et al.* (2023). Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, p. 100017.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach.

PAETZOLD G. & SPECIA L. (2016). SemEval 2016 task 11 : Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 560–569, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/S16-1085.

PAN C., SONG B., WANG S. & LUO Z. (2021). DeepBlueAI at SemEval-2021 task 1 : Lexical complexity prediction with a deep ensemble approach. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, p. 578–584, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.semeval-1.72.

PENG K., DING L., ZHONG Q., SHEN L., LIU X., ZHANG M., OUYANG Y. & TAO D. (2023). Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv :2303.13780*.

RAY P. P. (2023). Chatgpt : A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SHARDLOW M. (2013). A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, p. 103–109, Sofia, Bulgaria : Association for Computational Linguistics.

SHARDLOW M., COOPER M. & ZAMPIERI M. (2020). CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, p. 57–62, Marseille, France : European Language Resources Association.

SHARDLOW M., EVANS R., PAETZOLD G. H. & ZAMPIERI M. (2021). SemEval-2021 task 1 : Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, p. 1–16, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.semeval-1.1.

SHARDLOW M., EVANS R. & ZAMPIERI M. (2022). Predicting lexical complexity in english texts : the complex 2.0 dataset. *Language Resources and Evaluation*, **56**(4), 1153–1194. DOI : 10.1007/s10579-022-09588-2.

SPEER R. (2022). rspeer/wordfreq : v3.0. DOI : 10.5281/zenodo.7199437.

WILKENS R., ALFTER D., WANG X., PINTARD A., TACK A., YANCEY K. P. & FRANÇOIS T. (2022). FABRA : French aggregator-based readability assessment toolkit. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 1217–1233, Marseille, France : European Language Resources Association.

WILKENS R., WATRIN P., CARDON R., PINTARD A., GRIBOMONT I. & FRANÇOIS T. (2024). Exploring hybrid approaches to readability : experiments on the complementarity between linguistic features and transformers. In Y. GRAHAM & M. PURVER, Éds., *Findings of the Association for Computational Linguistics : EACL 2024*, p. 2316–2331, St. Julian's, Malta : Association for Computational Linguistics.

YIMAM S. M., BIEMANN C., MALMASI S., PAETZOLD G., SPECIA L., ŠTAJNER S., TACK A. & ZAMPIERI M. (2018). A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 66–78, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/W18-0507.

ZAMPIERI M., TAN L. & VAN GENABITH J. (2016). MacSaar at SemEval-2016 task 11 : Zipfian and character features for ComplexWord identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 1001–1005, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/S16-1155.

ZHANG B., DING D. & JING L. (2022). How would stance detection techniques evolve after the launch of chatgpt ? *arXiv preprint arXiv :2212.14548*.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Quel *workflow* pour les sciences du texte ?

Antoine Widlöcher
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
antoine.widlocher@unicaen.fr

RÉSUMÉ

Le triomphe des approches adossées à des méthodes d'apprentissage, dans de nombreuses branches de notre discipline, tend à occulter une part importante des domaines d'investigation pourtant intimement liée au traitement automatique des langues. Nous proposerons, pour commencer, de faire un pas dans la direction opposée, en faveur de ce que nous nommerons ici les *sciences du texte*, en les distinguant de l'ingénierie de la langue, dont l'omniprésence explique largement cette occultation. Nous voudrions ensuite contribuer à mettre en évidence la méthode propre à cette branche des savoirs, méthode commune pouvant permettre de faire sortir de l'isolement des travaux hétérogènes liés par un même rapport au texte. Nous voudrions enfin nous concentrer sur la phase de ce *workflow* qui demeure actuellement la plus difficile, celle de l'expérimentation sur corpus, et proposer un cadre pour la mise en place d'environnements d'expérimentation appropriés.

ABSTRACT

What workflow for text science ?

The triumphal success of approaches based on machine-learning methods, in many branches of our discipline, tends to marginalise a large part of the fields of research which are nevertheless intimately linked to natural language processing. First of all, we propose to take a step in the opposite direction, in favour of what we call here *text sciences*, distinguishing them from human language technologies, whose omnipresence largely explains this marginalisation. Then we would like to contribute to highlight the specific method of theses approaches, which share a common relationship with the text. Finally, we would like to focus on the phase of their *workflow* that currently remains the most difficult, that of corpus-based experimentation, and to propose a framework for setting up appropriate environments for experimentation.

MOTS-CLÉS : Sciences du texte, environnement d'expérimentation, expérimentation sur corpus.

KEYWORDS: Text science, experimental environement, corpus-based experimentation.

1 Sciences et ingénierie du texte

Une part très importante des travaux de notre communauté se porte désormais naturellement vers les approches adossées à des méthodes d'apprentissage, dont les succès fulgurants ne sont pas contestables. Que cette tendance forte ait d'ores et déjà transformé en profondeur notre discipline ne doit pas nous dispenser de méditer d'une part à la nature et à la portée de ces succès et d'autre part à la zone d'occultation qui résulte de cette forte mise en lumière. Pour tirer au clair la portée de ces succès incontestables, il nous semble important d'en souligner l'horizon applicatif. Qu'il s'agisse de classification de documents, de traduction, d'extraction d'information ou de résumé, pour citer

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 630–649.

quelques exemples bien documentés des *applications* du TAL, la tentation d'y voir le texte comme un obstacle à surmonter, ne doit pas être minimisée. Le texte n'y est pas alors regardé comme une fin, mais comme un moyen d'atteindre certains objectifs applicatifs, la pertinence d'une approche étant alors potentiellement mesurée à l'aune d'un critère étranger au texte lui-même et à sa compréhension.

D'autres travaux, au contraire, s'appuyant aussi sur une machinerie computationnelle, visent l'étude du texte lui-même, la mise en lumière de ses lois. Là où les précédents, qui relèvent selon nous pour cette raison de l'*ingénierie du texte*, visent quelque chose derrière le texte, au moyen du texte, ces derniers, de l'ordre de la *science du texte*, visent le texte lui-même et sa compréhension. Si cette ligne de démarcation au sein des « disciplines du texte » évoque celle qu'identifie (Rastier, 2001) entre *arts* et *sciences* du texte, elle renvoie davantage ici à une différence de visée entre les différents paradigmes d'étude, au sein des approches scientifiques. Non sans rappeler aussi de vieilles querelles entre TAL et linguistique computationnelle, cette cartographie disciplinaire doit être précisée. On pourrait dire que les approches visant le texte lui-même et celles qui sont guidées par des impératifs applicatifs ont le TAL en commun. Celui-ci renvoie à un ensemble de méthodes de traitement des données textuelles utilisable dans une perspective ou dans l'autre. Ainsi, la linguistique computationnelle ne s'oppose pas au TAL mais le précise, par sa visée spécifique, son absence d'autres objets que le texte lui-même et sa compréhension. Mais elle n'épuise pas le concept des sciences du texte, dont relèvent tout autant, par exemple, des études littéraires ou philologiques et de nombreux chantiers ouverts à l'interface des humanités numériques, là où les sciences humaines rencontrent le besoin d'explorer computationnellement le texte, pour le comprendre. Pour ces disciplines, l'identification d'un phénomène textuel ne suffit pas ; il faut encore comprendre comment et pourquoi il se produit. Cette dimension intrinsèquement explicative y va de pair avec l'omniprésence de l'interprète humain.

Faire entendre la voix des sciences du texte, dans un contexte disciplinaire où l'omniprésence des impératifs de l'ingénierie du texte ne cesse d'en fragiliser l'existence, est-ce à dire pour autant qu'une hétérogénéité radicale devrait interdire toute communication entre ces disciplines ? Évidemment non. D'une part parce qu'une communauté de moyens (de formalisation, de calcul...) rend évidemment possible des avancées communes. D'autre part parce que la science du texte pourra toujours en droit éclairer son ingénierie, comme l'ont souvent montré les approches linguistiquement informées du TAL, qui n'ont certes pas actuellement le vent en poupe... Enfin, car les moyens puissants élaborés pour son ingénierie peuvent constituer aussi des moyens d'observation utiles à la science du texte, pourvu que sa visée explicative soit bien entendue, ce à quoi les nombreux travaux actuels consacrés à l'explicabilité dans les méthodes d'apprentissage pourraient évidemment contribuer. Reste que les travaux relevant de la science du texte, travaux visant exclusivement le texte et sa compréhension, doivent pouvoir exister. Leur marginalisation relative actuelle impose de repenser leurs spécificités et leur fond théorique commun, ne serait-ce que pour mettre en évidence leur omniprésence dans notre communauté et pour mettre en lumière les moyens dont ils doivent disposer pour s'y épanouir.

2 Quel *workflow* pour les sciences du texte ?

Par *sciences du texte*, nous désignons le complexe théorique et expérimental qui vise l'étude et la compréhension du texte et de ses lois, à différentes échelles (du caractère au corpus) et dans différentes perspectives (de la forme à l'interprétation). Leur scientificité repose sur les éléments suivants :

1. Leur **démarche expérimentale** s'appuie sur l'articulation entre des phases inductives et hypothético-déductives menées par confrontation aux données de l'expérience, c'est-à-dire aux données d'un **corpus**.

2. L'exigence de validation sur corpus suppose la capacité à identifier les données de l'expérience dont l'étude est menée, c'est-à-dire la capacité à **constituer des observables** dont le modèle sera modèle, dont la théorie sera théorie...
3. L'ensemble de leur démarche doit satisfaire, dans chacune de ses phases, aux **exigences de reproductibilité**. Cela suppose qu'un degré de formalisation suffisant soit atteint dans l'énoncé des modèles, des hypothèses et des paramètres d'expérimentation, pour que la communauté puisse vérifier que, dans les mêmes conditions, les mêmes causes conduisent aux mêmes effets et identifier, le cas échéant, erreurs et biais dans l'interprétation des résultats, conformément à l'**impératif de réfutabilité** qui prolonge naturellement celui de reproductibilité.

De ces contraintes épistémologiques résulte un schéma de *workflow* dans lequel s'inscriront naturellement les travaux en sciences du texte. La figure 1 donne sa forme essentielle.

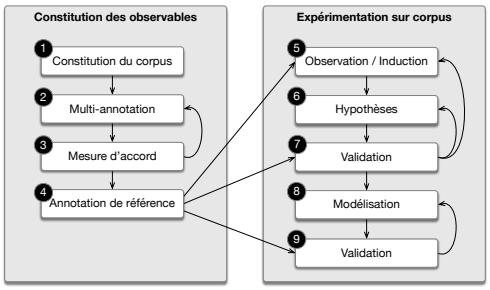


FIGURE 1 – Forme générale du *workflow* des sciences du texte

Nous distinguons ici deux grandes phases au sein de ce *workflow*. Les phases ❶ à ❹ correspondent à la constitution des observables, c'est-à-dire à la constitution d'un corpus de référence au sein duquel des occurrences du phénomène textuel ciblé ont été identifiées. Les phases ❺ à ❾ visent l'établissement d'un modèle du phénomène ciblé.

2.1 La constitution des observables

Les phases ❶ à ❹ sont très largement balisées, même si certains problèmes restent évidemment ouverts et si un défaut de systématicité est encore nettement observable à l'échelle de notre communauté. Des facteurs de blocage en résultent qui doivent être mis en évidence. La question de la constitution de corpus et celle de leur représentativité fait l'objet d'une littérature relativement abondante, dont la revue *Corpus* est par exemple le reflet dans la communauté francophone, depuis ses premiers numéros (Mellet, 2002). Un point important concerne aussi la diffusion des corpus. Sur ce point, on peut s'étonner que, si un assez net consensus se dégage concernant l'utilisation de schémas éprouvés de représentation des corpus textuels, et cela depuis fort longtemps, notamment autour de la TEI [1] (Ide & Véronis, 1995; Burnard *et al.*, 2006), le disponibilité de corpus de référence selon ce standard de fait demeure assez sporadique dans nos disciplines, alors que les sciences humaines et les disciplines littéraires, soucieuses d'établissement de données finement ciselées, se sont pour leur part nettement approprié les technologies associées. Si l'inscription dans la nébuleuse XML est acquise, l'absence de systématicité dans le recours à un vocabulaire commun est pourtant un évident facteur de ralentissement.

1. https://tei-c.org/

La méthodologie d'ensemble selon laquelle peuvent être menées les campagnes d'élaboration de ressources enrichies, annotées, fait l'objet d'une littérature très utile (Fort, 2016). L'importance de la multi-annotation ❷ a été largement soulignée pour garantir la fiabilité des données produites, en particulier sur les objets textuels peu étudiés ou difficiles à interpréter, et, au moins depuis (Artstein & Poesio, 2008), les moyens de mesurer l'émergence d'un consensus entre annotateurs ❸, condition *sine qua non* pour la constitution de données *de référence* ❹, sont étudiés en tant que tels. Différents environnements logiciels dédiés à l'annotation ont été proposées, parmi lesquels on peut notamment citer (Stenetorp *et al.*, 2012), (Widlöcher & Mathet, 2012) et plus récemment (Klie *et al.*, 2018). À mi-chemin entre l'annotation et l'exploration de corpus telle qu'elle sera définie ci-dessous, on peut également mentionner les travaux de (Landragin *et al.*, 2012), qui combinent ces deux phases.

Bien entendu, la diversité et l'hétérogénéité des phénomènes susceptibles d'être étudiés et annotés retardent l'élaboration de modèles communs de représentation des données enrichies, mais aussi, inévitablement, la mise en place de méthodes communes pour la comparaisons de structures annotées, préalable pourtant nécessaire à la mesure d'accord sur des données multi-annotées, dont l'absence bloque l'émergence de données de référence. De fait, si les moyens ne manquent pas pour la prise en charge de tâches d'annotation relevant de la pure catégorisation d'items textuels déjà identifiés, les propositions sont dramatiquement moins nombreuses pour des tâches imposant de surcroît le positionnement des objets dans le *continuum* textuel. Et si des propositions voient le jour, quoique très sporadiquement, pour la prise en charge des tâches dites d'*unitizing* (de repérage d'unités dans le texte), c'est-à-dire pour leur annotation puis pour la comparaison des productions de plusieurs annotateurs, le travail sur des structures plus complexes, notamment relationnelles (par exemple en syntaxe, rhétorique ou argumentation), impose souvent le recours à des méthodes *ad hoc*, élaborées pour chaque objet, en rendant fatalement difficile la confrontation des modèles et des théories.

2.2 Absence d'un cadre commun pour l'expérimentation sur corpus

Sur la seule question de la constitution des observables, un immense travail reste donc encore à accomplir, pour définir objets et méthodes communs. Nous voudrions toutefois nous concentrer sur les phases suivantes du *workflow*, celles de l'expérimentation sur corpus (❺ à ❾). Car si chacun, dans sa discipline, avance évidemment sur ce terrain, nous ne voyons pas cependant émerger un cadre commun qui permettrait, à l'échelle de la communauté, la reproduction simple des expériences menées ici ou là, la confrontation des approches, le partage des résultats. Que l'absence d'un tel cadre commun soit dommageable, tout le monde en conviendra. Nous voudrions savoir à quelles conditions son émergence pourrait être rendue possible et, inversement, quels éléments font blocage.

La question pourtant n'est pas neuve. Sans remonter aux origines de notre discipline – car en réalité les conditions que nous mentionnons sont presque imposées par la démarche scientifique elle-même – on retiendra notamment qu'il y a quinze ans déjà, (Enjalbert *et al.*, 2008), dans le prolongement de journées ATALA consacrées aux « architectures logicielles pour articuler les traitements sur corpus » (en 2005), notre communauté francophone se posait frontalement la question de la mise en place d'environnements d'expérimentation sur corpus, de la reproductibilité des expériences, du partage des ressources, de l'interopérabilité entre les systèmes... Et depuis lors au moins, différentes solutions méthodologiques et logicielles ont été proposées pour répondre à ces exigences. Nous voudrions ici proposer quelques repères dans cette nébuleuse, en cherchant surtout à identifier les paradigmes concurrents et les lignes de démarcation principales entre les différentes options envisagées, pour mieux saisir les raisons pour lesquelles principes et solutions communs tardent à émerger.

Pour nous orienter dans cette nébuleuse [2], il peut être utile de commencer par distinguer deux traditions restées jusqu'ici assez étanches l'une à l'autre, répondant à deux manières d'envisager les données textuelles. Nous proposons de formuler l'esprit de cette démarcation en nous appuyant sur la distinction souvent faite, notamment dans le champ des humanités numériques, entre *distant reading* (Moretti, 2013) et *close reading*, pour désigner l'hétérogénéité entre des approches considérant les données textuelles, souvent assez massives, d'une certaine hauteur et souvent au moyen de méthodes statistiques, et des approches restant davantage au contact des énoncés et des occurrences en contexte.

Relèvent clairement du premier paradigme les propositions faites dans le champs de l'analyse statistique des données textuelles (Lebart *et al.*, 1998) et notamment, surtout au niveau francophone, les travaux de la tradition lexicométrique issue de (Lafon, 1984), dont sont inspirés des environnements intégrés très complets d'exploration de corpus tels que la plate-forme TXM (Heiden, 2010).

Visant davantage l'exploration du corpus en restant au contact des énoncés, en pilotant généralement la recherche d'occurrences des phénomènes visés par l'expression de règles établies dans des formalismes dédiés à des niveaux d'analyse variés (échelles lexicale, syntagmatique, discursive...), d'autres environnements intégrés d'expérimentation sur corpus ont vu le jour, parmi lesquels ont peut citer notamment Gate (Cunningham *et al.*, 2002, 2011, 2013), Nooj (Silberztein, 2016), Unitex (Paumier *et al.*, 2021) et LinguaStream (Widlöcher & Bilhaut, 2008). L'enthousiasme suscité par ces environnements intégrés, puissants mais difficiles à prendre en main, semble avoir connu un certain infléchissement. Si Gate, Nooj et Unitex sont toujours maintenus et jouissent d'une communauté active (ce n'est pas le cas de LinguaStream), on a néanmoins le sentiment que la communication scientifique associée à ces plate-formes a sensiblement diminué ces dernières années, au-delà de cercles assez spécifiques, signe peut-être d'un relatif déphasage par rapport aux attentes de notre communauté. On ne voit pas, du moins, émerger un consensus large en faveur d'un environnement commun d'expérimentation en TAL qui reposerait sur ce principe.

On a vu au contraire se multiplier les approches *par librairie* visant, plutôt que l'élaboration d'un environnement intégré, l'exploitation depuis un langage d'interfaçage ou d'intégration tel que Python notamment, très apprécié pour le prototypage rapide et dans de nombreuses sciences expérimentales. À des outils dédiés à la langue comme NLTK (Bird *et al.*, 2009), régulièrement utilisé pour la recherche et l'enseignement, en vertu notamment de la pluralité des paradigmes d'analyse qu'il permet d'exploiter et le contrôle qu'il donne sur l'expression de règles dans différents formalismes, il convient d'ajouter d'une part des outils comme spaCy [3], certes dédiés à la matière textuelle, mais visant davantage la mise en production d'applications que l'expérimentation sur corpus, et d'autre part des outils dédiés à la science des données et à l'apprentissage machine, tel Scikit-learn (Pedregosa *et al.*, 2011), qui intègrent des éléments permettant de traiter des données textuelles, dont les spécificités sont d'ailleurs souvent écartées assez rapidement, au profit de représentations tabulaires et vectorielles évidemment plus en phase avec les méthodes courantes en apprentissage.

2.3 Quel cadre pour l'expérimentation sur corpus ?

La partie expérimentale de nos disciplines offre donc encore souvent le spectacle d'une collection d'approches difficiles à unifier dans un mouvement commun, où chaque travail avance son propre

2. Nous laissons de côté, dans ce rapide survol, des infrastructures d'assez bas niveau, comme notamment le *framework* UIMA (https://uima.apache.org/), et les questions importantes qu'elles posent en termes d'interopérabilité, pour nous concentrer en priorité sur les environnements plus immédiatement dédiés à l'analyse des textes.

3. https://spacy.io

formalisme, sa propre représentation du texte, des règles d'analyse... Conscient que toute tentative d'avancer à rebours de cette tendance naturellement entropique nous fait prendre le risque d'une fausse solution de plus, et donc finalement d'une augmentation du désordre, nous voudrions néanmoins envisager quelques pistes pour essayer d'y remédier. Nous les présenterons sous la forme d'une série de principes, non sans avoir d'abord souligné que nous maintenons largement les recommandations faites par (Widlöcher & Bilhaut, 2008), recommandations que nous complétons et que nous proposons d'amender, parfois en les radicalisant, parfois en relaxant certaines contraintes difficiles à satisfaire sans entrer en contradiction avec d'autres principes d'importance égale ou supérieure.

P1 - Hétéronomie fondamentale du chercheur De façon peut-être un peu provocante, nous voudrions aborder cette énumération des principes par des considérations de nature presque sociologique concernant certains *habitus* de notre communauté des sciences du texte. Nous pensons ici plus précisément à la tradition relativement forte de l'autonomie radicale du chercheur, par laquelle nos disciplines s'inscrivent d'ailleurs dans le prolongement des sciences humaines et de l'esprit, où l'autorité du savant suppose souvent la solitude. Nous entendons par là l'ambition (et souvent d'ailleurs la capacité admirable) du chercheur à maîtriser individuellement l'ensemble des phases du processus de construction intellectuelle, conceptuelle et expérimentale qu'implique l'étude des objets qu'il s'est fixés [4]. Or, pour admirable qu'elle soit, cette parfaite autonomie n'en demeure pas moins tout à fait exceptionnelle en pratique, et le risque est grand dès lors que cette ambition devienne contre-productive si l'on n'en mesure pas la limite. Si nous évoquons ce tropisme, c'est qu'il ne nous semble pas étranger à l'échec relatif des grands systèmes intégrés, qui reposent en partie sur l'hypothèse, trompeuse selon nous, que chacun pourra, en autonomie, mener ses expérimentations sur corpus. Nous voudrions promouvoir au contraire l'idée d'une hétéronomie fondamentale du chercheur, et prendre la juste mesure de la nécessité qui en résulte de clarifier les moyens d'une collaboration féconde entre les différents corps de métiers impliqués dans les sciences du texte. Être linguiste, statisticien ou algorithmicien, ce n'est résolument pas la même chose (dans une large majorité des cas), et, plutôt que de viser l'objectif illusoire d'une autonomie parfaite, nous devons plutôt nous interroger sur les moyens de rendre fertile la collaboration (*cf.* **P4 P5**).

P2 - Données textuelles de référence, données d'entrée et données de sortie Que le partage des données importe davantage que celui des outils, c'est là un fait qui a été maintes fois souligné. Quelles conséquences pratiques devons-nous en tirer ? D'abord, la nécessité de prendre au sérieux le mode de représentation faisant consensus pour la représentation des corpus textuels. De ce point de vue, il est clair que les technologies XML, tombées en désuétude chez les informaticiens, demeurent incontournables pour la représentation des données semi-structurées dont les corpus textuels sont la parfaite illustration. Il en résulte non seulement la nécessité de mettre en place des environnements logiciels capables de consommer de telles données en entrée (ce qui est souvent admis), mais aussi la nécessité de produire en sortie des données répondant à cette norme, notamment pour que les données textuelles enrichies issues de l'analyse demeurent compatibles avec les outils d'observation et les chaînes éditoriales définies pour la diffusion et la valorisation des données initiales. S'il n'est pas mécaniquement nécessaire que les processus d'analyse opèrent en conséquence à chaque étape sur des représentations XML des données, potentiellement coûteuses, il est en revanche nécessaire que toutes les représentations intermédiaires fassent systématiquement référence aux structures initiales, pour qu'à chaque instant, et surtout en fin de traitement, elle puissent y être projetées ou rapportées.

P3 - Complémentarité des *altitudes* d'observation Par *altitude* d'observation, nous renvoyons ici à la distinction évoquée ci-dessus entre *close* et *distant reading*. Si de nombreux chercheurs appré-

4. Ce dont l'usage des productions scientifiques à signature unique est le reflet éloquent.

hendent alternativement les données avec des méthodes distantes (statistiques, lexicométriques...) et des méthodes plus directement focalisées sur le repérage d'occurrences en contexte, il faut bien néanmoins reconnaître que l'adoption de ces différents points de vue relève de traditions, de méthodologies et conséquemment d'outils assez hétérogènes. S'il est raisonnable de supposer la complémentarité de ces altitudes d'observation, les premières étant notamment indispensables dans les phases inductives et de vérifications des hypothèses à petite échelle (celle du corpus ou du sous-corpus), quand les secondes interviennent dans les phases de modélisation de structures potentiellement complexes et dans le repérage d'occurrences à grande échelle (en contexte, à l'échelle de la phrase ou du discours), il est alors nécessaire de disposer de moyens efficaces de circulation entre ces niveaux. Cela impose d'abord que les descriptions d'objets analysés en contexte puissent être aisément collectées et synthétisées dans des représentations manipulables par les outils d'observations de plus haute altitude. Inversement, cela suppose que des éléments observés à haute altitude (des fréquences, des régularités, des attirances...) puissent être facilement reformulés en configurations observables *in situ*. On ne saurait trop insister sur l'importance du *retour au texte* pour les sciences du texte. Elle impose en particulier la disponibilité d'outils de visualisation adaptés aux modes conventionnels de représentation des données (*cf.* **P2**), pour que les objets soient observés dans leur contexte initial.

P4 - Abstraction progressive des formes de surface et variabilité des perspectives Avoir le texte pour matière n'implique pas que chaque étape et chaque niveau d'analyse doive reposer exclusivement sur sa forme initiale. Au contraire, pour les traitements computationnels, comme pour les travaux réalisés sans traitement mécanique, il est clair que certains niveaux d'analyse doivent pouvoir s'appuyer sur les résultats obtenus à d'autres niveaux, certains ordres classiques conduisant même d'ailleurs à des *pipelines* parfois figés à l'excès dont l'enchaînement [Tokenisation > POS Tagging > Analyse syntaxique > Analyse du discours] donne une bonne illustration. S'il nous semble important de garantir au contraire une assez grande liberté dans les enchaînements d'analyse à mettre en place (*cf. infra*), reste qu'une analyse d'un niveau quelconque doit pouvoir s'appuyer sur les sorties d'un niveau d'analyse préalable, et qu'une analyse subséquente devra pouvoir, à son tour, s'appuyer sur ses propres sorties. Il en résulte que chaque niveau d'analyse doit pouvoir exploiter, non seulement la matière textuelle initiale et ses formes de surface, mais surtout les représentations antérieurement calculées, cette forme d'indirection conduisant en pratique à une *abstraction progressive des formes de surface*. Chaque niveau d'analyse, humain ou computationnel, produit des *annotations*, souvent obtenues par abstraction depuis des annotations déjà produites, sur lesquelles les analyses subséquentes devront à leur tour pouvoir s'appuyer. Ces annotations doivent combiner la *localisation* dans le texte des phénomènes identifiés, en référence à la représentation initiale (*cf.* **P2**), et une *représentation symbolique* de leur interprétation, une caractérisation utilisable par les traitements subséquents. Cela n'implique pas que chaque niveau d'analyse doive tenir compte de toutes les représentations préalablement calculées. Au contraire, chacun devra pouvoir spécifier la *perspective* qui est la sienne, c'est-à-dire la manière dont il se rapporte au texte, en explicitant les abstractions sur lesquelles il s'appuie. Les avantages qui peuvent en résulter, en terme d'expressivité de chaque modèle d'analyse et en termes d'efficacité sur un plan combinatoire, sont potentiellement colossaux. La confrontation des points de vue (entre le linguiste, l'informaticien...) est toutefois évidemment nécessaire à l'exploitation de ces bénéfices (*cf.* **P1**).

P5 - Complémentarité des formalismes et modèles d'analyse L'étude d'objets textuels variés a naturellement conduit à l'émergence de multiples formalismes et modèles d'analyse, dont la pouvoir expressif et l'efficacité ont été établis pour les objets pour lesquels ils ont été pensés. Viser la mise en place d'un cadre expérimental commun, ce n'est évidemment pas proposer en la matière un réductionnisme total supposant l'omnipotence d'un formalisme particulier. Il est au contraire

nécessaire de faire jouer la complémentarité des formalismes et modèles d'analyse. Nous défendons du reste l'idée que, même si un formalisme et un modèle d'analyse ont généralement été élaborés pour l'étude d'objets particuliers, leur exploitation à d'autres niveaux peut s'avérer d'autant plus féconde que la variabilité des perspectives sur le texte donne une grande liberté dans sa lecture (*cf.* **P4**). L'exploitation des automates et expressions régulières sur des séquences quelconques, au-delà du niveau des chaînes de caractères pour lesquelles les modèles initiaux ont été pensés, illustre bien l'extension possible d'un domaine d'application. Pour que cette extension demeure possible, il est nécessaire que chaque formalisme et modèle d'analyse retenu ne soit pas inféodé à une certaine représentation du texte qu'il consomme, mais puisse au contraire opérer depuis une perspective quelconque sur le texte. Reste que l'identification d'un modèle approprié, en termes d'expressivité et d'efficacité, pour un problème textuel donné, demeure un problème complexe, imposant la collaboration entre les différents corps de métier (*cf.* **P1**).

P6 - Représentation unifiée des annotations, des extractions et des représentations synthétiques Dire qu'un formalisme et un modèle d'analyse quelconques doivent pouvoir opérer depuis une perspective quelconque à une échelle quelconque (*cf.* **P5**), c'est dire aussi que des enchaînements d'analyse doivent pouvoir être réalisés dans un ordre quelconque, que nous ne pouvons pas nous référer à des ordres classiques pour fixer les entrées/sorties de tel ou tel composant. Cela impose au contraire que les entrées/sorties de chaque niveau d'analyse soient encodées dans un modèle unifié, pouvant être produit et consommé à n'importe quel moment du processus d'analyse. En réponse au principe d'abstraction progressive et de variabilité des perspectives (*cf.* **P4**), chaque moment de l'analyse dédié au repérage d'occurrences en *close reading* (*cf.* **P3**) portera donc sur les annotations produites en amont, qui localisent dans les données initiales (*cf.* **P2**) et caractérisent les objets déjà reconnus, et produira de nouvelles annotations utilisables en aval, les unes et les autres étant représentées de manière homogène. Les extractions et représentations synthétiques élaborées en *distant reading* (*cf.* **P3**), elles aussi représentées de manière unifiée, s'appuieront elles aussi sur les annotations disponibles en amont (pour une perspective donnée) et les représentations synthétiques déjà établies, et seront elles-mêmes réutilisables en aval.

P7 - Déclarativité ciblée Les bonnes propriétés de la déclarativité pour la formalisation, l'étude et la capitalisation des règles d'analyse sont bien connues et nous la préconisons sans réserve pour l'ensemble des formalismes dédiés à la description des règles d'analyse (*cf.* **P5**). Le fait que l'approche déclarative masque intentionnellement les appareils procéduraux sous-jacents, pour l'application des règles, impose néanmoins clairement une concertation entre les corps de métier (*cf.* **P1**), ne serait-ce que pour que les conséquences algorithmiques restent sous contrôle. Au-delà de ce paramétrage des analyseurs, pour lequel la déclarativité doit être privilégiée, l'articulation de l'ensemble des traitements, dans des processus de type *pipeline* ou plus itératifs, pourra au contraire tirer bénéfice de l'adoption d'un paradigme plus impératif. En effet, pour le pilotage de la lecture des données d'entrée, pour la configuration des sorties, pour la gestion de flots d'exécution non strictement séquentiels, on tirera avantage du passage par un langage de programmation pour l'articulation des différentes phases d'analyse. Le recours à un langage d'intégration aura aussi l'avantage de simplifier l'utilisation combinée de librairies variées, pourvu que le langage retenu y donne effectivement accès.

P8 - Traces expérimentales La satisfaction des contraintes liées à la progression expérimentale, à la reproductibilité et à la réfutabilité impose évidemment pour commencer la disponibilité des données d'entrée et de sortie, qui doivent en conséquence être représentées dans des formats ouverts et documentés. La consultation des sorties, évidemment indispensable à l'évaluation du traitement mis en place, doit être soutenue par des outils de visualisation appropriés n'imposant idéalement ni l'installation d'un quelconque environnement logiciel complexe, ni la réitération de l'ensemble des

calculs ayant permis leur production. La mise en évidence du paramétrage du processus d'analyse bénéficiera d'abord du respect de l'exigence de déclarativité (*cf.* **P7**). Si l'articulation des différents traitements est pour sa part prise en charge programmatiquement, une API parfaitement claire devra être proposée. Toute expérimentation entreprise dans ce cadre devra pouvoir en conséquence produire une trace du cheminement suivi par le chercheur, trace où seront présentés de manière articulée les données d'entrée, les paramètres d'enchaînement, les paramètres d'analyse et les sorties, ainsi que, bien entendu, la justification des choix et l'interprétation des résultats.

3 Implémentation des ces principes dans la librairie Skhólion

Nous avons fait le choix ici de nous concentrer sur la présentation des principes qui nous semblent devoir être suivis pour la pratique expérimentale des sciences du texte. Pour rendre l'énoncé de ces principes plus concret et les illustrer, nous voudrions évoquer succinctement leur mise en œuvre au sein de la librairie Python Skhólion [5] que nous élaborons actuellement. Quelques illustrations de cette mise en œuvre sont données en annexes de cet article.

Dans l'esprit du principe **P1**, cette librairie vise à permettre la construction collective de dispositifs expérimentaux pour les sciences du texte. Elle doit permettre la collaboration efficace entre les différents acteurs de ces sciences et notamment entre le développeur et le spécialiste du texte, plutôt que de laisser à ce dernier le soin d'exploiter solitairement un environnement intégré puissant mais difficile à contrôler. En particulier, le principe selon lequel il plus aisé de vérifier la conformité d'un code donné à un problème posé, que d'élaborer *ex nihilo* la méthode de résolution, doit ici s'appliquer.

Conformément au principe **P2**, Skhólion opère sur des données textuelles d'entrée semi-structurées XML [6], auxquelles toutes les représentations produites feront référence [7], soit par la combinaison d'une expression XPath et de l'indication de la position de l'objet visé dans le nœud ciblé par l'expression, soit par référence à d'autres objets ainsi positionnés. En sortie, des données annotées sont produites par enrichissement des représentations initiales. En cours de traitement, des représentations variées sont utilisables, mais il est systématiquement possible de connaître l'ancrage des objets manipulés dans les données de référence.

La complémentarité des altitudes d'observation évoquée en **P3** est assumée, d'une part par la disponibilité de modèles d'analyse pour l'identification d'occurrences de phénomènes décrits dans différents formalismes, et d'autre part par la représentation des données, sans perte de leur ancrage, dans des structures adaptées à l'analyse de données. Nous nous appuyons notamment sur des *DataFrames* de la librairie Pandas (McKinney, 2010), qui permettent un accès direct aux outils puissants de cette librairie et des librairies sous-jacentes (en particulier NumPy (Harris *et al.*, 2020)), tout en permettant un pont vers les méthodes alimentées par des représentations tabulaires et vectorielles.

L'abstraction progressive des formes de surface du principe **P4** passe d'abord par la représentation des structures présentes dans les données initiales (notamment les structures typo-dispositionnelles). Toute unité porteuse de texte peut être segmentée en phrases, tokenisée et POS-taggée (un pont avec le Treetagger (Schmid, 1994) est assuré par défaut). L'ensemble des objets résultant peut être parcouru de différentes manières et utilisé pour produire des annotations, dont chacune, positionnée par rapport aux données de référence ou par rapport à des objets ainsi positionnés, est dotée d'une représentation

5. https://www.skholion.org
6. La librairie `lxml` (https://lxml.de) est largement utilisée.
7. La structuration initiale peut-être minimale, limitée par exemple à une décomposition en sections et paragraphes.

symbolique, sous la forme d'une structure de traits récursive. Les annotations produites pourront être exploitées par des analyseurs de plus haut niveau, qui pourront s'exprimer sur le texte qu'elles couvrent ou sur les représentations symboliques qu'elles portent, conformément à **P6**.

Conformément à **P7**, la manipulation du corpus, l'articulation des traitements et le paramétrage des sorties sont assurés de manière impérative, en Python, via l'API de Skhólion. En plus des analyseurs pouvant être directement écrits en Python, sur la base de cette API, pour le parcours des données et annotations disponibles, les modèles d'analyse proposés dans l'esprit de **P5**, **P6** et **P7**, encore en petit nombre pour le moment, permettent : 1) de construire une annotation correspondant à une structure présente dans les données d'entrée ; 2) de positionner librement une annotation dans le *continuum* textuel de toute unité disponible ; 3) de produire une annotation sur la base d'expressions régulières classiques sur la séquence de caractères de toute unité textuelle disponible et 4) d'utiliser la puissance des expressions régulières sur une séquence d'annotations produites en amont, en exprimant des contraintes sur les structures de traits associées, pour générer de nouvelles annotations. Ce premier ensemble a évidemment vocation à être étendu, conformément à **P5**, l'API de Skhólion devant simplifier l'intégration d'autres moyens d'analyse.

La démarche expérimentale est soutenue, dans l'esprit de **P8**, par l'utilisation systématique de formats ouverts pour la représentation des données, XML pour la représentation des données semi-structurées à dominante textuelle, JSON pour les autres données. En sortie de tout traitement, et notamment pour garantir un retour au texte conforme à **P3**, des représentations exploitant systématiquement les technologies du web (HTML, CSS, SVG et JavaScript) sont produites, qui peuvent être aisément consultées dans un navigateur, capitalisées et diffusées, sans aucune autre dépendance logicielle. Des visualisations sont proposées à l'échelle d'un texte ou à l'échelle du corpus, et des extractions d'objets ou de passages choisis sont aussi possibles. S'il est évidemment envisageable de travailler avec Skhólion dans un environnement de développement traditionnel, nous veillons à ce que l'utilisation de l'ensemble des outils proposés soit possible dans des Notebooks Jupyter (Kluyver *et al.*, 2016) et dans des environnements comme Jupyterlab, y compris pour la visualisation des résultats. De tels environnements, qui constituent le cadre privilégié d'utilisation de Skhólion, permettent de produire une trace du cheminement expérimental facile à capitaliser et à partager.

4 Conclusion

L'enthousiasme suscité dans notre communauté par les méthodes d'apprentissage et les LLM illustre parfaitement les potentialités et les risques où cet article trouve sa source. Car si la richesse des faits de langue que ces méthodes permettent de capturer justifie pleinement l'intérêt qu'on leur accorde, elles montrent aussi que bien des applications sont rendues possibles, qui ne mettent pas en pleine lumière les phénomènes linguistiques sous-jacents qu'elles exploitent. À travers ce plaidoyer pour les sciences du texte, nous voulions d'abord insister sur la nécessité de maintenir, au-delà de la réponse à des impératifs applicatifs (qui font souvent du texte un moyen), l'exigence de compréhension fine des phénomènes de langue (qui prend le texte pour fin). Les outils puissants auxquels l'ingénierie du texte a donné naissance peuvent aussi bien entendu contribuer à sa compréhension, pourvu que leur utilisation s'intègre dans un cadre expérimental dont cette compréhension est l'objectif clair. C'est ce cadre expérimental dont nous espérons pouvoir contribuer modestement à dessiner les contours, par la mise en lumière de principes pouvant guider sa mise en place et leur illustration dans une librairie naissante dédiée à l'expérimentation sur corpus, Skhólion.

Références

ARTSTEIN R. & POESIO M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, **34**(4), 555–596. DOI : http ://dx.doi.org/10.1162/coli.07-034-R2.

BIRD S., KLEIN E. & LOPER E. (2009). *Natural language processing with Python : analyzing text with the natural language toolkit*. O'Reilly.

BURNARD L., O'KEEFE K. O. & UNSWORTH J., Éds. (2006). *Electronic Textual Editing*. Modern Language Association.

CUNNINGHAM H., MAYNARD D., BONTCHEVA K. & TABLAN V. (2002). GATE : A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

CUNNINGHAM H., MAYNARD D., BONTCHEVA K., TABLAN V., ASWANI N., ROBERTS I., GORRELL G., FUNK A., ROBERTS A., DAMLJANOVIC D., HEITZ T., GREENWOOD M. A., SAGGION H., PETRAK J., LI Y. & PETERS W. (2011). *Text Processing with GATE (Version 6)*.

CUNNINGHAM H., TABLAN V., ROBERTS A. & BONTCHEVA K. (2013). Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS Computational Biology*, **9**(2), e1002854. DOI : 10.1371/journal.pcbi.1002854.

ENJALBERT P., BONTCHEVA K. & HABERT B., Éds. (2008). *Plate-formes pour le traitement automatique des langues*. Volume 49(2) de Revue TAL. France : ATALA (Association pour le Traitement Automatique des Langues).

FORT K. (2016). *Collaborative Annotation for Reliable Natural Language Processing : Technical and Sociological Aspects*. Wiley.

HARRIS C. R., MILLMAN K. J., WALT S. J. V. D., GOMMERS R., VIRTANEN P., COURNAPEAU D., WIESER E., TAYLOR J., BERG S., SMITH N. J., KERN R., PICUS M., HOYER S., KERKWIJK M. H. V., BRETT M., HALDANE A., RÍO J. F. D., WIEBE M., PETERSON P., GÉRARD-MARCHANT P., SHEPPARD K., REDDY T., WECKESSER W., ABBASI H., GOHLKE C. & OLIPHANT T. E. (2020). Array programming with NumPy. *Nature*, **585**(7825), 357–362. Publisher : Springer Science and Business Media LLC, DOI : 10.1038/s41586-020-2649-2.

HEIDEN S. (2010). The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In R. OTOGURO, K. ISHIKAWA, H. UMEMOTO, K. YOSHIMOTO & Y. HARADA, Éds., *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, p. 389–398, Sendai, Japon : Institute for Digital Enhancement of Cognitive Development, Waseda University.

IDE N. & VÉRONIS J., Éds. (1995). *Text Encoding Initiative : Background and Context*. Text, Speech and Language Technology. Dordrecht : Kluwer.

KLIE J.-C., BUGERT M., BOULLOSA B., CASTILHO R. E. D. & GUREVYCH I. (2018). The INCEpTION Platform : Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics : System Demonstrations*, p. 5–9 : Association for Computational Linguistics.

KLUYVER T., RAGAN-KELLEY B., PÉREZ F., GRANGER B., BUSSONNIER M., FREDERIC J., KELLEY K., HAMRICK J., GROUT J., CORLAY S., IVANOV P., AVILA D., ABDALLA S. & WILLING C. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. In F. LOIZIDES & B. SCHMIDT, Éds., *Positioning and Power in Academic Publishing : Players, Agents and Agendas*, p. 87 – 90 : IOS Press.

LAFON P. (1984). *Dépouillements et statistiques en lexicométrie*. Volume 24 de Travaux de linguistique quantitative. Genève : Paris : Slatkine ; Champion.

LANDRAGIN F., POIBEAU T. & VICTORRI B. (2012). ANALEC : a New Tool for the Dynamic Annotation of Textual Data. In *Proceedings of Language Resources and Evaluation (LREC 2012)*, p. 357–362, Istanbul, Turkey.

LEBART L., SALEM A. & BERRY L. (1998). *Exploring Textual Data*. Text, speech, and language technology. Kluwer Academic.

MCKINNEY W. (2010). Data Structures for Statistical Computing in Python. p. 56–61, Austin, Texas. DOI : 10.25080/Majora-92bf1922-00a.

MELLET S. (2002). Corpus et recherches linguistiques : Introduction. *Corpus*, (1). DOI : 10.4000/corpus.7.

MORETTI F. (2013). *Distant reading*. London ; New York : Verso.

PAUMIER S., GUENTHNER F., LAPORTE E., MALCHOK F., MARSCHNER C., MARTINEAU C., MARTÍNEZ C., MAUREL D., NAGEL S., NEME A., PETIT M., STIEHLER J. & VOLLANT G. (2021). UNITEX 3.3 Manuel d'utilisation.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

RASTIER F. (2001). *Arts et sciences du texte*. Formes sémiotiques. Paris : Presses universitaires de France, 1re édition.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the Conference on New Methods in Language Processing*, Manchester, UK.

SILBERZTEIN M. (2016). *Formalizing natural languages : the NooJ approach*. Collection Science cognitive et management des connaissances. London Hoboken : ISTE John Wiley and Sons.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). BRAT : a Web-based Tool for NLP-Assisted Text Annotation. In F. SEGOND, Éd., *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 102–107, Avignon, France : Association for Computational Linguistics.

WIDLÖCHER A. & BILHAUT F. (2008). Articulation des traitements en TAL. Principes méthodologiques et mise en œuvre dans la plate-forme LinguaStream. *Revue TAL (Traitement Automatique des Langues)*, **49**(2), 73–101. Place : France Publisher : ATALA (Association pour le Traitement Automatique des Langues).

WIDLÖCHER A. & MATHET Y. (2012). The Glozz Platform : a Corpus Annotation and Mining Tool. In C. CONCOLATO & P. SCHMITZ, Éds., *ACM Symposium on Document Engineering (DocEng'12)*, p. 171–180, Paris, France : ACM.

Annexes - Quelques illustrations du cadre proposé par Skhólion

Les exemples fournis ci-après ont pour unique objectif de donner une idée des informations accessibles depuis l'API et de la relative simplicité de mise en œuvre des traitements et des outils de visualisation proposés. Les algorithmes présentés ne sont pas toujours optimaux mais permettent d'illustrer en particulier l'exploitation des niveaux de segmentation et la possiblité de s'appuyer sur les annotations antérieurement produites.

```python
#
# Affichage simple d'un texte (CorpusItem) issu du corpus.
#

from skholion.corpus.map import CorpusMap
from skholion.xxml.navigator import Navigator
from skholion.gui.browser import Browser

corpus_map = CorpusMap("./corpus_data/")
fortune = corpus_map.get_corpus_item_by_key_name("Zola_La_Fortune_des_Rougon")

output_dir = "navigator/"
navigator_dir = output_dir + fortune.get_file_name_prefix()
navigator = Navigator(xml_corpus_item_source_path=fortune.get_html_quick_view_full_path(),
                      navigator_dir_output_path=navigator_dir)
navigator.make_all()

browser = Browser()
browser.open_local_path(navigator.get_main_file_path())

```

FIGURE 2 – Visualisation simple d'un item de corpus dans un navigateur web

```python
#
# Annotation par expressions régulières simples appliquées au contenu textuel d'un unique
# CorpusItem, puis affichage du texte annoté et des représentations symboliques associées.
#

from skholion.corpus.map import CorpusMap
from skholion.corpus.quickview import CorpusItemQuickView
from skholion.gui.color import ColorManager
from skholion.xxml.annotation.offset import OffsetAnnotator, OffsetAnnotation
from skholion.metamodel.characterization import Characterization
from skholion.xxml.navigator import Navigator
from skholion.gui.browser import Browser

import re

corpus_map = CorpusMap("./corpus_data/")
fortune = corpus_map.get_corpus_item_by_key_name("Zola_La_Fortune_des_Rougon")
fortune_text = fortune.get_anchored_text()
fortune_text_content = fortune_text.get_content()

offset_annotator = OffsetAnnotator(fortune.get_html_quick_view_full_path(),
                                   CorpusItemQuickView.CHARACTERIZATIONS_INJECTION_XPATH)

color_manager = ColorManager()

searched_items = ["(Rougon)", "(Macquart)", "(Lantier)", "(Mouret)", "(Saccard)", "(Coupeau)", "(Quenu)"]
for si in searched_items :
    patron = re.compile(si)
    for occurrence_position, match in enumerate(patron.finditer(fortune_text_content)):
        start_offset, end_offset = match.span()
        color = color_manager.get_color(match.group(1))
        characterization = Characterization("Famille", {"forme": match.group(1),
                                                        "occurrence" : str(occurrence_position+1),
                                                        "contexte":{"gauche":fortune_text_content[start_offset-100:start_offset],
                                                                    "droite":fortune_text_content[end_offset:end_offset+100]}})
        annotation = OffsetAnnotation(annotation_context_anchored_item=fortune_text,
                                      annotation_start_offset_in_context=start_offset,
                                      annotation_end_offset_in_context=end_offset,
                                      annotation_characterization=characterization,
                                      annotation_xml_type="span",
                                      annotation_xml_attributes={"style":"background-color:%s;" % color},
                                      annotation_text_content=match.group(1))
        offset_annotator.add_annotation(annotation)

offset_annotator.annotate()
output_xhtml_path = "/tmp/corpus_item.tmp.xhtml"
offset_annotator.dump(output_xhtml_path)

color_map_path = "/tmp/color_map.tmp.xhtml"
color_manager.write_html_color_map_file(output_path=color_map_path)

output_dir = "navigator/"
navigator_dir = output_dir + fortune.get_file_name_prefix()
navigator = Navigator(xml_corpus_item_source_path=output_xhtml_path,
                      offset_annotation_set=offset_annotator.offset_annotations_inserted_with_success,
                      color_map_source_path=color_map_path,
                      navigator_dir_output_path=navigator_dir)
navigator.make_all()

browser = Browser()
browser.open_local_path(navigator.get_main_file_path())

```

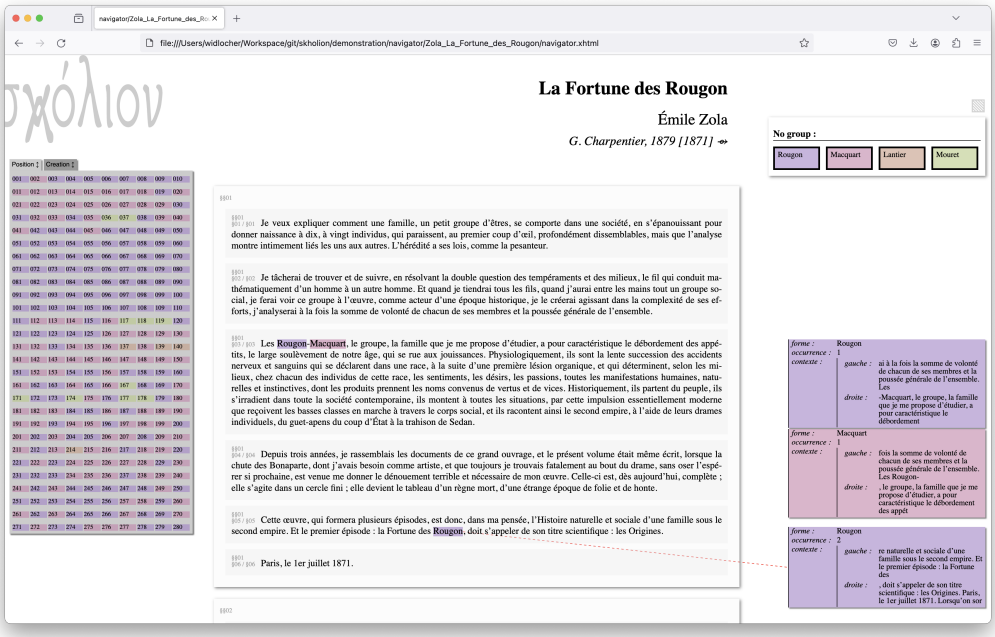


FIGURE 3 – Annotation par expressions régulières au niveau caractère, puis visualisation des annotations et des structures de traits associées

```python
#
# Annotation par expressions régulières simples appliquées au contenu textuel des phrases
# d'un ensemble de CorpusItems, puis affichage à l'aide d'un CorpusNavigator permettant
# de naviguer entre les textes.
#

from skholion.corpus.map import CorpusMap
from skholion.corpus.quickview import CorpusItemQuickView
from skholion.xxml.annotation.offset import OffsetAnnotator, OffsetAnnotation
from skholion.metamodel.characterization import Characterization
from skholion.xxml.navigator import Navigator, CorpusNavigator
from skholion.gui.color import ColorManager
from skholion.gui.browser import Browser

import re

corpus_map = CorpusMap("./corpus_data/")
germinal = corpus_map.get_corpus_item_by_key_name("Zola_Germinal")
fortune = corpus_map.get_corpus_item_by_key_name("Zola_La_Fortune_des_Rougon")
argent = corpus_map.get_corpus_item_by_key_name("Zola_L_Argent")
corpus = [germinal, fortune, argent]

corpus_level_color_manager = ColorManager()

for corpus_item in corpus :

    corpus_item_level_color_manager = ColorManager()
    text = corpus_item.get_sentences_segmented_anchored_text()
    offset_annotator = OffsetAnnotator(corpus_item.get_html_quick_view_full_path(),
                                       CorpusItemQuickView.CHARACTERIZATIONS_INJECTION_XPATH)

    for section in text:
        for paragraph in section:
            for sentence in paragraph:
                sentence_content = sentence.get_content()
                searched_items = ["(Rougon)", "(Macquart)", "(Lantier)", "(Mouret)", "(Saccard)", "(Coupeau)", "(Quenu)"]
                for si in searched_items :
                    patron = re.compile(si)
                    for match in patron.finditer(sentence_content):
                        color = corpus_level_color_manager.get_color(match.group(1))
                        start_offset, end_offset = match.span()
                        characterization = Characterization("Famille", {"forme": match.group(1), "phrase" : sentence_content})
                        annotation = OffsetAnnotation(annotation_context_anchored_item=sentence,
                                                      annotation_start_offset_in_context=start_offset,
                                                      annotation_end_offset_in_context=end_offset,
                                                      annotation_characterization=characterization,
                                                      annotation_xml_type="span",
                                                      annotation_xml_attributes={"style":"background-color:%s;" % color},
                                                      annotation_text_content=match.group(1))
                        offset_annotator.add_annotation(annotation)

                searched_items = ["(hériter)", "(hérédit\\we)", "(famille)", "(fille)", "(fils)", "(mère)", "(père)"]
                for si in searched_items :
                    patron = re.compile(si)
                    for match in patron.finditer(sentence_content):
                        color = corpus_item_level_color_manager.get_color(match.group(1), group_key="Parenté")
                        start_offset, end_offset = match.span()
                        characterization = Characterization("Famille", {"forme": match.group(1), "motif" : si})
                        annotation = OffsetAnnotation(annotation_context_anchored_item=sentence,
                                                      annotation_start_offset_in_context=start_offset,
                                                      annotation_end_offset_in_context=end_offset,
                                                      annotation_characterization=characterization,
                                                      annotation_xml_type="span",
                                                      annotation_xml_attributes={"style":"background-color:%s;" % color},
                                                      annotation_text_content=match.group(1))
                        offset_annotator.add_annotation(annotation)

    offset_annotator.annotate()
    output_xhtml_path = "/tmp/corpus_item.tmp.xhtml"
    offset_annotator.dump(output_xhtml_path)

    corpus_item_level_color_map_path = "/tmp/color_map.tmp.xhtml"
    corpus_item_level_color_manager.write_html_color_map_file(output_path=corpus_item_level_color_map_path)

    output_dir = "navigator/"

    navigator_dir = output_dir + corpus_item.get_file_name_prefix()
    navigator = Navigator(xml_corpus_item_source_path=output_xhtml_path,
                          offset_annotation_set=offset_annotator.offset_annotations_inserted_with_success,
                          color_map_source_path=corpus_item_level_color_map_path,
                          navigator_dir_output_path=navigator_dir)
    navigator.make_all()

corpus_level_color_map_path = "/tmp/color_map.tmp.xhtml"
corpus_level_color_manager.write_html_color_map_file(output_path=corpus_level_color_map_path)
corpus_navigator = CorpusNavigator(navigator_source_and_output_path=output_dir,
                                   color_map_source_path=corpus_level_color_map_path)
corpus_navigator.make_all()

browser = Browser()
browser.open_local_path(corpus_navigator.get_main_file_path())
```

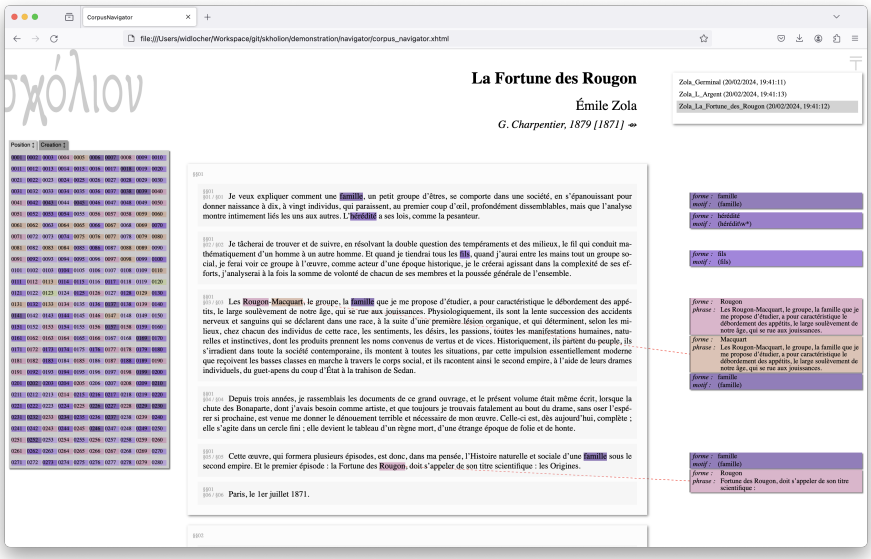


FIGURE 4 – Application d'expressions régulières au contenu textuel des phrases d'un corpus composé de plusieurs textes

```python
#
# On annote tous les verbes à l'indicatif, puis on annote les séquences ininterrompues
# de verbes au présent.
#

from skholion.corpus.map import CorpusMap
from skholion.corpus.quickview import CorpusItemQuickView
from skholion.metamodel.characterization import Characterization
from skholion.xxml.annotation.offset import OffsetAnnotator
from skholion.linguistics.partofspeech import PartOfSpeech, VerbalPartOfSpeech
from skholion.analysis.anchoreditemannotator import AnchoredItemAnnotator
from skholion.analysis.regexoa import RegexOnAnnotationsSolver
from skholion.analysis.unitgrouper import UnitGrouper
from skholion.gui.color import ColorManager
from skholion.xxml.navigator import Navigator
from skholion.gui.browser import Browser

corpus_map = CorpusMap("./corpus_data/")
corpus_item = corpus_map.get_corpus_item_by_key_name("Zola_La_Fortune_des_Rougon")
text = corpus_item.get_sentences_and_tokens_segmented_pos_tagged_anchored_text()
quick_view_html_path = corpus_item.get_html_quick_view_full_path()

offset_annotator = OffsetAnnotator(quick_view_html_path, CorpusItemQuickView.CHARACTERIZATIONS_INJECTION_XPATH,
                                   characterization_injection_mode=OffsetAnnotator.CHARACTERIZATION_INJECTION_MODE_COMPACT)

color_manager = ColorManager()
for section in text :
    for paragraph in section :
        for sentence in paragraph :
            for token in sentence :
                if (token.part_of_speech[PartOfSpeech.FEATURE_NAME_TAG]==PartOfSpeech.TAG_VERB
                and token.part_of_speech[VerbalPartOfSpeech.FEATURE_NAME_MOOD]==VerbalPartOfSpeech.MOOD_INDICATIVE):
                    main_tense = token.part_of_speech[VerbalPartOfSpeech.FEATURE_NAME_MAIN_TENSE]
                    characterization = Characterization("Verbe", {"type":"verb", "content":token.content, "main-tense": main_tense})
                    color = color_manager.get_color(main_tense)
                    token_annotation = AnchoredItemAnnotator.get_unit_from_anchored_item(anchored_item=token,
                                                                                         annotation_characterization=characterization,
                                                                                         annotation_xml_attributes={"style": f"background-
color:{color};"})
                    offset_annotator.add_annotation(token_annotation)

roas = RegexOnAnnotationsSolver(offset_annotator.offset_annotations)
annotations_input_sequence = roas.prepare_annotation_set()
pattern = "Annotation:{{@main-tense:%s}}+" % VerbalPartOfSpeech.MAIN_TENSE_PRESENT
pattern = roas.prepare_pattern(pattern)
matches = roas.find_all(pattern, annotations_input_sequence)

color = color_manager.get_color("present-sequence")
for match in matches :
    unit_1, unit_2, match_charaterization = match
    characterization = Characterization("Present-sequence", {"type":"present-sequence"})
    new_annotation = UnitGrouper.get_unit_from_unit_to_unit(unit_1=unit_1,
                                                            unit_2=unit_2,
                                                            annotation_characterization=characterization,
                                                            annotation_xml_attributes={"style":f"background-color:{color};padding:2px;"})
    offset_annotator.add_annotation(new_annotation)

offset_annotator.annotate()

output_xhtml_path = "/tmp/corpus_item.tmp.xhtml"
offset_annotator.dump(output_xhtml_path)
color_map_path = "/tmp/color_map.tmp.xhtml"
color_manager.write_html_color_map_file(output_path=color_map_path)
output_dir = "navigator/"
navigator_dir = output_dir + corpus_item.get_file_name_prefix()
navigator = Navigator(xml_corpus_item_source_path=output_xhtml_path,
                      offset_annotation_set=offset_annotator.offset_annotations_inserted_with_success,
                      color_map_source_path=color_map_path,
                      navigator_dir_output_path=navigator_dir)
navigator.make_all()
browser = Browser()
browser.open_local_path(navigator.get_main_file_path())
```

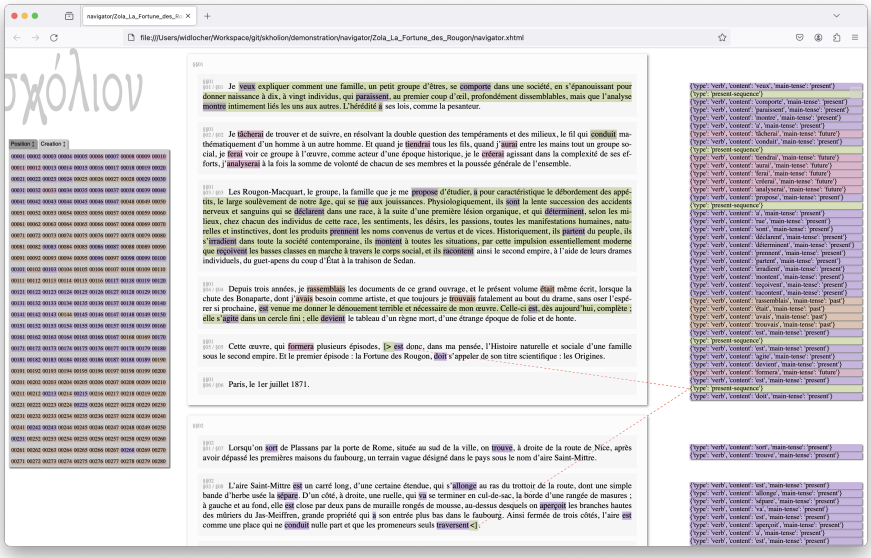


FIGURE 5 – Annotation de tokens POS-taggés, puis application de motifs REGEXOA (*REGEX On Annotations*) pour l'annotation d'une séquence d'annotations de plus bas niveau

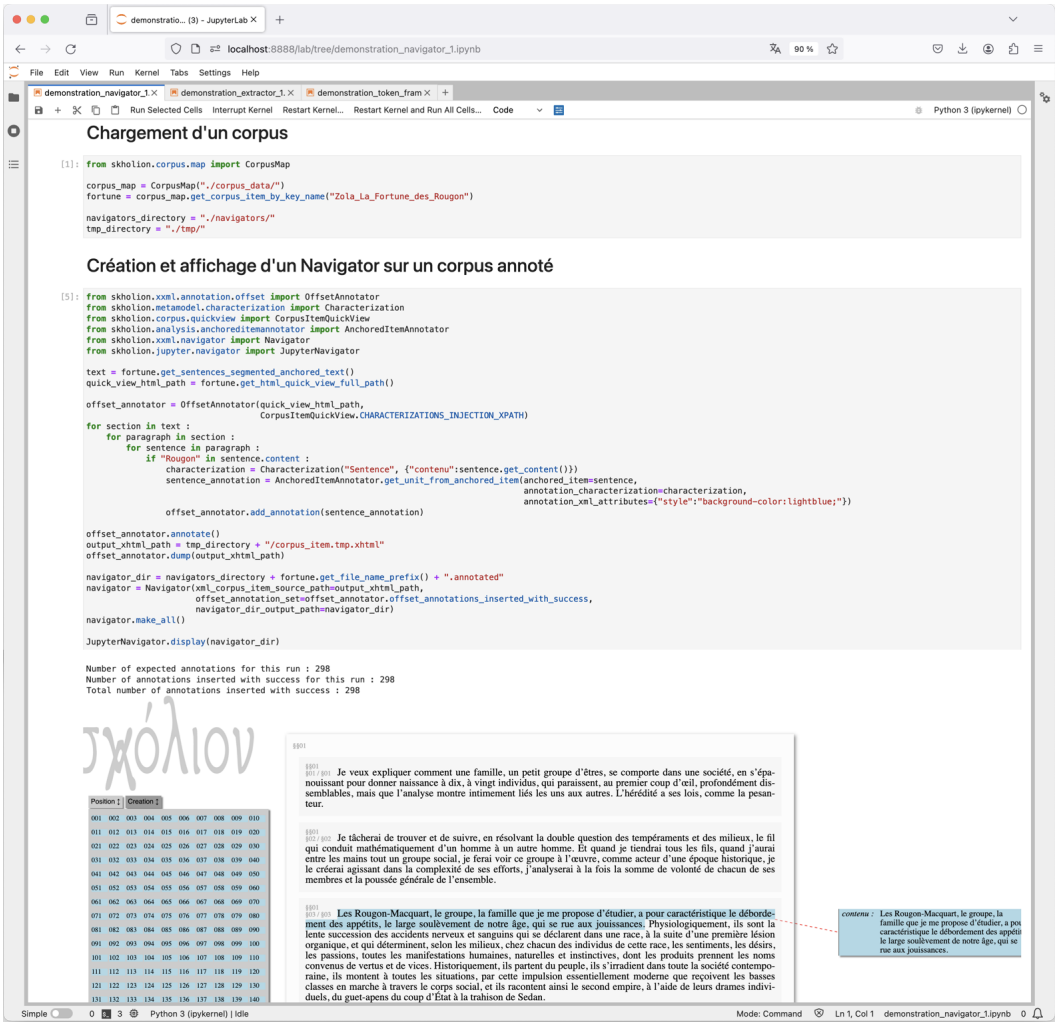


FIGURE 6 – Annotation de phrases et visualisation du texte annoté dans JupyterLab

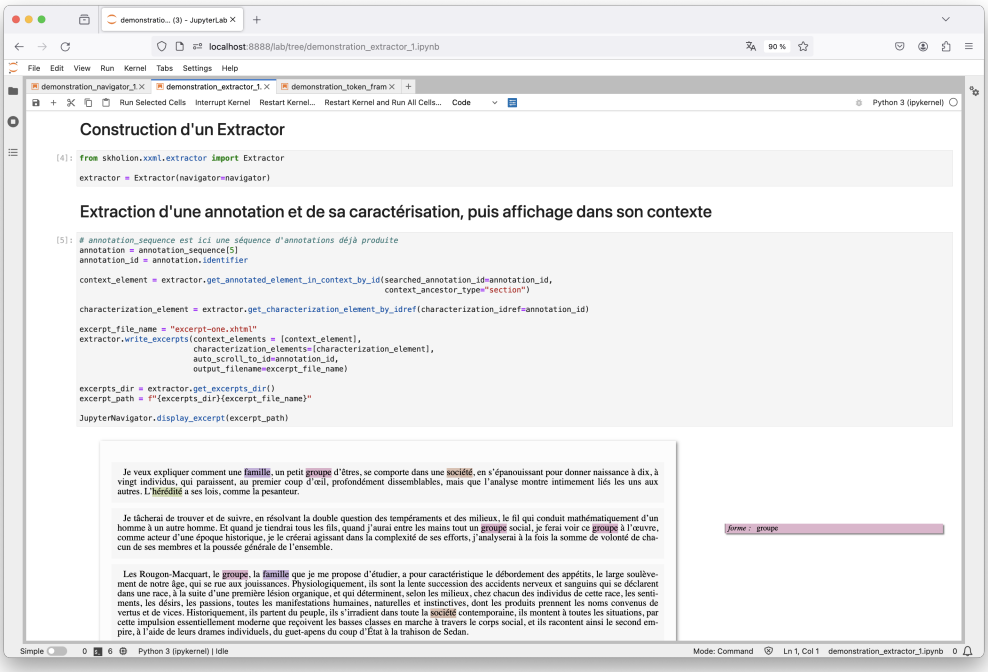


FIGURE 7 – Extraction d'une annotation et visualisation dans son contexte

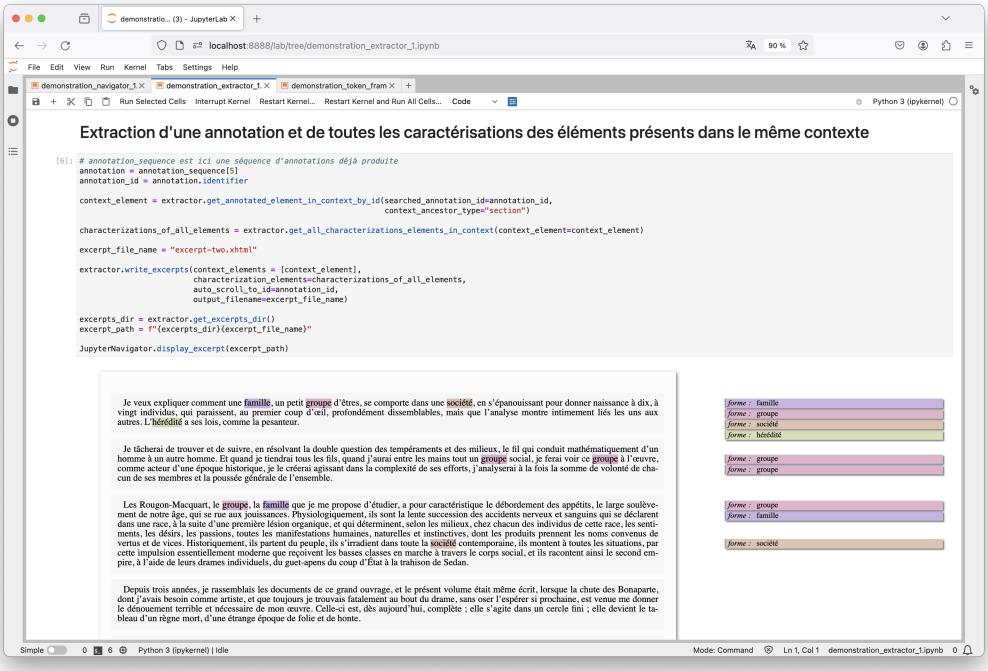


FIGURE 8 – Extraction d'une annotation et visualisation dans son contexte, en intégrant les descriptions des autres objets présents dans ce contexte

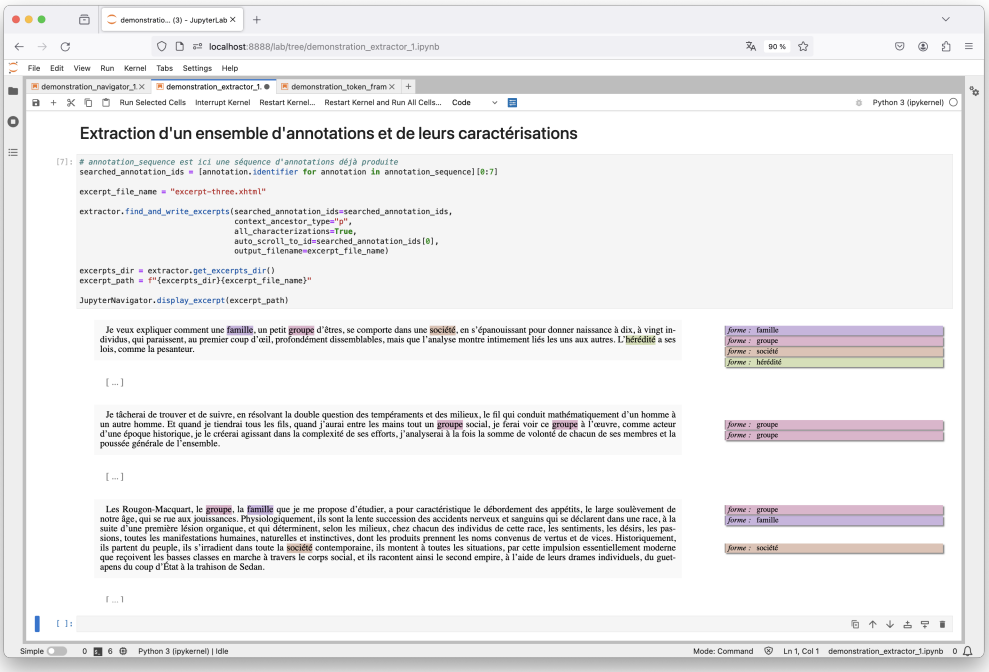


FIGURE 9 – Extraction d'une sélection d'annotations

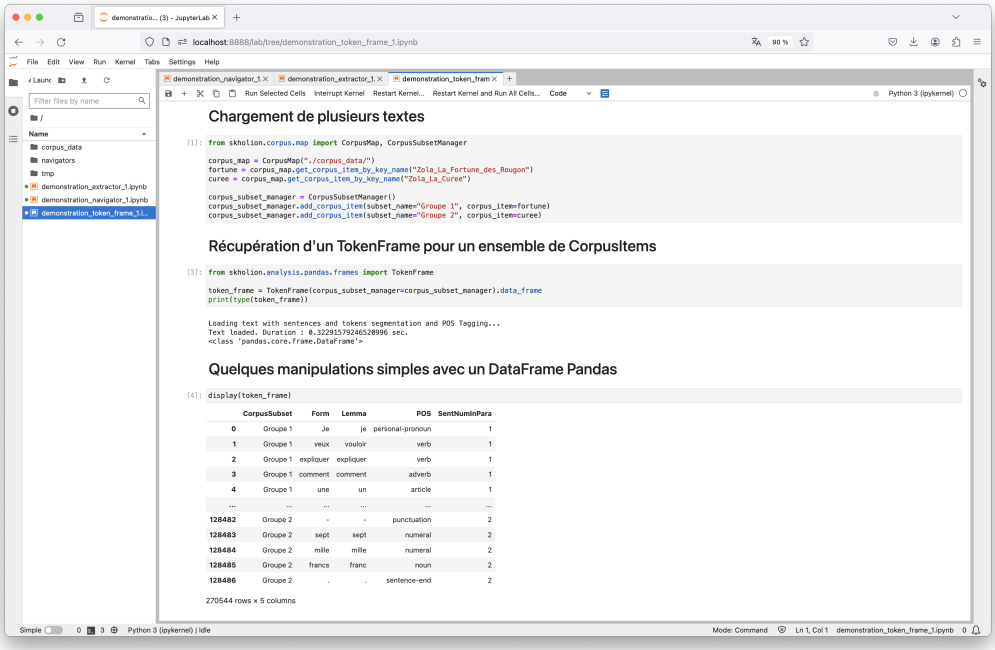


FIGURE 10 – Récupération d'un TokenFrame et manipulation du DataFrame Pandas sous-jacent

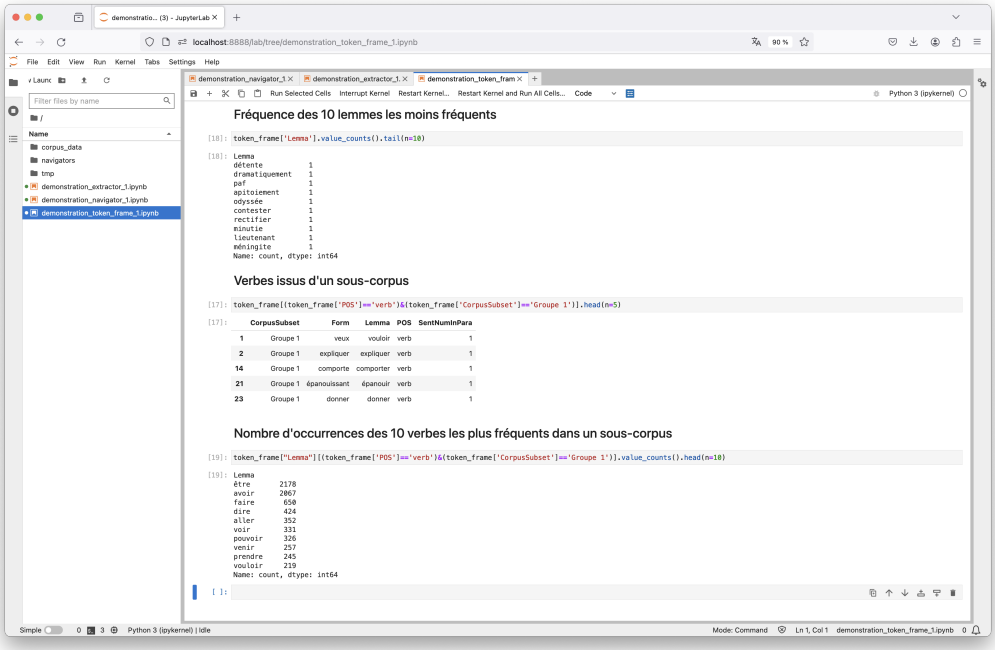


FIGURE 11 – Exemples de manipulation d'un DataFrame avec Pandas

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Repérage et caractérisation automatique des émotions dans des textes : traiter aussi leurs modes d'expression indirects

Aline Étienne[1] Delphine Battistelli[1] Gwénolé Lecorvé[2]
(1) Univ. Paris-Nanterre, CNRS, MoDyCo – Nanterre, France (2) Orange – Lannion, France
acm.etienne@gmail.com, del.battistelli@gmail.com, gwenole.lecorve@orange.com

RÉSUMÉ

Cet article présente un modèle capable de prédire (A) si une phrase contient l'expression d'une émotion, (B) selon quel(s) mode(s) cette émotion est exprimée, (C) si elle est basique ou complexe, et (D) quelle est sa catégorie exacte. Notre principale contribution est d'intégrer le fait qu'une émotion puisse s'exprimer selon différents modes : depuis un mode direct, essentiellement lexicalisé, jusqu'à un mode plus indirect, où des émotions vont être seulement suggérées, mode dont les approches en TAL ne tiennent généralement pas compte. Nos expériences sur des textes en français pour les enfants mènent à des résultats tout à fait acceptables en comparaison de ce sur quoi des annotateurs humains experts en psycholinguistique s'accordent et à des résultats meilleurs que ceux produits par GPT-3.5 via du *prompting*. Ceci offre une perspective intéressante de prise en compte des émotions comme facteur d'analyse automatique de la complexité dans les textes, cadre plus général de nos travaux.

ABSTRACT

Spotting and characterization emotions in texts : also addressing indirect modes of expression.

This paper presents a model able to predict whether a sentence expresses an emotion, (B) the mode(s) in which it is expressed, (C) whether it is basic or complex, and (D) its emotional category. One of our major contributions is to integrate the fact that an emotion can be expressed in different modes : from a direct mode, essentially lexicalized, to a more indirect mode, where emotions will only be suggested, a mode that NLP approaches generally don't take into account. Experiments on French texts intended for children show acceptable results compared to the human annotators' agreement, and the lower results obtained by a large language model without fine-tuning. They offer an interesting perspective on how to take emotions into account as a factor in the automatic analysis of complexity in texts, which is the more general framework of our work.

MOTS-CLÉS : Émotions, mode d'expression, complexité, étiquetage multi-tâches.

KEYWORDS: Emotions, expression mode, complexity, multi-task labeling.

1 Introduction

En traitement automatique des langues (TAL), la tâche de repérage des émotions est assez souvent abordée à propos de textes produits dans un cadre d'interactions ou conversations (par ex. (Poria *et al.*, 2019)), orales comme écrites (chats, forums, tweets) avec des jeux de données souvent multimodaux (par ex. (Busso *et al.*, 2008; Poria *et al.*, 2018; Chen *et al.*, 2018). La visée est généralement de pouvoir

Le modèle et les données sont disponibles sur https://huggingface.co/TextToKids.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 650–676.

identifier les émotions ressenties par les locuteurs en situation dialogique. L'analyse des émotions dans des textes autres que de type conversationnel, comme par exemple des textes journalistiques et encyclopédiques ou encore des romans est un domaine moins développé en TAL. Elle relève de fait d'un autre type de visée qui n'est plus celui de caractériser l'état émotionnel du locuteur mais plutôt des personnages qui composent ces textes. Comme le soulignent certains travaux en psycholinguistique, les émotions sont dédiées dans ces types de textes - de manière plus ou moins contrôlée par le scripteur - à capter l'attention du lecteur. Elles permettent en outre de créer du lien entre les situations décrites et donc sont un facteur clé dans la compréhension (par ex. pour les enfants dans Davidson *et al.*, 2001). Encore faut il que ces émotions elles-mêmes soient identifiées et comprises. Là se joue une question qui devient alors celle d'envisager les émotions comme un facteur de complexité, au moins relative pour reprendre la terminologie de Ehret *et al.* (2023), c'est-à-dire qui tient compte de la difficulté perçue par les locuteurs en termes d'apprentissage ou de compréhension de la langue. Un texte sera ainsi d'autant plus complexe qu'il contient d'émotions considérées comme complexes par un type de locuteur donné. Dans le cas d'enfants par exemple, on sait que certaines catégories émotionnelles ne sont pas accessibles avant un certain âge et que leur mode d'expression (direct *vs.* indirect ou implicite) joue également un rôle dans l'accessibilité à leur signification.

De ces réflexions sur la question des émotions comme facteur de complexité, nous avons choisi d'orienter nos travaux vers une meilleure prise en compte de la diversité des modes d'expression des émotions. Nous présentons un modèle qui introduit ainsi la notion de mode d'expression en plus des informations habituelles sur les catégories émotionnelles (par ex., joie, peur, etc.). En pratique, le modèle classe les émotions dans les textes à travers quatre tâches : (A) prédire si une phrase contient ou non une émotion ; (B) si oui, comment elle est exprimée (le *mode*) ; (C) s'il s'agit d'une catégorie d'émotion basique ou complexe ; et (D) dans quelle catégorie émotionnelle elle se situe. Le modèle est construit à partir du modèle CamemBERT (Martin *et al.*, 2020) et de données dérivées d'un schéma d'annotation psycho-linguistiquement motivé comprenant différents types de sources. L'évaluation montre que le modèle proposé surpasse des approches fondées sur des resources expertes, des architectures non neuronales (SVM et XGBoost), et de l'inférence en-contexte (*in-context learning* d'un grand modèle de langue (GPT-3.5). De plus, l'évaluation humaine menée en complément montre que les erreurs de prédiction faites par le modèle proposé se situent généralement dans les mêmes proportions que celles faites par les humains.

Dans la suite, la section 2 dresse une panorama de la littérature dédiée à l'identification des émotions dans les textes, notamment en TAL. Les sections 3, 4 et 5 détaillent respectivement les tâches traitées, les données associées, puis le modèle proposé. La section 6 rapporte enfin les expériences et résultats.

2 Cadre d'analyse des émotions et travaux connexes

Cette section fournit un bref aperçu du cadre d'analyse des émotions dans lequel nous nous situons et qui justifie le choix du schéma et des données (annotées avec ce schéma). Elle positionne également notre travail parmi les études en TAL sur l'identification automatique des émotions.

2.1 L'analyse des émotions en tant que facteur de complexité d'un texte

En psycho-linguistique, le rôle-clé des émotions des personnages sur la compréhension des textes est un sujet bien documenté (par ex. Dijkstra *et al.*, 1995; Dyer, 1983). Parmi les travaux récents,

deux facteurs d'influence ont été mis en évidence dans la compréhension des émotions par les enfants, et donc des textes eux-mêmes : le *type d'émotion* exprimé, basique ou complexe – les émotions complexes (par ex. la fierté, la honte) étant plus difficiles à saisir car elles nécessitent une connaissance des normes sociales – (Davidson, 2006; Blanc & Quenette, 2017) ; ainsi que *la manière dont les émotions sont exprimées* (Creissen & Blanc, 2017)), directement *via* une étiquette émotionnelle, indirectement à travers la mention d'un comportement émotionnel, ou à travers la description d'une situation émotionnelle, cette dernière étant la plus difficile à comprendre. Bien sûr, la notion de catégorie émotionnelle est également abordée en psycho-linguistique, et il a été montré que certaines catégories prennent plus de temps à être maîtrisées par les enfants (par ex. Baron-Cohen *et al.*, 2010).

Du côté du TAL, plusieurs travaux (voir par ex. Bostan & Klinger, 2018; Acheampong *et al.*, 2020; Öhman, 2020) soulignent la grande hétérogénéité des schémas d'annotation des émotions – et des corpus annotés –, mettant ainsi clairement en évidence la difficulté de modéliser les émotions et *in fine* de les analyser. Cette hétérogénéité concerne tous les aspects de ces travaux : depuis les notions (par ex. le nombre et les types de catégories émotionnelles) et le type de données étudiées (journaux, tweets...) jusqu'aux procédures d'annotation (*crowdsourcing*, annotation par des experts) et méthodes d'évaluation mises en œuvre (par ex. avec ou sans accord entre annotateurs). Bien que certains travaux s'efforcent de prendre en compte des ensembles plus larges de notions et d'indices linguistiques pour analyser les émotions (par ex. Casel *et al.*, 2021; Kim & Klinger, 2019), le concept le plus couramment utilisé reste la notion de *catégorie émotionnelle*, souvent abordée à travers une liste d'émotions de base introduite soit par Ekman (1992) (colère, dégoût, peur, joie, tristesse et surprise) ou Plutchik (1980) (catégories d'Ekman, anticipation et confiance), avec un accent sur une manière d'exprimer les émotions : le lexique émotionnel. Comme souligné dans (Klinger, 2023) et dans (Troiano *et al.*, 2023), quelques approches très récentes en TAL visent cependant à acquérir une compréhension plus profonde des unités textuelles qui soutiennent l'évocation d'émotions en dehors des termes lexicaux directement émotionnels (par ex. «heureux», «colère»). Ces approches s'inspirent alors de modèles psychologiques et/ou linguistiques des émotions. Nous adoptons ici la même approche car nous visons à la fois les modes d'expression directs et indirects des émotions dans les textes. Comme Troiano *et al.* (2023), nous cherchons à évaluer à quel point les modèles computationnels peuvent capter des émotions exprimées indirectement (par ex., via la description de situations qui sont associées à des émotions eu égard à des normes sociales et à des conventions). Plus précisément, notre travail fait le choix du cadre proposé par Etienne *et al.* (2022). Ce cadre repose sur un schéma d'annotation détaillé des émotions et propose un corpus annoté manuellement dont la taille est compatible avec les expériences d'apprentissage automatique. À notre connaissance, c'est le seul travail qui correspond à l'objectif explicite d'analyser les émotions dans les textes en traitant à la fois les modes d'expression directs et indirects.

2.2 Identification automatique des émotions

En TAL, l'analyse des émotions dans les textes est généralement traitée comme une tâche de classification. L'hétérogénéité précédemment évoquée des schémas d'annotation et des corpus annotés se reflète alors dans la diversité des classes prédites, de la granularité des éléments à classifier et des méthodes pour développer et évaluer les classificateurs.La manière dont les résultats sont présentés varie donc également d'un article à l'autre, ce qui rend la comparaison des performances plus difficile.

L'accent est souvent mis sur la classification en émotions de base (Strapparava & Mihalcea, 2007; Mohammad, 2012; Abdaoui *et al.*, 2017; Demszky *et al.*, 2020; Öhman *et al.*, 2020; Bianchi *et al.*,

2021), bien que certains travaux utilisent un mélange d'émotions de base et complexes (Balahur *et al.*, 2012; Fraisse & Paroubek, 2015; Abdaoui *et al.*, 2017; Mohammad *et al.*, 2018; Liu *et al.*, 2019; Demszky *et al.*, 2020). De plus, il existe une longue histoire de construction et d'utilisation de lexiques émotionnels et la diversité des marqueurs linguistiques des émotions n'est pas systématiquement prise en compte, bien qu'elle soit mentionnée dans plusieurs travaux (Alm *et al.*, 2005; Mohammad, 2012; Kim & Klinger, 2018; Demszky *et al.*, 2020)). Certains travaux étudient malgré tout d'autres moyens d'expression. par ex., Kim & Klinger (2019) analysent les expressions non verbales des émotions par les personnages dans un corpus de fanfictions (par ex. les regards, les gestes). Balahur *et al.* (2012) visent à détecter les émotions indirectes. Ces travaux ont pour limite de ne se focaliser à chaque fois que sur un seul mode d'expression, laissant ainsi de côté les complémentarités entre modes. Pour leur part, sur la base du modèle de processus de composants émotionnels de Scherer (2005), Casel *et al.* (2021) ont annoté puis prédit plusieurs composantes des émotions, tels que les symptômes physiologiques et les expressions motrices des émotions, ou l'évaluation cognitive des événements. Bien que (Casel *et al.*, 2021) traitent d'un ensemble plus large d'indices, ceux-ci ne sont pas rigoureusement motivés linguistiquement. Par conséquent, en s'appuyant sur (Etienne *et al.*, 2022), la véritable originalité de notre travail réside dans la prise en compte de différents modes d'expression des émotions.

Historiquement, les modèles *Support Vector Machine* (SVM) ont été largement utilisés pour classer des phrases (Aman & Szpakowicz, 2007; Mohammad, 2012)) ou des textes (Abdaoui *et al.*, 2017; Balahur *et al.*, 2012; Fraisse & Paroubek, 2015; Mohammad, 2012) selon la catégorie émotionnelle qu'ils expriment. Jusqu'à l'avènement des plongements, les entrées étaient principalement symboliques : sacs de mots ou *n*-grammes, caractéristiques basées sur des ressources émotionnelles telles que WordNetAffect (Aman & Szpakowicz, 2007; Balahur *et al.*, 2012; Strapparava & Mihalcea, 2007) ou lexiques émotionnels (Strapparava & Mihalcea, 2007; Abdaoui *et al.*, 2017; Kim & Klinger, 2018). Aujourd'hui, les réseaux neuronaux (Kim & Klinger, 2018) et les architectures Transformer (Liu *et al.*, 2019; Demszky *et al.*, 2020; Öhman *et al.*, 2020; Bianchi *et al.*, 2021) dominent évidemment l'état de l'art. En ce qui concerne le français, aucun modèle basé sur l'architecture Transformer n'a encore été proposé à notre connaissance.

3 Tâches

Construit dans la perspective globale de permettre l'analyse des émotions en tant que facteur de complexité, notre travail tient ainsi compte de deux éléments clé pour aborder la complexité d'une émotion : sa catégorie et son mode d'expression. L'objectif est de proposer un modèle Transformer pour 4 tâches de classification (notés A, B, C et D) au niveau de la *phrase*, par opposition au niveau du texte (cela peut, par exemple, permettre d'étudier comment la présence d'émotions évolue le long d'un texte). Les phrases peuvent contenir plusieurs émotions et les classifications sont donc multi-étiquettes. Toutes les tâches sont apprises ensemble, menant à un modèle unique.

Tâche A : Présence d'émotion La première tâche vise à prédire la présence d'informations émotionnelles dans une phrase donnée (prédiction binaire).

Tâche B : Mode d'expression Le mode d'expression se concentre sur les moyens linguistiques utilisés pour transmettre la présence d'une émotion dans un texte. Suivant Etienne *et al.* (2022), 4 modes sont considérés : les **émotions désignées** directement indiquées par un terme du lexique émotionnel (par ex. *heureux*, *effrayé*) ; les **émotions comportementales** qui s'appuient sur la

	Phrase à classifier (+ phrases voisines)	emot.	comport.	désignée	montrée	suggérée	basique	complexe	admiration	autre	colère	culpabilité	dégoût	embarras	fierté	jalousie	joie	peur	surprise	tristesse
(1)	Nicolas Hulot n'appartient à aucun parti politique. **Il a refusé trois fois le poste de ministre de l'Ecologie avant d'accepter la proposition d'Emmanuel Macron.** Mais ça ne s'est pas très bien passé.																			
(2)	Les extraterrestres ont une autre grande qualité : ils sont mystérieux, viennent d'un monde inconnu et donc, ils font peur ! **Or, la peur est une émotion très puissante, qu'on ressent dans notre corps.** Quand on regarde des films effrayants, on peut éprouver un certain plaisir parce qu'on domine notre peur.	√		√			√											√		
(3)	Le jour du marché, le paysan vendit très bien son blé et sa paille, tandis que les diables ne tirèrent pas un liard de leurs tas de chaume. **– Tu m'as trompé, cria le diable.** Mais l'an prochain, il n'en sera pas de même.	√	√			√	√				√									
(4)	Cet été, Nolita a pour la première fois depuis longtemps dû manger une merguez, parce qu'il n'y avait rien d'autre. **« Ça m'a écœurée, et j'ai culpabilisé, raconte-t-elle.** Je me suis forcée. »	√		√			√	√				√	√							
(5)	Aux JO de Rome, l'événement historique a lieu lors du marathon : l'Ethiopien Abebe Bikila est le premier sportif venu d'Afrique noire à devenir champion olympique. **Surtout, il réussit l'exploit de gagner… pieds nus !** Il avait en effet pris l'habitude, chez lui en Ethiopie, de courir ainsi.	√			√	√	√	√						√		√		√	

TABLE 1 – Exemples de phrases en contexte et étiquettes de référence pour les tâches A, B, C et D.

description d'un comportement émotionnel, telles que des manifestations physiologiques (par ex. *pleurer*, *sourire*) ou d'autres comportements (par ex. *gifler quelqu'un*) ; les **émotions montrées** qui sont exprimées par des caractéristiques linguistiques de surface très hétérogènes des énoncés qui reflètent principalement l'état émotionnel du scripteur (par ex. interjections, phrases courtes) ; les **émotions suggérées** qui émanent de la description d'une situation généralement associée à un sentiment émotionnel d'après les normes sociales et conventions (par ex. *voir un bon ami après une longue période* suggère la joie).

Tâche C : Type d'émotion La tâche C vise à prédire la présence de types d'émotions *basique* et *complexe* (2 prédictions binaires simultanées). À notre connaissance, cette notion n'a encore jamais été étudiée en tant que telle dans l'analyse automatique des émotions (bien que les catégories émotionnelles *basiques* et *complexes* aient été utilisées en TAL (*cf.* section 2.2)). Cela est probablement dû au fait que le type d'une émotion exprimée est directement lié à sa catégorie émotionnelle. Cependant, le type d'émotion est en lui-même un marqueur de complexité, comme nous l'avons vu.

Tâche D : Catégorie émotionnelle Conformément à Etienne *et al.* (2022), la tâche D est conçue pour étiqueter 11+1 catégories émotionnelles, à savoir les 6 émotions de base d'Ekman (*colère*, *dégoût*, *peur*, *joie*, *tristesse* et *surprise*) et 5 émotions complexes (*admiration*, *embarras*, *culpabilité*, *jalousie* et *fierté*). Une dernière catégorie, nommée *autre*, est utilisée pour capturer les marqueurs qui expriment toute autre émotion (par ex. haine, mépris, amour, *etc.*).

4 Données

Notre corpus est dérivé de celui fourni dans (Etienne *et al.*, 2022). Il se compose de 1 594 textes en français (28K phrases, 515K mots) destinés aux enfants de 6 à 14 ans, répartis en 3 types : très principalement des textes journalistiques (91 % des phrases), des articles encyclopédiques (9 %) et des romans (1 %). Originellement, le corpus est accompagné d'annotations d'experts délimitant des unités émotionnelles (segments) dans les textes. Chaque unité est décrite par son mode d'expression et sa catégorie émotionnelle. Nous avons fusionné ces annotations depuis le niveau des segments vers le niveau de la phrase. Ainsi, une phrase donnée peut couvrir plusieurs unités émotionnelles. La présence d'émotions et les types d'émotions ont été dérivés des étiquettes de mode d'expression et

Tâche	Étiquettes	Prop. phr. (%)		
		entr.	dév.	test
(A) Prés. émotion	émotion	20,2	15,8	17,6
(B) Mode d'expression	comport.	4,6	3,6	4,3
	désigné	5,3	5,2	5,7
	montré	3,6	2,3	3,5
	suggéré	7,1	5,8	6,3
(C) Type d'émotion	basique	15,4	12,6	13,9
	complexe	2,0	2,1	2,3
(D) Catégorie émotionnelle	admiration	0,6	1,1	1,0
	autre	5,0	3,2	3,7
	colère	4,6	3,2	3,4
	culpabilité	0,1	0,0	0,1
	dégoût	0,2	0,3	0,2
	embarras	0,6	0,6	0,6
	fierté	0,7	0,4	0,9
	jalousie	0,0	0,0	0,0
	joie	3,2	2,3	3,6
	peur	3,8	3,3	3,8
	surprise	3,0	3,1	2,5
	tristesse	2,5	2,0	2,5

TABLE 2 – Répartition des étiquettes

Tâche	Modèle	Macro R	Macro P	Macro F1
(A) Présence d'une émotion	SVM	0,481	0,659	0,556
	XGBoost	0,223	0,700	0,338
	GPT-3.5	0,622	0,443	0,518
	notre modèle	**0,764**	**0,741**	**0,752**
(B) Mode d'expression	SVM	0,267	0,721	0,368
	XGBoost	0,218	**0,730**	0,314
	GPT-3.5	0,513	0,101	0,152
	notre modèle	**0,626**	0,665	**0,645**
(C) Type d'émotion	SVM	0,211	0,343	**0,261**
	XGBoost	0,120	**0,659**	0,200
	GPT-3.5	**0,756**	0,123	0,199
	notre modèle	0,557	0,662	**0,601**
(D) Catégorie émotionnelle	SVM	0,125	0,487	0,186
	XGBoost	0,192	**0,565**	0,272
	GPT-3.5	**0,697**	0,109	0,174
	notre modèle	0,397	0,463	**0,420**

TABLE 3 – Performances des modèles (moyennes sur 3 exécutions, tous les écarts-types sont inférieurs à $0,02$).

de catégorie émotionnelle. Au final, chaque phrase est associée à un vecteur de 19 booléens. Des exemples de phrases en contexte sont fournis dans le tableau 1.

Les données sont divisées en ensembles d'entraînement, de développement et de test (70/10/20% des phrases, respectivement), tel que toutes les phrases d'un texte se trouvent dans le même sous-ensemble, ceci afin d'éviter un biais d'entraînement sur les particularités des textes (par ex., le nom d'un personnage). Le tableau 2 présente la proportion des étiquettes au sein du corpus. Globalement, les proportions sont comparable d'un sous-ensemble à un autre. Plusieurs déséquilibres apparaissent au sein des tâches. **(A)** Seulement 15-20% des phrases sont émotionnelles. **(B)** Les modes d'expression sont assez uniformément répartis, 'désigné' étant le moins fréquent (3% des phrases) et *suggéré* le plus courant (6%). Les sommes des pourcentages de chaque mode sont supérieures aux pourcentages de l'étiquette *émotionnelle* car une phrase peut certaines émotions sont véhiculés par plusieurs modes et une phrase peut aussi contenir plusieurs unités émotionnelles dont les modes respectifs diffèrent. **(C)** Les étiquettes des types d'émotions sont très déséquilibrées, avec une nette dominance des émotions basiques. La catégorie émotionnelle 'autre' (tâche D) n'est associée à aucun type d'émotion, d'où le fait que la somme des pourcentages *basique* et *complexe* est inférieure à celle des phrases émotionnelles. **(D)** Les étiquettes des catégories émotionnelles sont déséquilibrées, avec des pourcentages toujours en dessous de 5% des phrases. Les catégories *colère*, *peur*, *joie*, *tristesse*, *surprise* et *autre* sont dominantes, alors que d'autres sont très rares (*dégoût*, *culpabilité* et *jalousie*).

5 Modèle proposé

Le modèle proposé résulte d'un affinage (*fine-tuning*) de la version de base du modèle pré-entraîné CamemBERT (Martin *et al.*, 2020). Il s'agit d'un modèle de type encodeur (BERT) de 110 millions de paramètres et 12 couches BERT. Il a été pré-entraîné sur 138 Go de textes français (Suárez *et al.*, 2019). Bien que des modèles de langue plus récents et plus grands (génératifs) comme Llama2 ou Mistral conduiraient probablement à de meilleurs résultats, le choix d'un modèle de taille raisonnable

est motivé par deux raisons. Premièrement, notre objectif est de montrer que, contrairement à plusieurs autres tâches en TAL, la caractérisation fine des émotions dans les textes ne peut pas être réalisée en sollicitant des modèles de langage génériques (c.-à-d. non spécialisés) de grande taille *via* de l'apprentissage en-contexte (c.-à-d. sans affinage). Ensuite, notre travail vise une solution légère, de sorte que la caractérisation des émotions puisse être intégrée comme un processeur pour l'analyse de la complexité des textes dans une collection massive de textes d'un moteur de recherche public. Ainsi, même si l'affinage de modèles plus grands fait partie de nos perspectives, l'article ne l'aborde pas.

Nous affinons le modèle CamemBERT en remplaçant sa dernière couche de prédiction de jetons par une couche de classification binaire de la taille du nombre d'étiquettes, avec l'entropie croisée binaire comme fonction de perte. L'affinage porte sur tous les poids du modèle, c.-à-d. qu'aucune couche n'est gelée. Suite à des travaux de prototypage sur l'ensemble de développement, le modèle final n'est pas directement appris à partir de CamemBERT. Un premier affinage est conduit sur la seule tâche A pendant 3 époques (couche de classification de taille 1), puis le modèle final est affiné sur toutes les tâches en partant de ce modèle intermédiaire pendant 6 époques supplémentaires (la couche de classification finale est remplacée par une couche vierge de taille 19). L'optimiseur est Adam avec un taux d'apprentissage de 10^{-5} (sans décroissance) et des lots de 8 exemples.

D'autres expérimentations (non rapportés dans cet article) ont été conduites sur l'ensemble de développement, par exemple sur le choix d'une fenêtre ou non autour des phrases, la pondération des classes ou non, ou encore le choix d'un premier affinage uniquement sur la tâche A ou non. Au final, les résultats présentés sont ceux de la meilleure stratégie obtenues sur l'ensemble de développement en moyenne les résultats sur 3 exécution de l'apprentissage avec des initialisation aléatoire différentes. Notamment, une pondération entre classes est adoptée afin que ne favorise pas trop les classes majoritaires. Le facteur de pondération maximal est borné à 50 pour, à l'inverse, ne pas donner non plus trop d'importance aux classes très rares. Enfin, le modèle prend en entrée un triplet de phrases où la phrase cible à étiqueter est entourée de sa phrase précédente et suivante sous la forme `avant : {précédente}</s>actuelle : {cible}</s>après : {suivante}</s>`. Pour plus de détails, le lecteur est invité à consulter (Etienne, 2023).

6 Évaluations automatique et humaine

6.1 Comparaison avec d'autres modèles

Le modèle proposé est comparé à trois autres types de modèles. Des modèles **SVM** ont été entraînés car c'est une approche historique dans le domaine. Deux types de descripteurs d'entrée ont été utilisés : (i) des sac-de-jetons où les jetons proviennent du tokeniseur de CamemBERT, restreints à ceux de l'ensemble d'entraînement, résultant en des vecteurs d'entrée de dimension 18 437 ; (ii) des plongements de phrases de taille 768 obtenus avec SentenceTransformer (Reimers & Gurevych, 2019) et CamemBERT [1]. Des modèles **XGBoost** ont été entraînés car c'est une technique plus récente, légère et compétitive pour de nombreuses tâches de classification, en particulier avec des données déséquilibrées (Chen & Guestrin, 2016). Les descripteurs d'entrée sont les mêmes que pour les SVM. Notre approche est comparée à **GPT-3.5** (Ouyang *et al.*, 2022). Pour un échantillon d'entrée donné, GPT-3.5 est sollicité de manière incrémentale pour l'annoter avec des étiquettes binaires (oui/non). Consécutivement pour chaque tâche et étiquette, une description en langage naturel de ce qui est

1. https://huggingface.co/dangvantuan/sentence-camembert-base

Réf.	Lg.	Étiquettes	Modèle	Lexique	Granularité	Macro-F1 du meill modèle
notre modèle	Fra	colère, dég., joie, peur, surpr., trist.	Transformer	aucun	triplets de phr.	0,52
(Öhman et al., 2020)	Ang	*idem* + conf., anticipation	Transformer	aucun	phr.	0,54
(Kim and Klinger, 2018)	Fra	*idem* + conf., anticipation	symb.	NRC	triplets de phr.	0,31
			MLP	aucun	triplets de phr.	0,31
(Fraisse, Paroubek, 2015)	Fra	colère, peur, tristesse	SVM	perso.	paragr.	0,31

TABLE 4 – Éléments de comparaison avec des travaux proches.

Tâche	Étiquette	Approche	Macro-F1
(A) Présence d'une émotion	émotionelle	notre modèle	**0.752**
		TextBlob	0.299
		Emotaix	0.445
(B) Mode d'expression	comport.	notre modèle	**0.626**
		Emotaix	0.041
	désignée	notre modèle	**0.807**
		Emotaix	0.559
(D) Catégories émot. (mode *désigné* seul.)	toutes	notre modèle	**0.466**
		Emotaix	0.425
Polarité émotionnelle	positive	notre modèle	**0.575**
		TextBlob	0.163
	négative	notre modèle	**0.678**
		TextBlob	0.168

TABLE 5 – Éléments de comparaison avec des outils disponibles pour le français.

attendu est fournie au modèle avant de demander de répondre, accompagnée d'exemples issus de l'ensemble d'entraînement pour chaque étiquette. Différentes amorces ont été testées (*cf.* détails dans l'annexe A). Celle retenue reporte de 2 à 4 exemples positifs par étiquette. Contrairement à SVM, XGBoost et notre modèle, cette approche n'est pas économe mais elle se passe d'entraînement.

Le tableau 3 résume les performances sur l'ensemble de test des meilleurs modèles pour chaque tâche – pour SVM et XGBoost, les descripteurs sac-de-jetons ; pour GPT-3.5, les amorces sans exemples négatifs – et les compare à notre modèle. Les modèles sont évalués à travers les scores de rappel (R), de précision (P) et de score F1. Dans l'ensemble, il apparaît que notre modèle proposé surpasse significativement les SVM, XGBoost et GPT-3.5 en termes de scores F1 pour toutes les tâches, avec des valeurs qui sont presque le double du modèle le mieux classé pour les tâches B, C et D. Il semble surtout que tous les autres modèles tendent à favoriser soit le rappel (GPT-3.5) soit la précision (SVM, XGBoost), tandis que notre modèle est équilibré. Enfin, les faibles résultats de GPT-3.5 montrent que la tâche est difficile et nécessite un affinage.

6.2 Comparaison avec les travaux connexes

À défaut de travaux véritablement similaires aux nôtres, cette section rapporte des résultats complémentaires pour donner une meilleure intuition de la performance de notre modèle.

Travaux comparables les plus proches Le tableau 4 résume les performances des trois travaux les plus proches que nous avons pu trouver dans la littérature. Ils ont été choisis parce qu'ils prédisent tous des étiquettes sur une granularité proche de celui de la phrase. (Öhman *et al.*, 2020) permet une comparaison avec un autre modèle Transformer ; (Fraisse & Paroubek, 2015) avec un autre travail en français ; et (Kim & Klinger, 2018) avec un méthode qui travaille au niveau des marqueurs linguistiques (par opposition au niveau phrastique ou textuel). Tous se concentrent uniquement sur les catégories émotionnelles. Les résultats montrent que notre modèle est compétitif.

Implémentations fondées sur des ressources existantes Deux ressources disponibles en français sont intéressantes pour l'identification des émotions : TextBlob (`https://textblob.readthedocs.io/`), une bibliothèque d'analyse des sentiments qui intègre un lexique français où

les termes sont associés à un poids négatif et positif reflétant leur polarité ; Emotaix (Piolat & Bannour, 2009), un autre lexique comprenant des associations (i) de termes avec des catégories émotionnelles pour le seul mode désigné, et (ii) d'autres termes avec le mode comportemental (mais cette fois sans information sur la catégorie émotionnelle). Plusieurs tâches gérées par notre modèle ont été répliquées *via* TextBlob et Emotaix. Pour tenir compte des différences entre ces ressources et notre modèle proposé, la tâche B a été limitée aux seuls modes comportemental et désigné et la tâche D au mode désigné. De plus, notre modèle a été testé sur une tâche de prédiction de la polarité émotionnelle sur notre ensemble de test puisque TextBlob est conçu pour cet usage. Pour prédire la polarité *via* notre modèle, les catégories ont été prédites et empiriquement projetés vers la polarité positive ou négative (par ex., *colère* est *négative*, *joie* est *positive*). Comme le montre le tableau 5, notre modèle se comporte nettement mieux que TextBlob et Emotaix, y compris dans la tâche de polarité émotionnelle pour laquelle il n'a pas été spécifiquement conçu. La seule tâche pour laquelle la concurrence demeurre est la prédiction des catégories lorsque le mode est désigné, ce qui est la situation la plus facile par rapport à la prise en compte de tous les modes.

6.3 Résultats par étiquette

Le tableau 6 présente les résultats de notre classifieur sur toutes les étiquettes de toutes les tâches. Des observations supplémentaires peuvent être faites comme suit. En ce qui concerne les modes d'expression (B), les émotions désignées sont très bien reconnues (F1 > 0,8), contrairement aux émotions suggérées (F1 < 0,5). Cela n'est pas surprenant car les émotions désignées sont les plus faciles à identifier pour un annotateur humains, tandis que les émotions suggérées ont la part d'interprétation la plus grande. La performance pour les types d'émotions (C) semble pour sa part liée aux résultats sur les catégories émotionnelles, puisque l'étiquette *basique* est, conformément à l'intuition, mieux reconnue que l'étiquette *complexe*. Enfin, concernant les catégories émotionnelles (D), trois d'entre elles ne sont jamais prédites (*culpabilité*, *dégoût* et *jalousie*). Ce sont les étiquettes les plus rares de l'ensemble d'entraînement, probablement trop pour que le modèle apprenne à les prédire. De fait, les catégories émotionnelles les mieux prédites sont les émotions de base, plus fréquentes, à savoir les étiquettes *surprise*, *peur* et *colère* (*cf.* tableau 2). Cependant, alors que *surprise* est l'étiquette la mieux prédite de la tâche D, ce n'est pas la plus représentée dans l'ensemble d'entraînement. Au contraire, *tristesse* n'est pas bien reconnue, même si c'est l'une des catégories émotionnelles les plus fréquentes. De nos analyses complémentaires, cela semblerait s'expliquer par des interactions parfois fortes entre les notions de mode d'expression et de catégorie émotionnelle. Par exemple, la catégorie *surprise*, qui est principalement *montrée* dans le corpus, est en moyenne 14 fois mieux reconnue lorsqu'elle est exprimée par ce mode par rapport aux autres modes. De même, la *colère*, principalement *comportementale* dans le corpus, est 4 fois mieux prédite dans ce mode.

6.4 Évaluation humaine

Étant donné la difficulté des tâches considérées, il est opportun de recouper l'évaluation automatique avec une analyse humaine, notamment pour donner une intuition de ce que représentent les erreurs de prédiction observées. Une expérience de validation perceptive a ainsi été menée avec trois experts en complexité textuelle et en émotions. Chacun d'eux a été informé des tâches et des définitions des étiquette en psycho-linguistique et en linguistique. Ils ont ensuite été chacun confrontés à 150 phrases de l'ensemble de test et à leurs étiquettes de catégorie émotionnelle et de mode d'expression. Ces

Tâche	MacroR	MacroP	MacroF1	Étiquettes	R	P	F1
(A) Présence d'une émotion	0,764	0,741	0,752	émotion	0,764	0,741	0,752
(B) Mode d'expression	0,626	0,665	0,645	comportem.	0,601	0,653	0,626
				désigné	0,811	0,803	0,807
				montré	0,667	0,726	0,695
				suggéré	0,426	0,479	0,451
(C) Type d'émotion	0,557	0,662	0,601	basique	0,705	0,733	0,719
				complexe	0,409	0,591	0,484
(D) Catégorie émotionnelle	0,397	0,463	0,420	admiration	0,281	0,457	0,348
				autre	0,745	0,592	0,660
				colère	0,670	0,685	0,677
				culpabilité	0,000	0,000	0,000
				dégoût	0,000	0,000	0,000
				embarras	0,364	0,600	0,453
				fierté	0,333	0,615	0,432
				jalousie	0,000	0,000	0,000
				joie	0,530	0,709	0,606
				peur	0,717	0,661	0,688
				surprise	0,697	0,739	0,717
				tristesse	0,428	0,504	0,463

TABLE 6 – Performances détaillées de notre modèle

Source de l'étiq.	Opinion de l'évaluateur	Proportion (nb. d'étiquettes) catég. émot.	mode d'expr.
Humain & Modèle	Accord	95,5 % (105)	97,7 % (129)
	Désaccord	4,5 % (5)	2,3 % (3)
Modèle	Accord	92,1 % (58)	91,1 % (41)
	Désaccord	7,9 % (5)	8,9 % (4)
Modèle	Accord	76,5 % (39)	90,2 % (37)
	Désaccord	23,5 % (12)	9,8 % (4)

TABLE 7 – Accord des experts vis-à-vis des prédictions communes à l'annotateur humain et à notre modèle, ou spécifiques à chacun.

étiquettes provenaient soit des annotations humaines de référence, soit des prédictions de notre modèle. Pour chaque étiquette, les experts devaient dire s'ils étaient d'accord ou non avec l'annotation proposée. Bien sûr, ils n'avaient pas connaissance de l'origine des étiquettes.

Le tableau 7 rapporte les taux d'accord des experts avec les étiquettes proposées, en fonction de la source de l'étiquette. Bien que l'accord le plus fort soit lorsque les étiquettes humaines et celles du modèle concordent (*humain & modèle*), les scores d'accord sont globalement très élevés, en particulier pour le mode d'expression. Ces résultats tendent donc à montrer que, même lorsque le modèle prédit différemment de la référence, la prédiction est généralement considérée comme pertinente par les experts humains. Cela démontre que notre modèle est capable de généraliser correctement et que les scores F1 des expériences précédentes sous-estiment la qualité perçue des prédictions du modèle.

7 Conclusion et perspectives

Nous avons proposé un modèle d'analyse des émotions dans des textes qui est original en TAL car il prend en compte leurs modes d'expression directs mais aussi indirects. De plus, les expériences montrent que ce modèle a de bonnes performances par rapport à d'autres approches, des travaux comparables et des solutions à partir de ressources sur étagère. L'évaluation humaine a montré que ce niveau est presque équivalent à ce que les humains peuvent faire.

À l'avenir, des prédictions intra-phrastiques, délimitant des unités, permettant l'inclusion d'autres notions, comme celle d'expérienceur, sont des pistes pour sophistiquer encore l'analyse. L'affinage de modèles génératifs semble alors nécessaire. Sur un autre plan, une application directe de notre modèle est l'analyse de la complexité – contexte plus large de notre travail – , puisque les étiquettes prédites reflètent des marqueurs de complexité. Plus largement, notre travail pourrait contribuer à la recherche en psychologie pour étudier le lien entre le langage émotionnel et l'état psychologique du scripteur/locuteur, dans la lignée des études rapportées dans (Tausczik & Pennebaker, 2010).

Remerciements

Ce travail a en partie été financé par l'Agence Nationale de la Recherche (ANR) dans le cadre du projet TextToKids (ANR-19-CE38-0014).

A Détails de l'inférence avec GPT-3.5

GPT-3.5 a été utilisé en mode conversationnel. Les amorces sont donc une alternances de messages entre l'*utilisateur* et l'*assistant*, précédés d'un message global dit du *système*. Les messages de l'utilisateur parcourt toutes les étiquettes de toutes les tâches A à D en expliquant le sens de chaque étiquette, ceux de l'assistant sont une réponse binaire (« oui » / « non ») pour signaler la présence ou absence de la classe en question. Deux types de messages sont considérés pour l'utilisateur : soit les explications de chaque étiquette sont accompagnées d'exemple positifs, soit elles sont accompagnées d'exemple positifs *et* négatifs (c.-à-d. des contre-exemples). Seule la tâche A (présence ou non d'une information émotionnelle) fait exception puisqu'elle est, en effet, toujours accompagnée de contre-exemples, quelque soit le type d'amorce. La section A.1 donne le détail des résultats sur l'ensemble de test pour chaque approche en comparaison de notre modèle. Les sections A.2 et A.3 montrent ensuite les détails des deux types d'amorces. Nous utilisons la version 0311 de GPT-3.5 pour toutes les expériences.

A.1 Détails des résultats de GPT-3.5

La table 8 donne le détail des résultats sur l'ensemble de test pour chaque approche en comparaison de notre modèle. Dans l'ensemble ces résultats montre que notre modèle s'en sort mieux et que l'approche sans contre-exemples et meilleure que celle avec contre-exemples. Le problème principale de GPT-3.5 semble être qu'il prédit trop d'étiquettes (rappel élevé mais précision faible). Nous pouvons néanmoins noter que GPT-3.5 semble mieux s'en sortir sur les classes rares car notre modèle ne les prédit pas.

A.2 Sans contre-exemples

Système:
Tu joues le rôle d'un expert linguiste qui annote des phrases en t'intéressant à leur dimension émotionnelle.

L'annotation porte au niveau de la phrase et prend la forme de questions successives. Pour comprendre le contexte, la phrase à annoter est donnée avec sa phrase précédente et sa phrase suivante, mais la réponse à chaque question doit uniquement porter sur la seule phrase à annoter, et non sur la phrase précédente ou suivante.

- Phrase précédente: Nicolas Hulot n'appartient à aucun parti politique.
- Phrase à annoter: Il a refusé trois fois le poste de ministre de l'Ecologie avant d'accepter la proposition d'Emmanuel Macron.
- Phrase suivante: Mais ça ne s'est pas très bien passé.

Utilisateur:
Définition: une phrase est dite "émotionnelle" si elle exprime explicitement ou

Tâche	Notre modèle				GPT3.5 sans contre-exemples				GPT3.5 avec contre-exemples				Étiquettes
	macro-F1	R	P	F1	macro-F1	R	P	F1	macro-F1	R	P	F1	
A	**0,752**	**0,764**	**0,741**	**0,752**	0,518	0,622	0,443	0,518	0,382	0,744	0,257	0,382	Émotionnel
B	**0,645**	**0,601**	**0,653**	**0,626**	0,152	0,218	0,154	0,181	0,132	0,416	0,069	0,118	Comportemental
		0,811	**0,803**	**0,807**		0,404	0,095	0,154		0,442	0,126	0,196	Désigné
		0,667	**0,726**	**0,695**		**0,897**	0,066	0,122		0,656	0,054	0,100	Montré
		0,426	**0,479**	**0,451**		**0,534**	0,089	0,153		0,372	0,066	0,112	Suggéré
C	**0,601**	0,705	**0,733**	**0,719**	0,199	**0,709**	0,203	0,315	0,220	0,466	0,286	0,354	Base
		0,409	**0,591**	**0,484**		**0,803**	0,043	0,082		0,260	0,051	0,085	Complexe
D	**0,420**	0,281	0,457	0,348	0,174	**0,825**	0,036	0,069	0,125	0,526	0,028	0,052	Admiration
		0,745	0,592	0,660		**0,798**	0,045	0,085		0,606	0,040	0,074	Autre
		0,670	0,685	0,677		0,665	0,234	0,346		0,590	0,268	0,369	Colère
		0,000	0,000	0,000		**1,000**	**0,222**	0,364		**1,000**	0,003	0,006	Culpabilité
		0,000	0,000	0,000		**0,800**	0,073	0,133		**0,800**	**0,131**	0,225	Dégoût
		0,364	0,600	0,453		0,424	0,110	0,175		**0,758**	0,025	0,048	Embarras
		0,333	**0,615**	**0,432**		0,958	0,023	0,045		**0,771**	0,020	0,039	Fierté
		0,000	0,000	0,000		0,000	0,000	0,000		0,000	0,000	0,000	Jalousie
		0,530	0,709	**0,606**		**0,837**	0,110	0,194		0,683	0,148	0,243	Joie
		0,717	0,661	0,688		0,731	0,153	0,253		**0,741**	0,104	0,182	Peur
		0,697	**0,739**	0,717		**0,873**	0,058	0,109		0,789	0,050	0,094	Surprise
		0,428	**0,504**	0,463		**0,449**	0,247	0,319		0,399	0,110	0,173	Tristesse

TABLE 8 – Comparaison des résultats pour notre modèle et les deux approches testées avec GPT-3.5 pour les tâches A, B, C et D.

implicitement une émotion, qu'elle soit exprimée par le narrateur ou un personnage. Par exemple:
- émotionnelle: "Cette information a beaucoup énervé Marie."
- émotionnelle: "Andrée a sautillé partout en chantant."
- émotionnelle: "Oh, non... C'est vraiment dommage !"
- émotionnelle: "Ces deux amis se retrouvent après une longue séparation."
- non émotionnelle: "Avant d'arriver devant une salle de classe, les enseignants, eux aussi, sont sur les bancs de l'école."
- non émotionnelle: "De 2007 à 2012, il a été le Premier ministre de l'ancien président Nicolas Sarkozy."
- non émotionnelle: "Récemment, une nouvelle autorisation a été délivrée pour un deuxième test dans le courant de l'année 2019."
- non émotionnelle: "Avant de sortir, Billy prépare un dîner orange : une soupe de potiron, des cuisses de canard à l'orange avec une purée de carottes et une tarte à la citrouille."

Question: La phrase à annoter est-elle **émotionnelle** ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: La catégorie émotionnelle "colère" recouvre les émotions suivantes: agacement, colère, contestation, désaccord (si émotion suggérée), désapprobation, énervement, fureur/rage, indignation, insatisfaction, irritation, mécontentement, réprobation et révolte. Par exemple :
- "C'est notamment pour cette raison que des "gilets jaunes", les personnes qui manifestent et bloquent des routes dans le pays depuis plusieurs semaines, sont en colère."
- "- Ton commentaire est déplacé, jeune homme ! a-t-elle dit d'un air pincé."

Question: Si la phrase à annoter est émotionnelle, est-ce que la catégorie

émotionnelle **colère** est présente ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: La catégorie émotionnelle "dégoût" recouvre les émotions suivantes: dégoût, lassitude et répulsion. Par exemple :
- "Beurk !"
- "Ça peut paraître dégoûtant, mais on peut manger des insectes."

Question: Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **dégoût** est présente ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: La catégorie émotionnelle "joie" recouvre les émotions suivantes: amusement, enthousiasme, exaltation, joie et plaisir. Par exemple :
- "Pour fêter ses buts, il lui arrive souvent de danser."
- "– Je suis bien aise de vous voir, me dit le roi sur un ton amical."

Question: Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **joie** est présente ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: La catégorie émotionnelle "peur" recouvre les émotions suivantes: angoisse, appréhension, effroi, horreur, inquiétude, méfiance, peur, stress et timidité. Par exemple :
- "Le Front national, qui est d'extrême droite, faisait peur, à cause des idées qu'il défendait."
- "Il y avait un grand silence dans la maison."

Question: Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **peur** est présente ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: La catégorie émotionnelle "surprise" recouvre les émotions suivantes: étonnement, stupeur, surprise. Par exemple :
- "Finalement, ils ont été pris en charge... par les agriculteurs locaux, dans un camion benne !"
- "Tous, étonnés, se taisent."

Question: Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **surprise** est présente ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: La catégorie émotionnelle "tristesse" recouvre les émotions suivantes: blues, chagrin, déception, désespoir, peine, souffrance et tristesse. Par exemple :
- "Sa mère venait de mourir et son père était au front."
- "L'âne continuait à examiner la peinture d'un regard plutôt attristé."

Question: Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **tristesse** est présente ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: La catégorie émotionnelle "admiration" recouvre les émotions suivantes: admiration. Par exemple :
- "De nos jours, ce site exceptionnel permet de montrer toute la richesse de la civilisation romaine et la façon dont les villes et la société étaient organisées."
- "– Tes enfants sont vraiment merveilleux, ma chérie, dit-elle à sa fille."

Question: Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **admiration** est présente ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: La catégorie émotionnelle "culpabilité" recouvre les émotions suivantes: culpabilité. Par exemple :
- "Et je l'avais bien mérité."
- "Surtout, il ne faut pas se sentir coupable de ne pas avoir réagi."

Question: Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **culpabilité** est présente ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: La catégorie émotionnelle "embarras" recouvre les émotions suivantes: embarras, gêne, honte, humiliation et timidité. Par exemple :
- "Après cette humiliante défaite, Napoléon abdique une nouvelle fois, ce qui marque définitivement la fin de l'Empire et de sa période de retour appelée "les Cent jours"."
- "Légèrement décontenancée, la prof s'est raclé la gorge et commencé la lecture."

Question: Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **embarras** est présente ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: La catégorie émotionnelle "fierté" recouvre les émotions suivantes: fierté et orgueil. Par exemple :
- "Flavia entre dans la cour comme une conquérante, entourée de ses supporters."

- "Magawa peut être fier de lui, car il vient de recevoir une médaille d'or."

Question: Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **fierté** est présente ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: La catégorie émotionnelle "jalousie" recouvre les émotions suivantes: jalousie. Par exemple :
- "Mais quand Flavia découvre le jeune génie du piano, elle se sent comme écrasée."
- "On dirait presque qu'il fait partie de l'instrument."

Question: Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **jalousie** est présente ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: La catégorie émotionnelle "autre" recouvre les émotions suivantes: amour, courage, curiosité, désir, détermination, envie, espoir, haine, impuissance, mépris et soulagement. Par exemple :
- "Dans chaque camp, ils se sont mobilisés pour donner envie aux gens de voter comme eux."
- "Ils n'apprécient pas du tout l'attitude des dirigeants, notamment celle du président, "qu'ils jugent méprisant, déconnecté de la réalité, du quotidien", note le sociologue Alexis Spire."

Question: Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **autre** est présente ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: Les émotions suivantes sont dites "de base" : Colère, Dégoût, Joie, Peur, Surprise, Tristesse.

Question: Si la phrase à annoter est émotionnelle, contient-elle une **émotion de base** ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: Les émotions suivantes sont dites "complexes": Admiration, Culpabilité, Embarras, Fierté, Jalousie.

Question: Si la phrase à annoter est émotionnelle, contient-elle une **émotion complexe** ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:

Définition: Une émotion est dite du mode "désigné" lorsqu'elle est exprimée par un terme du lexique émotionnel. Par exemple :
- "Pierre est heureux d'être bientôt à la retraite.", où la joie de Pierre est désignée par le terme "heureux".
- "Cette information a beaucoup énervé Marie.", où la colère de Marie est désignée par le terme "énervé".

Question: Si la phrase à annoter est émotionnelle, est-ce que le mode **désigné** est utilisé ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: Une émotion est dite du mode "comportemental" lorsqu'elle est exprimée par la description d'une manifestation physique (physiologique ou comportementale) de l'émotion. Par exemple :
- "Paul sanglote.", où la tristesse de Paul est exprimée par le comportement "sanglote".
- "Andrée a sautillé partout en chantant.", où la joie de Andrée est exprimée par le comportement "sautillé partout en chantant".

Question: Si la phrase à annoter est émotionnelle, est-ce que le mode **comportemental** est utilisé ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: Une émotion est dite du mode "montré" lorsqu'elle est exprimée par des caractéristiques linguistiques de l'énoncé qui traduisent l'état émotionnel dans lequel se trouvait l'énonciateur au moment de l'énonciation. Par exemple :
- "Oh, chouette ! Quelle bonne idée !", car la joie de l'énonciateur est traduite au sein de l'énoncé par les interjections "oh" et "chouette", les énoncés averbaux et les points d'exclamations.
- "Oh, non... C'est vraiment dommage !", car la tristesse de l'énonciateur est traduite au sein de l'énoncé par l'interjection "oh", l'énoncé averbal, les points de suspension et le point d'exclamation.

Question: Si la phrase à annoter est émotionnelle, est-ce que le mode **montré** est utilisé ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

Utilisateur:
Définition: Une émotion est dite du mode "suggéré" lorsqu'elle est exprimée par la description d'une situation associée de manière conventionnelle à un ressenti émotionnel. Par exemple :
- "Le père de Jeanne est mort hier à cause d'un cancer.", où la tristesse de Jeanne est suggérée par la description du décès, il y a peu de temps, de son père (une personne proche d'elle).
- "Ces deux amis se retrouvent après une longue séparation.", où la joie des deux amis est suggérée par la description de leurs retrouvailles après un temps long.

Question: Si la phrase à annoter est émotionnelle, est-ce que le mode **suggéré** est utilisé ?

Réponse (oui/non):

Assistant:
<réponse du modèle>

A.3 Avec contre-exemples

Système:
Tu joues le rôle d'un expert linguiste qui annote des phrases d'après leurs dimensions émotionnelle.

Les différentes annotations sont toute binaires (absence ou présence d'une propriété). Elles vont porter sur la nature émotionnelle ou non des phrases et, si oui, le mode d'expression de la ou des émotions présentes (désignée, comportementale, montrée ou suggérée), la ou les catégories émotionnelles (joie, peur, colère, tristesse, etc.) et le ou les types d'émotion ("de base" ou "complexe"). Chaque propriété est décrite par une définition et des exemples.

L'annotation La phrase à annoter est entourée des balises <annotate>...</annotate>.

Utilisateur:
Définition : une phrase est dite "émotionnelle" si elle exprime explicitement ou implicitement une émotion, qu'elle soit exprimée par le narrateur ou un personnage.

Question : La phrase à annoter est-elle **émotionnelle** ?

Exemples :

- <annotate>Avant de sortir, Billy prépare un dîner orange : une soupe de potiron, des cuisses de canard à l'orange avec une purée de carottes et une tarte à la citrouille.</annotate> -> non
- <annotate>Cette information a beaucoup énervé Marie.</annotate> -> oui
- <annotate>Andrée a sautillé partout en chantant.</annotate> -> oui
- <annotate>Récemment, une nouvelle autorisation a été délivrée pour un deuxième test dans le courant de l'année 2019.</annotate> -> non - <annotate>Oh, non... C'est vraiment dommage !</annotate> -> oui
- <annotate>De 2007 à 2012, il a été le Premier ministre de l'ancien président Nicolas Sarkozy.</annotate> -> non
- <annotate>Ces deux amis se retrouvent après une longue séparation. -> oui
- <annotate>Avant d'arriver devant une salle de classe, les enseignants, eux aussi, sont sur les bancs de l'école.</annotate> -> non

Annotation (oui/non) :
- Nicolas Hulot n'appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l'Ecologie avant d'accepter la proposition d'Emmanuel Macron.</annotate> Mais ça ne s'est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : La catégorie émotionnelle "colère" recouvre les émotions suivantes: agacement, colère, contestation, désaccord (si émotion suggérée), désapprobation, énervement, fureur/rage, indignation, insatisfaction, irritation, mécontentement, réprobation et révolte.

Question : Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **colère** est présente ?

Exemples :
- <annotate>De 2007 à 2012, il a été le Premier ministre de l'ancien président

Nicolas Sarkozy.</annotate> -> non
- <annotate>C’est notamment pour cette raison que des "gilets jaunes", les personnes qui manifestent et bloquent des routes dans le pays depuis plusieurs semaines, sont en colère.</annotate> -> oui.
- <annotate>Tous, étonnés, se taisent.</annotate> -> non.
- <annotate>- Ton commentaire est déplacé, jeune homme ! a-t-elle dit d’un air pincé.</annotate> -> oui.
- <annotate>Après cette humiliante défaite, Napoléon abdique une nouvelle fois, ce qui marque définitivement la fin de l’Empire et de sa période de retour appelée "les Cent jours".</annotate> -> non.

Annotation (oui/non) :
- Nicolas Hulot n’appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l’Ecologie avant d’accepter la proposition d’Emmanuel Macron.</annotate> Mais ça ne s’est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : La catégorie émotionnelle "dégoût" recouvre les émotions suivantes: dégoût, lassitude et répulsion.

Question : Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **dégoût** est présente ?

Exemples :
- <annotate>Ça peut paraître dégoûtant, mais on peut manger des insectes.</annotate> -> oui.
- <annotate>Beurk !</annotate> -> oui.
- <annotate>Finalement, ils ont été pris en charge... par les agriculteurs locaux, dans un camion benne !</annotate> -> non.
- <annotate>Le Front national, qui est d’extrême droite, faisait peur, à cause des idées qu’il défendait.</annotate> -> non.
- <annotate>Avant d’arriver devant une salle de classe, les enseignants, eux aussi, sont sur les bancs de l’école.</annotate> -> non

Annotation (oui/non) :
- Nicolas Hulot n’appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l’Ecologie avant d’accepter la proposition d’Emmanuel Macron.</annotate> Mais ça ne s’est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : La catégorie émotionnelle "joie" recouvre les émotions suivantes: amusement, enthousiasme, exaltation, joie et plaisir.

Question : Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **joie** est présente ?

Exemples :
- <annotate>Dans chaque camp, ils se sont mobilisés pour donner envie aux gens de voter comme eux.</annotate> -> non.
- <annotate>- Je suis bien aise de vous voir, me dit le roi sur un ton amical.</annotate> -> oui.
- <annotate>Beurk !</annotate> -> non.
- <annotate>Avant d’arriver devant une salle de classe, les enseignants, eux aussi, sont sur les bancs de l’école.</annotate> -> non
- <annotate>Pour fêter ses buts, il lui arrive souvent de danser.</annotate> -> oui.

Annotation (oui/non) :
- Nicolas Hulot n’appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l’Ecologie avant d’accepter la proposition d’Emmanuel Macron.</annotate> Mais ça ne s’est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : La catégorie émotionnelle "peur" recouvre les émotions suivantes: angoisse, appréhension, effroi, horreur, inquiétude, méfiance, peur, stress et timidité.

Question : Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **peur** est présente ?

Exemples :
- <annotate>Le Front national, qui est d’extrême droite, faisait peur, à cause des idées qu’il défendait.</annotate> -> oui.
- <annotate>Dans chaque camp, ils se sont mobilisés pour donner envie aux gens de voter comme eux.</annotate> -> non.
- <annotate>Ça peut paraître dégoûtant, mais on peut manger des insectes.</annotate> -> non.
- <annotate>Récemment, une nouvelle autorisation a été délivrée pour un deuxième test dans le courant de l’année 2019.</annotate> -> non
- <annotate>Il y avait un grand silence dans la maison.</annotate> -> oui.

Annotation (oui/non) :
- Nicolas Hulot n’appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l’Ecologie avant d’accepter la proposition d’Emmanuel Macron.</annotate> Mais ça ne s’est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : La catégorie émotionnelle "surprise" recouvre les émotions suivantes: étonnement, stupeur, surprise.

Question : Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **surprise** est présente ?

Exemples :
- <annotate>Finalement, ils ont été pris en charge... par les agriculteurs locaux, dans un camion benne !</annotate> -> oui.
- <annotate>Avant d’arriver devant une salle de classe, les enseignants, eux aussi, sont sur les bancs de l’école.</annotate> -> non
- <annotate>Mais quand Flavia découvre le jeune génie du piano, elle se sent comme écrasée.</annotate> -> non.
- <annotate>Beurk !</annotate> -> non.
- <annotate>Tous, étonnés, se taisent.</annotate> -> oui.

Annotation (oui/non) :
- Nicolas Hulot n’appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l’Ecologie avant d’accepter la proposition d’Emmanuel Macron.</annotate> Mais ça ne s’est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : La catégorie émotionnelle "tristesse" recouvre les émotions suivantes: blues, chagrin, déception, désespoir, peine, souffrance et tristesse.

Question : Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **tristesse** est présente ?

Exemples :
- <annotate>Avant d'arriver devant une salle de classe, les enseignants, eux aussi, sont sur les bancs de l'école.</annotate> -> non
- <annotate>Le Front national, qui est d'extrême droite, faisait peur, à cause des idées qu'il défendait.</annotate> -> non.
- <annotate>Sa mère venait de mourir et son père était au front.</annotate> -> oui.
- <annotate>Légèrement décontenancée, la prof s'est raclé la gorge et commencé la lecture.</annotate> -> non.
- <annotate>L'âne continuait à examiner la peinture d'un regard plutôt attristé.</annotate> -> oui.

Annotation (oui/non) :
- Nicolas Hulot n'appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l'Ecologie avant d'accepter la proposition d'Emmanuel Macron.</annotate> Mais ça ne s'est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : La catégorie émotionnelle "admiration" recouvre les émotions suivantes: admiration.

Question : Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **admiration** est présente ?

Exemples :
- <annotate>Tous, étonnés, se taisent.</annotate> -> non.
- <annotate>De nos jours, ce site exceptionnel permet de montrer toute la richesse de la civilisation romaine et la façon dont les villes et la société étaient organisées.</annotate> -> oui.
- <annotate>Magawa peut être fier de lui, car il vient de recevoir une médaille d'or.</annotate> -> non.
- <annotate>Avant de sortir, Billy prépare un dîner orange : une soupe de potiron, des cuisses de canard à l'orange avec une purée de carottes et une tarte à la citrouille.</annotate> -> non
- <annotate>- Tes enfants sont vraiment merveilleux, ma chérie, dit-elle à sa fille.</annotate> -> oui.

Annotation (oui/non) :
- Nicolas Hulot n'appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l'Ecologie avant d'accepter la proposition d'Emmanuel Macron.</annotate> Mais ça ne s'est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : La catégorie émotionnelle "culpabilité" recouvre les émotions suivantes: culpabilité.

Question : Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **culpabilité** est présente ?

Exemples :
- <annotate>Et je l'avais bien mérité.</annotate> -> oui.
- <annotate>Tous, étonnés, se taisent.</annotate> -> non.
- <annotate>Surtout, il ne faut pas se sentir coupable de ne pas avoir réagi.</annotate> -> oui.
- <annotate>Tous, étonnés, se taisent.</annotate> -> non.

- <annotate>Avant d'arriver devant une salle de classe, les enseignants, eux aussi, sont sur les bancs de l'école.</annotate> -> non

Annotation (oui/non) :
- Nicolas Hulot n'appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l'Ecologie avant d'accepter la proposition d'Emmanuel Macron.</annotate> Mais ça ne s'est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : La catégorie émotionnelle "embarras" recouvre les émotions suivantes: embarras, gêne, honte, humiliation et timidité.

Question : Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **embarras** est présente ?

Exemples :
- <annotate>Le Front national, qui est d'extrême droite, faisait peur, à cause des idées qu'il défendait.</annotate> -> non.
- <annotate>- Tes enfants sont vraiment merveilleux, ma chérie, dit-elle à sa fille.</annotate> -> non.
- <annotate>Avant d'arriver devant une salle de classe, les enseignants, eux aussi, sont sur les bancs de l'école.</annotate> -> non
- <annotate>Après cette humiliante défaite, Napoléon abdique une nouvelle fois, ce qui marque définitivement la fin de l'Empire et de sa période de retour appelée "les Cent jours".</annotate> -> oui.
- <annotate>Légèrement décontenancée, la prof s'est raclé la gorge et commencé la lecture.</annotate> -> oui.

Annotation (oui/non) :
- Nicolas Hulot n'appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l'Ecologie avant d'accepter la proposition d'Emmanuel Macron.</annotate> Mais ça ne s'est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : La catégorie émotionnelle "fierté" recouvre les émotions suivantes: fierté et orgueil.

Question : Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **fierté** est présente ?

Exemples :
- <annotate>Avant de sortir, Billy prépare un dîner orange : une soupe de potiron, des cuisses de canard à l'orange avec une purée de carottes et une tarte à la citrouille.</annotate> -> non
- <annotate>On dirait presque qu'il fait partie de l'instrument.</annotate> -> non.
- <annotate>Magawa peut être fier de lui, car il vient de recevoir une médaille d'or.</annotate> -> oui.
- <annotate>Flavia entre dans la cour comme une conquérante, entourée de ses supporters.</annotate> -> oui.
- <annotate>Il y avait un grand silence dans la maison.</annotate> -> non.

Annotation (oui/non) :
- Nicolas Hulot n'appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l'Ecologie avant d'accepter la proposition d'Emmanuel Macron.</annotate> Mais ça ne s'est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : La catégorie émotionnelle "jalousie" recouvre les émotions suivantes: jalousie.

Question : Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **jalousie** est présente ?

Exemples :
- <annotate>On dirait presque qu'il fait partie de l'instrument.</annotate> -> oui.
- <annotate>Et je l'avais bien mérité.</annotate> -> non.
- <annotate>Et je l'avais bien mérité.</annotate> -> non.
- <annotate>Mais quand Flavia découvre le jeune génie du piano, elle se sent comme écrasée.</annotate> -> oui.
- <annotate>Avant d'arriver devant une salle de classe, les enseignants, eux aussi, sont sur les bancs de l'école.</annotate> -> non

Annotation (oui/non) :
- Nicolas Hulot n'appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l'Ecologie avant d'accepter la proposition d'Emmanuel Macron.</annotate> Mais ça ne s'est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : La catégorie émotionnelle "autre" recouvre les émotions suivantes: amour, courage, curiosité, désir, détermination, envie, espoir, haine, impuissance, mépris et soulagement.

Question : Si la phrase à annoter est émotionnelle, est-ce que la catégorie émotionnelle **autre** est présente ?

Exemples :
- <annotate>De nos jours, ce site exceptionnel permet de montrer toute la richesse de la civilisation romaine et la façon dont les villes et la société étaient organisées.</annotate> -> non.
- <annotate>L'âne continuait à examiner la peinture d'un regard plutôt attristé.</annotate> -> non.
- <annotate>Récemment, une nouvelle autorisation a été délivrée pour un deuxième test dans le courant de l'année 2019.</annotate> -> non
- <annotate>Ils n'apprécient pas du tout l'attitude des dirigeants, notamment celle du président, "qu'ils jugent méprisant, déconnecté de la réalité, du quotidien", note le sociologue Alexis Spire.</annotate> -> oui.
- <annotate>Dans chaque camp, ils se sont mobilisés pour donner envie aux gens de voter comme eux.</annotate> -> oui.

Annotation (oui/non) :
- Nicolas Hulot n'appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l'Ecologie avant d'accepter la proposition d'Emmanuel Macron.</annotate> Mais ça ne s'est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : Les émotions suivantes sont dites "de base" : Colère, Dégoût, Joie, Peur, Surprise, Tristesse.

Question : Si la phrase à annoter est émotionnelle, contient-elle une **émotion de base** ?

Annotation (oui/non) :
- Nicolas Hulot n'appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l'Ecologie avant d'accepter la proposition d'Emmanuel

Macron.</annotate> Mais ça ne s'est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : Les émotions suivantes sont dites "complexes": Admiration, Culpabilité, Embarras, Fierté, Jalousie.

Question : Si la phrase à annoter est émotionnelle, contient-elle une **émotion complexe** ?

Annotation (oui/non) :
- Nicolas Hulot n'appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l'Ecologie avant d'accepter la proposition d'Emmanuel Macron.</annotate> Mais ça ne s'est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : Une émotion est dite du mode "désigné" lorsqu'elle est exprimée par un terme du lexique émotionnel.

Question : Si la phrase à annoter est émotionnelle, est-ce que le mode **désigné** est utilisé ?

Exemples :
- <annotate>Pierre est heureux d'être bientôt à la retraite.</annotate> -> oui (car la joie de Pierre est désignée par le terme "heureux").
- <annotate>Oh, non... C'est vraiment dommage !</annotate> -> non.
- <annotate>Avant d'arriver devant une salle de classe, les enseignants, eux aussi, sont sur les bancs de l'école.</annotate> -> non
- <annotate>Oh, non... C'est vraiment dommage !</annotate> -> non.
- <annotate>Cette information a beaucoup énervé Marie.</annotate> -> oui (car la colère de Marie est désignée par le terme "énervé").

Annotation (oui/non) :
- Nicolas Hulot n'appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l'Ecologie avant d'accepter la proposition d'Emmanuel Macron.</annotate> Mais ça ne s'est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : Une émotion est dite du mode "comportemental" lorsqu'elle est exprimée par la description d'une manifestation physique (physiologique ou comportementale) de l'émotion.

Question : Si la phrase à annoter est émotionnelle, est-ce que le mode **comportemental** est utilisé ?

Exemples :
- <annotate>Cette information a beaucoup énervé Marie.</annotate> -> non.
- <annotate>Paul sanglote.</annotate> -> oui (car la tristesse de Paul est exprimée par le comportement "sanglote").
- <annotate>Avant d'arriver devant une salle de classe, les enseignants, eux aussi, sont sur les bancs de l'école.</annotate> -> non
- <annotate>Le père de Jeanne est mort hier à cause d'un cancer.</annotate> -> non.
- <annotate>Andrée a sautillé partout en chantant.</annotate> -> oui (car la joie de Andrée est exprimée par le comportement "sautillé partout en chantant").

Annotation (oui/non) :
- Nicolas Hulot n'appartient à aucun parti politique. <annotate>Il a refusé trois

fois le poste de ministre de l'Ecologie avant d'accepter la proposition d'Emmanuel Macron.</annotate> Mais ça ne s'est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : Une émotion est dite du mode "montré" lorsqu'elle est exprimée par des caractéristiques linguistiques de l'énoncé qui traduisent l'état émotionnel dans lequel se trouvait l'énonciateur au moment de l'énonciation.

Question : Si la phrase à annoter est émotionnelle, est-ce que le mode **montré** est utilisé ?

Exemples :
- <annotate>Andrée a sautillé partout en chantant.</annotate> -> non.
- <annotate>Paul sanglote.</annotate> -> non.
- <annotate>Oh, chouette ! Quelle bonne idée !</annotate> -> oui (car la joie de l'énonciateur est traduite au sein de l'énoncé par les interjections "oh" et "chouette", les énoncés averbaux et les points d'exclamations).
- <annotate>Oh, non... C'est vraiment dommage !</annotate> -> oui (car la tristesse de l'énonciateur est traduite au sein de l'énoncé par l'interjection "oh", l'énoncé averbal, les points de suspension et le point d'exclamation.)
- <annotate>Avant d'arriver devant une salle de classe, les enseignants, eux aussi, sont sur les bancs de l'école.</annotate> -> non

Annotation (oui/non) :
- Nicolas Hulot n'appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l'Ecologie avant d'accepter la proposition d'Emmanuel Macron.</annotate> Mais ça ne s'est pas très bien passé. ->

Assistant:
réponse du modèle

Utilisateur:
Définition : Une émotion est dite "suggérée" lorsqu'elle est exprimée par la description d'une situation associée de manière conventionnelle à un ressenti émotionnel.

Question : Si la phrase à annoter est émotionnelle, est-ce que le mode **suggéré** est utilisé ?

Exemples :
- <annotate>Oh, chouette ! Quelle bonne idée !</annotate> -> non.
- <annotate>Le père de Jeanne est mort hier à cause d'un cancer.</annotate> -> oui (car où la tristesse de Jeanne est suggérée par la description du décès, il y a peu de temps, de son père, une personne proche d'elle).
- <annotate>Ces deux amis se retrouvent après une longue séparation.</annotate> -> oui (car la joie des deux amis est suggérée par la description de leurs retrouvailles après un temps long).
- <annotate>De 2007 à 2012, il a été le Premier ministre de l'ancien président Nicolas Sarkozy.</annotate> -> non
- <annotate>Andrée a sautillé partout en chantant.</annotate> -> non.

Annotation (oui/non) :
- Nicolas Hulot n'appartient à aucun parti politique. <annotate>Il a refusé trois fois le poste de ministre de l'Ecologie avant d'accepter la proposition d'Emmanuel Macron.</annotate> Mais ça ne s'est pas très bien passé. ->

Assistant:
réponse du modèle

Références

ABDAOUI A., AZÉ J., BRINGAY S. & PONCELET P. (2017). Feel : a french expanded emotion lexicon. *Language Resources and Evaluation*, **51**(3), 833–855. Publisher : Springer.

ACHEAMPONG F. A., WENYU C. & NUNOO-MENSAH H. (2020). Text-based emotion detection : Advances, challenges, and opportunities. *Engineering Reports*, **2**(7), e12189. Publisher : Wiley Online Library.

ALM C. O., ROTH D. & SPROAT R. (2005). Emotions from Text : Machine Learning for Text-based Emotion Prediction. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, p. 579–586, Vancouver, British Columbia, Canada : Association for Computational Linguistics.

AMAN S. & SZPAKOWICZ S. (2007). Identifying expressions of emotion in text. In *Proceedings of the International Conference on Text, Speech and Dialogue (TSD)*, p. 196–205 : Springer.

BALAHUR A., HERMIDA J. M. & MONTOYO A. (2012). Detecting implicit expressions of emotion in text : A comparative analysis. *Decision support systems*, **53**(4), 742–753. Publisher : Elsevier.

BARON-COHEN S., GOLAN O., WHEELWRIGHT S. & GRANADER Y. (2010). Emotion Word Comprehension from 4 to 16 Years Old : A Developmental Survey. *Frontiers in Evolutionary Neuroscience*, **0**. Publisher : Frontiers, DOI : 10.3389/fnevo.2010.00109.

BIANCHI F., NOZZA D. & HOVY D. (2021). Feel-it : Emotion and sentiment classification for the italian language. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, p. 76–83.

BLANC N. & QUENETTE G. (2017). La production d'inférences émotionnelles entre 8 et 10 ans : quelle méthodologie pour quels résultats ? *Enfance*, **4**(4), 503–511. Place : Paris Publisher : NecPlus, DOI : 10.3917/enf1.174.0503.

BOSTAN L.-A.-M. & KLINGER R. (2018). An Analysis of Annotated Corpora for Emotion Classification in Text. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, p. 2104–2119, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

BUSSO C., BULUT M., LEE C.-C., KAZEMZADEH A., MOWER E., KIM S., CHANG J. N., LEE S. & NARAYANAN S. S. (2008). Iemocap : Interactive emotional dyadic motion capture database. *Language resources and evaluation*, **42**, 335–359.

CASEL F., HEINDL A. & KLINGER R. (2021). Emotion recognition under consideration of the emotion component process model. In *Proceedings of the Conference on Natural Language Processing*, p. 49–61, Düsseldorf, Germany : KONVENS 2021 Organizers.

CHEN S.-Y., HSU C.-C., KUO C.-C., KU L.-W. *et al.* (2018). Emotionlines : An emotion corpus of multi-party conversations. *arXiv preprint arXiv :1802.08379*.

CHEN T. & GUESTRIN C. (2016). Xgboost : A scalable tree boosting system. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining*, p. 785–794.

CREISSEN S. & BLANC N. (2017). Quelle représentation des différentes facettes de la dimension émotionnelle d'une histoire entre l'âge de 6 et 10ans ? apports d'une étude multimédia. *Psychologie Française*, **62**(3), 263–277. Cognition et multimédia : les atouts du numérique en situation d'apprentissage, DOI : https ://doi.org/10.1016/j.psfr.2015.07.006.

DAVIDSON D. (2006). The Role of Basic, Self-Conscious and Self-Conscious Evaluative Emotions in Children's Memory and Understanding of Emotion. *Motivation and Emotion*, **30**(3), 232–242. DOI : 10.1007/s11031-006-9037-6.

DAVIDSON D., LUO Z. & BURDEN M. J. (2001). Children's recall of emotional behaviours, emotional labels, and nonemotional behaviours : Does emotion enhance memory ? *Cognition and Emotion*, **15**(1), 1–26. Place : United Kingdom Publisher : Taylor & Francis, DOI : 10.1080/0269993004200105.

DEMSZKY D., MOVSHOVITZ-ATTIAS D., KO J., COWEN A. S., NEMADE G. & RAVI S. (2020). GoEmotions : A Dataset of Fine-Grained Emotions. *CoRR*, **abs/2005.00547**. arXiv : 2005.00547.

DIJKSTRA K., ZWAAN R. A., GRAESSER A. C. & MAGLIANO J. P. (1995). Character and reader emotions in literary texts. *Poetics*, **23**(1-2), 139–157. Publisher : Elsevier.

DYER M. G. (1983). The role of affect in narratives. *Cognitive science*, **7**(3), 211–242. Publisher : Wiley Online Library.

EHRET K., BERDICEVSKIS A., BENTZ C. & BLUMENTHAL-DRAMÉ A. (2023). Measuring language complexity : challenges and opportunities. *Linguistics Vanguard*, **9**(s1), 1–8.

EKMAN P. (1992). An argument for basic emotions. *Cognition and Emotion*, **6**(3-4), 169–200. Publisher : Routledge _eprint : https ://doi.org/10.1080/02699939208411068, DOI : 10.1080/02699939208411068.

ETIENNE A. (2023). *Analyse automatique des émotions dans les textes : contributions théoriques et applicatives dans le cadre de l'étude de la complexité des textes pour enfants*. Thèse de doctorat, Université de Nanterre-Paris X.

ETIENNE A., BATTISTELLI D. & LECORVÉ G. (2022). A (Psycho-)Linguistically Motivated Scheme for Annotating and Exploring Emotions in a Genre-Diverse Corpus. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Marseille, France.

FRAISSE A. & PAROUBEK P. (2015). Utiliser les interjections pour détecter les émotions. In *Actes de la conférence sur le Traitement Automatique des Langues Naturelles*, p. 279–290.

KIM E. & KLINGER R. (2018). Who feels what and why ? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, p. 1345–1359.

KIM E. & KLINGER R. (2019). An Analysis of Emotion Communication Channels in Fan-Fiction : Towards Emotional Storytelling. In *Proceedings of the Workshop on Storytelling*, p. 56–64, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-3406.

KLINGER R. (2023). Where are we in event-centric emotion analysis ? bridging emotion role labeling and appraisal-based approaches.

LIU C., OSAMA M. & ANDRADE A. D. (2019). DENS : A Dataset for Multi-class Emotion Analysis. *CoRR*, **abs/1910.11769**. arXiv : 1910.11769.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE , SEDDAH D. & SAGOT B. (2020). CamemBERT : a Tasty French Language Model. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 7203–7219, Online : Association for Computational Linguistics.

MOHAMMAD S. (2012). #Emotional Tweets. In *Proceedings of the Joint Conference on Lexical and International Workshop on Semantic Evaluation (SemEval)*, p. 246–255, Montréal, Canada : Association for Computational Linguistics.

MOHAMMAD S., BRAVO-MARQUEZ F., SALAMEH M. & KIRITCHENKO S. (2018). SemEval-2018 Task 1 : Affect in Tweets. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, p. 1–17, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/S18-1001.

OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C., MISHKIN P., ZHANG C., AGARWAL S., SLAMA K., RAY A. *et al.* (2022). Training language models to follow instructions

with human feedback. *Proceedings of the Advances in Neural Information Processing Systems*, **35**, 27730–27744.

PIOLAT A. & BANNOUR R. (2009). An example of text analysis software (emotaix-tropes) use : The influence of anxiety on expressive writing. *Current psychology letters. Behaviour, brain & cognition*, **25**(2, 2009).

PLUTCHIK R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion*, p. 3–33. Elsevier.

PORIA S., HAZARIKA D., MAJUMDER N., NAIK G., CAMBRIA E. & MIHALCEA R. (2018). Meld : A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv :1810.02508*.

PORIA S., MAJUMDER N., MIHALCEA R. & HOVY E. (2019). Emotion recognition in conversation : Research challenges, datasets, and recent advances. *IEEE Access*, **7**, 100943–100953.

REIMERS N. & GUREVYCH I. (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3982–3992.

SCHERER K. R. (2005). What are emotions ? And how can they be measured ? *Social science information*, **44**(4), 695–729. Publisher : Sage Publications Sage CA : Thousand Oaks, CA.

STRAPPARAVA C. & MIHALCEA R. (2007). SemEval-2007 Task 14 : Affective Text. In *Proceedings of the International Workshop on Semantic Evaluations (SemEval)*, p. 70–74, Prague, Czech Republic : Association for Computational Linguistics.

SUÁREZ P. J. O., SAGOT B. & ROMARY L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on the Challenges in the Management of Large Corpora* : Leibniz-Institut für Deutsche Sprache.

TAUSCZIK Y. R. & PENNEBAKER J. W. (2010). The psychological meaning of words : LIWC and computerized text analysis methods. *Journal of language and social psychology*, **29**(1), 24–54. Publisher : Sage Publications Sage CA : Los Angeles, CA.

TROIANO E., OBERLÄNDER L. & KLINGER R. (2023). Dimensional modeling of emotions in text with appraisal theories : Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, **49(1)**.

ÖHMAN E. (2020). Emotion annotation : Rethinking emotion categorization. *Proceedings of the CEUR Workshop*, **2865**, 134–144. Publisher : CEUR-WS.

ÖHMAN E., PÀMIES M., KAJAVA K. & TIEDEMANN J. (2020). XED : A Multilingual Dataset for Sentiment Analysis and Emotion Detection. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, p. 6542–6552, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.575.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

TCFLE-8 : un corpus de productions écrites d'apprenants de français langue étrangère et son application à la correction automatisée de textes

Rodrigo Wilkens[1] Alice Pintard[1] David Alfter[2] Vincent Folny[3]
Thomas François[1]
(1) CENTAL, IL&C, Université catholique de Louvain, Belgique prenom.nom@uclouvain.be,
(2) University of Gothenburg david.alfter@gu.se
(3) France Éducation International Folny@france-education-international.fr
{rodrigo.wilkens, alice.pintard, thomas.francois}@uclouvain.be,
Folny@france-education-international.fr, david.alfter@gu.se

RÉSUMÉ

La correction automatisée de textes (CAT) vise à évaluer automatiquement la qualité de textes écrits. L'automatisation permet une évaluation à grande échelle ainsi qu'une amélioration de la cohérence, de la fiabilité et de la normalisation du processus. Ces caractéristiques sont particulièrement importantes dans le contexte des examens de certification linguistique. Cependant, un goulot d'étranglement majeur dans le développement des systèmes CAT est la disponibilité des corpus. Dans cet article, nous visons à encourager le développement de systèmes de correction automatique en fournissant le corpus TCFLE-8 [1], un corpus de 6 569 essais collectés dans le contexte de l'examen de certification *Test de Connaissance du Français* (TCF). Nous décrivons la procédure d'évaluation stricte qui a conduit à la notation de chaque essai par au moins deux évaluateurs selon l'échelle du Cadre européen commun de référence pour les langues (CECR) et à la création d'un corpus équilibré. Nous faisons également progresser les performances de l'état de l'art pour la tâche de CAT en français en expérimentant deux solides modèles de référence.

ABSTRACT

TCFLE-8 : a Corpus of Learner Written Productions for French as a Foreign Language and its Application to Automated Essay Scoring [2]

Automated Essay Scoring (AES) aims to automatically assess the quality of essays. Automation enables large-scale assessment, improvements in consistency, reliability, and standardization. Those characteristics are of particular relevance in the context of language certification exams. However, a major bottleneck in the development of AES systems is the availability of corpora. In this paper, we aim to foster the development of AES by providing the TCFLE-8 corpus, a corpus of 6.5k essays collected in the context of the French Knowledge Test (TCF) certification exam. We report the strict quality procedure that led to the scoring of each essay by at least two raters according to the levels of the Common European Framework of Reference for Languages (CEFR) and to the creation of a

1. TCFLE-8 est disponible à l'adresse https://www.france-education-international.fr/corpus
2. Cet article est une adaptation d'une publication en anglais : Wilkens, R., Pintard, A., Alfter, D., Folny, V., & François, T. (2023). TCFLE-8 : a Corpus of Learner Written Productions for French as a Foreign Language and its Application to Automated Essay Scoring. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 3447-3465).

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 677–685.

balanced corpus. We also advance the state-of-the-art performance for the AES task in French by experimenting with two strong baselines.

MOTS-CLÉS : correction automatisée de textes, français langue étrangère, TCF, corpus d'apprenants.

KEYWORDS: automated essay scoring, French as a foreign language, TCF, learner corpus.

1 Introduction

La correction automatisée de textes (CAT – AES pour *automatic essay scoring* en anglais) vise à développer des algorithmes capables d'évaluer la qualité de productions écrites de la même manière que des évaluateurs humains. Les origines de ce domaine remontent aux recherches de Page (1966). Depuis lors, de nombreux chercheurs se sont penchés sur cette question et plusieurs synthèses du domaine ont été publiées récemment (Ramesh & Sanampudi, 2022; Lagakis & Demetriadis, 2021; Klebanov & Madnani, 2021; Uto, 2021; Klebanov & Madnani, 2020; Ke & Ng, 2019; Shermis *et al.*, 2013). En bref, le début du 21e siècle fut marqué par des systèmes reposant soit sur des corpus annotés et sur l'extraction automatisée de caractéristiques linguistiques qui étaient ensuite combinées via des algorithmes d'apprentissage automatisé (Burstein *et al.*, 1998; Rudner & Liang, 2002; Dikli, 2006), soit sur des approches non supervisées recourant souvent à l'analyse sémantique latente (Landauer & Dumais, 1997). Les récentes avancées en CAT ont été rendues possibles par les algorithmes d'apprentissage profond et les grands modèles de langage (Ramesh & Sanampudi, 2022). Néanmoins, ces approches ont également exacerbé le besoin de grands corpus annotés.

Par conséquent, les équipes de recherche en CAT se sont intéressées aux travaux menées par ailleurs sur les corpus d'apprenants, une branche de la linguistique de corpus qui collecte des productions authentiques et informatisées d'apprenants à grande échelle. Des travaux pionniers tels que le *International Corpus of Learner English* (ICLE) (Granger, 1993) et le *European Science Foundation L2 Database* (Perdue, 1993) ont démontré le potentiel de ces collections de données d'apprenants pour la recherche sur l'acquisition d'une langue seconde, mais ce n'est que récemment que davantage de corpus d'apprenants ont été exploités en CAT.

Malheureusement, il n'existe pas de corpus suffisamment important pour le français, ce qui rend la situation de la CAT en français loin d'être encourageante. Les premiers systèmes ont, par conséquent, utilisé des approches non supervisées : Lemaire & Dessus (2001) a utilisé l'analyse sémantique latente pour comparer des essais en français langue maternelle (FLM) avec des passages de référence tirés de manuels, tandis que Zaghouani (2002) capture automatiquement plusieurs caractéristiques d'essais en FLM, lesquelles sont combinées de manière heuristique. Plus récemment, Parslow (2015) a entraîné un classificateur « Naive Bayes » sur un corpus de 200 essais écrits en français langue étrangère (FLE). Enfin, Ranković *et al.* (2020) ont été les premiers à affiner BERT pour la CAT en FLE, mais ils n'ont pas publié leur jeu de données où, de plus, une seule langue maternelle est représentée.

Par conséquent, afin de soutenir le développement de solutions en CAT pour le français, le besoin d'un large corpus fiablement annoté en matière de compétence est patent. Dans cet article, nous apportons deux contributions principales. Premièrement, nous fournissons à la communauté le corpus TCFLE-8 [3], composé de 6 569 productions d'apprenants. Ces textes ont été collectés dans le cadre

3. Ce nom combine le nom de l'examen d'où proviennent les textes, le *Test de connaissance du français*, et l'acronyme FLE. Le 8 fait référence aux huit différentes langues usuelles représentées dans le corpus.

de l'examen officiel du Test de connaissance du français (TCF), l'un des principaux examens de certification pour le FLE. La section 2 décrit en détails les propriétés du TCF et de son évaluation par FEI ainsi que le processus de nettoyage et d'échantillonnage du corpus. Le corpus résultant de ces opérations est présenté à la Section 3, qui décrit les différentes informations disponibles, parmi lesquelles la langue usuelle des candidats, l'évaluation du niveau de compétence à l'écrit des candidats, assignés par au moins 2 évaluateurs sur l'échelle du Cadre européen commun de référence pour les langues (CECR) (Council of Europe, 2001), des informations sur la tâche à accomplir et une annotation automatisée de nombreuses variables linguistiques réalisées par FABRA (Wilkens *et al.*, 2022). La seconde contribution consiste en une série d'expériences en correction automatique de textes sur la base de ce corpus, qui visent à proposer un solide point de référence en vue de futures recherches. Ces expériences sont décrites à la section 4.

2 Méthodologie de conception du corpus

2.1 Collecte des données

TCFLE-8 étant un corpus de productions écrites de candidats au TCF. Il a été collecté par l'un des opérateurs effectuant la certification officielle en FLE : *France Education International* (FEI). FEI est un opérateur français placé sous la tutelle du ministère de l'Éducation nationale et de la Jeunesse. Avec un effectif de plus de 250 salariés et un réseau de plus de 1 000 experts, FEI intervient dans différents domaines de la coopération en matière d'éducation et de formation et contribue à la promotion de la langue française et de la francophonie. FEI propose une large gamme de certifications en français alignées sur les six niveaux du CECR : diplôme initial de langue française (DILF), diplôme d'études en langue française (DELF), diplôme d'études approfondies en langue française (DALF) et test de connaissance du français (TCF). Environ 650 000 candidats se présentent chaque année à l'un de ces examens dans plus de 180 pays.

Comme son nom l'indique, TCFLE-8 est basé sur le TCF, un test linéaire aligné sur les six niveaux du CECR. Le TCF est principalement utilisé dans les contextes d'admission à des études universitaires, de migration et d'accès à la citoyenneté. Sa composante écrite, composée de trois tâches indépendantes, est passée chaque année par 120 000 candidats, dont 60 % passent l'examen sur ordinateur.

Les trois tâches visent à tester les capacités des candidats à s'exprimer en français à l'écrit et nécessitent de rédiger, par exemple, un message, un article, un courrier ou un texte comparant deux points de vue [4]. Ces tâches sont corrigées par des évaluateurs experts. FEI dispose d'un panel d'une centaine de correcteurs, recrutés sur la base de leur profil professionnel (enseignants expérimentés ayant une expérience préalable de l'évaluation en français). Les candidats évaluateurs passent un test psychométrique validant leurs compétences en évaluation de l'écrit et suivent une formation de deux jours. À l'issue de cette procédure, le recrutement est confirmé ou non. Pour garantir la fidélité des corrections à long terme, les indices de fidélité des évaluateurs sont évalués périodiquement et une décision est prise quant à leur maintien dans le panel. En outre, pour garantir la fidélité au niveau des candidats, FEI adopte une approche de double notation indépendante. En cas de désaccord, un troisième évaluateur vient en renfort pour évaluer indépendamment les trois productions. Le niveau final du candidat est établi sur la base de la fréquence des niveaux du CECR attribués aux trois

4. Plus de détails sur la nature des tâches sont disponibles sur le site du TCF : https://www.france-education-international.fr/test/tcf-tout-public?langue=fr

productions du candidat.

Malgré cette stricte procédure, la compétence langagière étant multidimensionnelle (Bachman, 1990; Bachman & Palmer, 2010; Oller & Hinofotis, 1980; Vollmer & Sang, 1983) et mesurable (Vollmer & Caroll, 1983), mesurer les compétences rédactionnelles implique de prendre en compte différentes facettes : les compétences du candidat, l'indulgence ou la sévérité de l'évaluateur et la difficulté de la tâche. À cette fin, « l'utilisation de modèles de Rasch à multi-facettes (MRMF) est une approche psychométrique qui établit un cadre cohérent pour tirer des conclusions fiables, valides et justes des évaluations effectuées par les évaluateurs, répondant ainsi au problème des évaluations humaines faillibles » (Eckes, 2009). Nous avons donc appliqué le MRMF à l'ensemble de la base de données des examens TCF afin d'identifier les évaluations humaines faillibles et d'éviter de les intégrer dans le corpus.

2.2 Nettoyage des données

Les données collectées par FEI ont dû être nettoyées à différents égards. Tout d'abord, la détection des valeurs aberrantes nous a conduit à supprimer les réponses des candidats qui n'atteignaient pas le niveau A1, étaient des copies de la question, ou étaient trop courtes, trop longues ou encore hors sujet. Ensuite, nous avons exploité les informations du modèle multi-facettes de Rasch afin de détecter les textes pour lesquels les évaluateurs humains semblaient ne pas avoir fourni un jugement fiable. À cette fin, nous avons comparé les scores CECR originaux des évaluateurs FEI et les scores ajustés par la méthode MRMF et avons supprimé tous les essais dont la valeur des résidus standardisés était supérieure à 4. En outre, nous avons également supprimé les productions dont l'évaluation semblait peu fiable (par exemple, pour les candidats qui se situent à la limite entre deux niveaux). Pour ce faire, nous avons supprimé tous les cas où les deux évaluateurs n'étaient pas d'accord entre eux ni avec la note finale du candidat, et nous avons également supprimé les cas où il y avait une distance de trois niveaux du CECR entre la note la plus basse et la note la plus élevée attribuée à l'une des trois tâches.

Après ce processus, nous avons attribué à chaque production le niveau CECR du candidat, lorsqu'au moins un des évaluateurs avait également donné ce niveau à la production. Par ailleurs, si les deux évaluateurs avaient attribué le même niveau à la production, nous lui avons attribué ce niveau (même si ce niveau n'était pas identique au niveau global du candidat). Toute production ne répondant pas à l'un de ces deux critères a été supprimée.

Après l'élimination des valeurs aberrantes, l'étape suivante a consisté à obtenir un échantillon représentatif de l'ensemble des tâches du TCF disponibles. Pour une représentation équilibrée, le niveau CECR du texte est une variable évidente à contrôler. En outre, nous avons contrôlé la langue usuelle [5], dans le but d'obtenir une représentativité des langues usuelles les plus fréquentes. Comme les cinq premières étaient toutes européennes et que la sixième était le kabyle, une langue afro-asiatique, nous avons également inclus le chinois et le japonais afin d'obtenir une meilleure représentativité des différentes familles typologiques de langues. Nous avons donc lancé une procédure d'échantillonnage aléatoire stratifié en contrôlant les 6 niveaux du CECR et la langue usuelle du candidat. La Table 1 décrit le corpus résultant de cette procédure. Pour finir, le corpus a été anonymisé et pseudo-anonymisé à l'aide de l'outil MAPA (Gianola *et al.*, 2020), afin de préserver l'identité des auteurs des textes qui contiennent parfois des informations personnelles.

5. La langue usuelle est la langue que le candidat a indiqué dans le formulaire d'inscription comme étant celle qu'il utilise habituellement.

Langue	A1	A2	B1	B2	C1	C2	Total
JPN	8	135	171	170	48	2	534
CHI	34	165	244	189	45	4	681
SPA	124	187	175	182	178	58	904
ARA	135	160	163	153	160	135	906
POR	102	187	182	191	172	38	872
ENG	125	163	167	165	169	128	917
RUS	103	198	183	196	180	29	889
KAB	58	180	181	181	175	91	866
Total	689	1375	1466	1427	1127	485	6569

TABLE 1 – Nombre de textes en fonction de la langue usuelle et du niveau CECR (les codes des langues suivent la norme ISO639-2)

3 Présentation du corpus

À la fin du processus de compilation, le corpus TCFLE-8 comprend 6 569 essais (581 333 mots). La Table 1 présente des chiffres plus précis sur la proportion de textes par niveau du CECR et par langue usuelle. Les niveaux extrêmes (A1 et C2) sont moins représentés dans le corpus. Cela s'explique par deux facteurs : (1) peu d'apprenants de niveau A1 cherchent à passer une certification en langue, car ce niveau est rarement suffisant à des fins officielles (par exemple, pour l'obtention d'un emploi ou d'un visa), et (2) il est extrêmement difficile d'atteindre le niveau C2 dans une langue étrangère.

Dans la version anglaise originale de cet article (publié à EMNLP), le lecteur trouvera davantage de détails sur le corpus, notamment en matière de représentation de genres, de distribution des tâches et de longueur des productions. En résumé, on note que 58% des textes ont été écrits par des femmes et que cette proportion varie selon les niveaux CECR. Au niveau des tâches, les trois types de tâches sont relativement uniformément représentés, ce qui était espéré au vu de la procédure d'échantillonnage. Par ailleurs, une comparaison systématique entre TCFLE-8 et les autres corpus existants révèle qu'il s'agit du plus grand corpus d'apprenants de FLE adapté à la CAT – à la fois en termes de taille et de représentativité des L1 (ici, langue usuelle) –, du troisième plus large corpus de productions écrites de candidats à notre connaissance, toute langue confondue, et que ses couches d'annotation fournissent les informations les plus riches. En effet, non seulement, il couvre les 6 niveaux du CECR, mais a également fait l'objet d'une annotation linguistique visant à décrire les compétences des apprenants avec plus de 400 variables (chacune associées à 18 agrégateurs statistiques, comme la moyenne, la médiane, l'écart-type, etc. ce qui donne plus de 5 000 caractéristiques). Cette annotation a été réalisée automatiquement à l'aide de la boîte à outils FABRA (Wilkens *et al.*, 2022).

Par ailleurs, en complément du niveau CECR consolidé, issu de la procédure décrite plus haut, TCFLE-8 comprend également le niveau CECR atteint par chaque candidat au terme des trois productions écrites. Ce score correspond au niveau officiel du CECR attribué au candidat pour la partie écrite de l'examen du TCF. Le Kappa quadratique de Cohen pondéré (KQP) entre ces deux scores (niveau CECR du texte et niveau CECR du candidat) atteint 0,98. On s'attend à ce que cette valeur soit élevée, mais pas égale à 1, en raison des cas où les candidats ne peuvent pas maintenir un niveau constant de qualité durant l'épreuve. En outre, les notes attribuées par les deux évaluateurs de FEI dans le cadre de la procédure de double évaluation sont également disponibles. Elles ont un KQP de 0,71 entre elles et de 0,84 avec le niveau CECR consolidé de l'essai. Enfin, la nature de la tâche et sa position

dans la séquence des trois tâches du TCF sont également rapportées. Ces informations permettent de contextualiser la réponse du candidat.

4 Résultats pour la CAT en FLE

Dans cette section, nous rapportons rapidement nos expériences visant à évaluer l'utilité du corpus TCFLE-8 pour l'entraînement des systèmes de CAT. En d'autres termes, prédire le niveau CECR des textes rédigés par les candidats au TCF revient donc à rendre possible l'automatisation de la correction des productions écrites du TCF, mais nous espérons que le corpus puisse soutenir plus largement la correction de textes automatisée pour le français.

À cette fin, nous explorons deux approches : l'apprentissage profond, étant donné que la plupart des systèmes AES s'appuient sur des réseaux neuronaux (Ramesh & Sanampudi, 2022), et l'apprentissage automatique basé sur des variables. Pour le modèle d'apprentissage profond, nous avons utilisé CamemBERT (Martin *et al.*, 2020). Pour l'apprentissage non-neuronal, nous utilisons XGBoost, d'une part, et un simple modèle de régression logistique, d'autre part. Les performances de ces modèles sont présentées à la Table 2. Il apparaît clairement que le modèle basé sur CamemBERT obtient une meilleure exactitude et un meilleur score F1 que les deux autres modèles. Malgré tout, on peut constater qu'il y a encore une marge de progression lorsque l'on compare les résultats de CamemBERT avec l'évaluation des experts de FEI (colonne « Évaluateurs »). Néanmoins, ce modèle est proche de la performance des évaluateurs lorsque l'on considère la relation d'ordinalité entre les niveaux. Celle-ci est capturée à l'aide de la métrique κ quadratique pondéré (KQP), qui évalue l'accord entre le système et les annotateurs, en tenant compte de la distance entre les 6 niveaux du CECR, ainsi que de la mesure d'exactitude contiguë, calculée de la même manière que l'exactitude contiguë, mais où les erreurs d'un niveau de différence ne sont pas prises en compte.

	CamemBERT	**XGBoost**	**Logistique**	**Évaluateurs**
KQP	**0,88** (0,01)	0,79 (0,02)	0,69 (0,02)	0,93 (0,01)
Exactitude	**0,57** (0,01)	0,46 (0,01)	0,37 (0,01)	0,76 (0,01)
$\text{Exactitude}_{\text{Contiguë}}$	**0,98** (0,01)	0,92 (0,02)	0,80 (0,01)	0,99 (0,01)
$\text{F1}_{\text{pondérée}}$	**0,56** (0,01)	0,46 (0,02)	0,36 (0,02)	0,76 (0,01)
A1_{F1}	**0,63** (0,01)	0,59 (0,04)	0,54 (0,06)	0,76 (0,02)
A2_{F1}	**0,57** (0,04)	0,53 (0,01)	0,40 (0,05)	0,76 (0,03)
B1_{F1}	**0,56** (0,04)	0,45 (0,05)	0,32 (0,02)	0,75 (0,01)
B2_{F1}	**0,56** (0,04)	0,43 (0,03)	0,34 (0,03)	0,76 (0,02)
C1_{F1}	**0,56** (0,04)	0,42 (0,03)	0,30 (0,05)	0,77 (0,02)
C2_{F1}	**0,48** (0,09)	0,19 (0,07)	0,31 (0,02)	0,80 (0,04)

TABLE 2 – Moyenne et écart-type des performances des 3 modèles ainsi que la performance des évaluateurs humains sur TCFLE-8.

TCFLE-8 étant un nouveau corpus pour la langue française, nous ne pouvons pas comparer nos résultats avec les travaux précédents, en raison d'une différence considérable au niveau de la taille de ces corpus. Dans la littérature en CAT pour le français nous n'avons identifié que deux articles portant sur l'identification de la compétence écrite en FLE (cf. Section 1). Tout d'abord, Parslow (2015) a rapporté des scores F1 allant de $0,51$ à $0,74$ pour les niveaux A1 à B2. Deuxièmement, Ranković

et al. (2020) ont utilisé les couches intermédiaires de CamemBERT comme caractéristiques pour prédire le niveau dans un corpus de 100 essais et ont rapporté des MSE allant de $0,35$ à $0,55$.

5 Conclusion

Dans ce travail, nous avons présenté TCFLE-8, un corpus de 6 569 essais de candidats écrits pendant le test de connaissance du français (TCF), incluant 8 langues usuelles différentes. Cet article a décrit la collecte des données par France Education International (FEI) et les différentes étapes de nettoyage des données. Au final, nous obtenons le plus grand corpus en français ciblant le FLE pour la CAT. Ce corpus, ainsi que ses métadonnées (essais, métadonnées et annotations) sont à la disposition de la communauté. En explorant l'utilité de TCFLE-8 pour la tâche de CAT en FLE, nous avons appliqué différents algorithmes d'apprentissage automatique. CamemBERT apparaît comme le plus précis des trois.

Enfin, l'intérêt du corpus TCFLE-8 dépasse les frontières de l'AES, car le fait qu'il s'agisse d'un grand corpus d'apprenants, annoté avec les niveaux du CECR, ouvre de nombreuses pistes de recherches en TAL, mais aussi en acquisition du langage et en linguistique de corpus. Ainsi, il pourrait soutenir des recherches pour le développement de matériel pédagogique, qu'il s'agisse de dictionnaires (Longman, 2002), d'activités axées sur les difficultés et les erreurs courantes des apprenants (Kaszubski, 1998; Reppen, 2010), de logiciels d'apprentissage des langues assisté par ordinateur (Granger, 2003) ou d'aides à l'écriture en L2 (Link *et al.*, 2014). Avec 8 langues usuelles différentes, ce corpus pourrait également être utile pour des études interlinguistiques ciblant les mécanismes de transfert et l'influence de la L1 sur la production de la L2 (Golden *et al.*, 2017; Werner *et al.*, 2020) ou pour l'identification automatique de la langue maternelle (Tetreault *et al.*, 2013). Enfin, une autre application possible de ce nouveau corpus est la détection et la correction d'erreurs (Dahlmeier *et al.*, 2013), que nous étudions actuellement dans le cadre d'un travail futur sur TCFLE-8.

Références

BACHMAN L. & PALMER A. (2010). *Language assessment in practice : developing language assessments and justifying their use in the real world*. Oxford applied linguistics. Oxford : Oxford Univ. Press.

BACHMAN L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

BURSTEIN J., KUKICH K., WOLFF S., LU C. & CHODOROW M. (1998). Computer analysis of essays. In *NCME Symposium on automated Scoring*.

COUNCIL OF EUROPE (2001). *Common European Framework of Reference for Lan- guages : Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

DAHLMEIER D., NG H. T. & WU S. M. (2013). Building a Large Annotated Corpus of Learner English : The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 22–31, Atlanta, Georgia : Association for Computational Linguistics.

DIKLI S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, **5**(1).

ECKES T. (2009). *Quantitative Data Analysis for Language Assessment Volume I : Fundamental Techniques*. Routledge, 1 édition. DOI : 10.4324/9781315187815.

GIANOLA L., AJAUSKS Ē., ARRANZ V., GIBERT O. D. & MELERO M. (2020). Automatic removal of identifying information in official eu languages for public administrations : The mapa project. In *33rd International Conference on Legal Knowledge and Information Systems (JURIX 2020) : proceedings, Dec 2020, Brno, Prague, Czech Republic*, volume 334, p. 223–226 : IOS Press.

GOLDEN A., JARVIS S. & TENFJORD K. (2017). *Crosslinguistic Influence and Distinctive Patterns of Language Learning : Findings and Insights from a Learner Corpus*. Multilingual Matters.

GRANGER S. (1993). The International Corpus of Learner English. In *The European English Messenger*, p.3̃4.

GRANGER S. (2003). Error-tagged Learner Corpora and CALL : A Promising Synergy. *CALICO Journal*, **20**(3), 465–480.

KASZUBSKI P. (1998). Learner corpora : The cross-roads of linguistic norm. *TALC98 Proceedings*, p. 24–27.

KE Z. & NG V. (2019). Automated essay scoring : A survey of the state of the art. In *IJCAI*, volume 19, p. 6300–6308.

KLEBANOV B. B. & MADNANI N. (2020). Automated evaluation of writing–50 years and counting. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, p. 7796–7810.

KLEBANOV B. B. & MADNANI N. (2021). Automated Essay Scoring. *Synthesis Lectures on Human Language Technologies*, **14**(5), 1–314.

LAGAKIS P. & DEMETRIADIS S. (2021). Automated essay scoring : A review of the field. In *2021 International Conference on Computer, Information and Telecommunication Systems (CITS)*, p. 1–6 : IEEE.

LANDAUER T. K. & DUMAIS S. T. (1997). A solution to Plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, **104**(2), 211.

LEMAIRE B. & DESSUS P. (2001). A System to Assess the Semantic Content of Student Essays. *Journal of Educational Computing Research*, **24**(3), 305–320. DOI : 10.2190/G649-0R9C-C021-P6X3.

LINK S., DURSUN A., KARAKAYA K. & HEGELHEIMER V. (2014). Towards Better ESL Practices for Implementing Automated Writing Evaluation. *Calico Journal*, **31**(3).

LONGMAN (2002). *Longman Essential Activator*. Harlow : Pearson ESL.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

OLLER, JR J. W. & HINOFOTIS F. B. (1980). Two mutually exclusive hypotheses about second language ability : factor analytic studies of a variety of language subtests.

PAGE E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, **47**(5), 238–243.

PARSLOW N. (2015). Automated Analysis of L2 French Writing : a preliminary study. Mémoire de master. Publisher : Unpublished.

PERDUE C. (1993). Comment rendre compte de la "logique" de l'acquisition d'une langue étrangère par l'adulte. *Études de Linguistique Appliquée*, **92**(1), 8–23.

RAMESH D. & SANAMPUDI S. K. (2022). An automated essay scoring systems : a systematic literature review. *Artificial Intelligence Review*, **55**(3), 2495–2527.

RANKOVIĆ B., SMIRNOW S., JAGGI M. & TOMASIK M. J. (2020). Automated Essay Scoring in Foreign Language Students Based on Deep Contextualised Word Representations. In *LAK20-10th International Conference on Learning Analytics & Knowledge*. Issue : CONF.

REPPEN R. (2010). *Using Corpora in the Language Classroom*. Cambridge University Press.

RUDNER L. M. & LIANG T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, **1**(2).

SHERMIS M. D., BURSTEIN J. & BURSKY S. A. (2013). Introduction to automated essay evaluation. In *Handbook of automated essay evaluation*, p. 23–37. Routledge.

TETREAULT J., BLANCHARD D. & CAHILL A. (2013). A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 48–57, Atlanta, Georgia : Association for Computational Linguistics.

UTO M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika*, **48**(2), 459–484.

VOLLMER, J H. & CAROLL J. B. (1983). Psychometric theory and language testing. In J. W. OLLER, Éd., *Issues in language testing research*, p. 29–79. Rowley, Mass. : Newbury House. 00000.

VOLLMER, J H. & SANG F. (1983). Competing hypotheses about second language ability : a plea of caution. In J. W. OLLER, Éd., *Issues in language testing research*. Rowley, Mass. : Newbury House. 00000.

WERNER V., FUCHS R. & GÖTZ S. (2020). L1 influence vs. universal mechanisms : An SLA-driven corpus study on temporal expression. In *Learner Corpus Research Meets Second Language Acquisition*, p. 39–66. Cambridge University Press.

WILKENS R., ALFTER D., WANG X., PINTARD A., TACK A., YANCEY K. P. & FRANÇOIS T. (2022). Fabra : French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 1217–1233.

ZAGHOUANI W. (2002). AUTO-ÉVAL : vers un modèle d'évaluation automatique des textes. In *Actes du colloque des étudiants en sciences du langage*, p.1̃6, Montréal, Canada : Université du Québec à Montréal.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Technologies de la parole et données de terrain : le cas du créole haïtien

William N. Havard[1,2], Renauld Govain[3], Daphne Gonçalves Teixeira[1], Benjamin Lecouteux[2], Emmanuel Schang[1]
[1] LLL, Université d'Orléans, CNRS, 45000 Orléans, France
[2] LIG, Université Grenoble Alpes, CNRS, Grenoble INP, 38000 Grenoble, France
[3] LangSé, Université d'État d'Haïti, Port-au-Prince, Haïti
william.havard@univ-orleans.fr

RÉSUMÉ

Nous utilisons des données de terrain en créole haïtien, récoltées il y a 40 ans sur cassettes puis numérisées, pour entraîner un modèle natif d'apprentissage auto-supervisé (SSL) de la parole (WAV2VEC2) en haïtien. Nous utilisons une approche de pré-entraînement continu (CPT) sur des modèles SSL pré-entraînés de deux langues étrangères : la langue lexificatrice – le français – et une langue non apparentée – l'anglais. Nous comparons les performances de ces trois modèles SSL, et de deux autres modèles SSL étrangers directement affinés, sur une tâche de reconnaissance de la parole. Nos résultats montrent que le modèle le plus performant est celui qui a été entraîné en utilisant une approche CPT sur la langue lexificatrice, suivi par le modèle natif. Nous concluons que l'approche de "mobilisation des archives" préconisée par (Bird, 2020) est une voie prometteuse pour concevoir des technologies vocales pour de nouvelles langues.

ABSTRACT

Speech Technologies with Fieldwork Recordings : the case of Haitian Creole

We use fieldwork recordings in Haitian Creole, collected 40 years ago on cassettes and then digitised, to train a native self-supervised learning (SSL) model of speech (WAV2VEC2) in Haitian. We use a continuous pre-training (CPT) approach on pre-trained SSL models of two foreign languages : the lexifier language – French – and an unrelated language – English. We compare the performance of these three SSL models, and of two other directly finetuned foreign SSL models, on a speech recognition task. Our results show that the best-performing model is the one trained using a CPT approach on the lexifier language, followed by the native model. We conclude that the "mobilise the archive" approach advocated by (Bird, 2020) is a promising avenue for designing speech technologies for new languages.

MOTS-CLÉS : créole haïtien, enregistrement de terrain, modèles auto-supervisés, reconnaissance de la parole.

KEYWORDS: Haitian Creole, fieldwork recordings, self-supervised model, speech recognition.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 686–694.

1 Introduction

La plupart des langues peu dotées ne le sont souvent que du point de vue des informaticiens [1] : elles disposent souvent de nombreuses ressources collectées au fil des ans par des linguistes, des missionnaires religieux et, plus généralement, par la communauté des locuteurs elle-même (Bird, 2020). Les données ne sont souvent pas facilement accessibles (p. ex. sous un format numérique), mais elles existent néanmoins. La question à laquelle nous tentons de répondre dans cet article est la suivante : jusqu'où pouvons-nous aller avec les modèles de traitement de la parole état-de-l'art en utilisant *uniquement* des données de terrain *déjà existantes* ?

Par "données de terrain", nous entendons des données qui n'ont pas été collectées à l'origine pour servir de données d'entraînement pour des applications informatiques (p. ex. la reconnaissance automatique de la parole, RAP), mais qui ont été collectées à des fins linguistiques (p. ex. l'étude des variations dialectales). Dans cet article, nous nous concentrons sur des données orales en créole haïtien (*kreyòl ayisyen*), constituées d'entretiens enregistrés entre des linguistes et leurs collaborateurs. Le créole haïtien est un créole à base lexicale française (le français est sa langue lexificatrice, c'est à dire, la langue lui a apporté la plupart de son vocabulaire, voir Hazael-Massieux 2012), parlé par 13M de locuteurs (Simons & Fennig, 2023) à Haïti et par la diaspora haïtienne, principalement aux États-Unis d'Amérique.

La majorité des données que nous utilisons dans cet article (voir la section 2) a été collectée il y a 40 ans avec des magnétophones pour étudier les variations dialectales en haïtien, en mettant l'accent sur les variations lexicales. Contrairement aux livres audio couramment utilisés pour entraîner les modèles neuronaux (p. ex. Librispeech, Panayotov *et al.* 2015) qui jouissent d'une haute qualité d'enregistrement, les données que nous utilisons sont particulièrement bruitées : réverbération, echo, bruits ambiants (p. ex. poules, coqs, poussins, voitures, passants, etc.). Pourtant, ce type de données représente la majorité des données disponibles pour la plupart des langues du monde. La collecte et la transcription des données étant un processus coûteux, [2] ne pourrions-nous pas utiliser — comme le préconise (Bird, 2020) dans l'approche consistant à "mobiliser les archives" (*mobilise the archive*) — des données de terrain déjà existantes (et potentiellement anciennes) et les ré-utiliser pour des applications informatiques ?

Questions de recherche. Plus précisément, les questions que nous abordons dans cet article sont les suivantes : **(a)** Des données de terrain, bien que bruités (mais écologiques) seraient-elles utilisables pour entraîner des modèles d'apprentissage auto-supervisé (SSL) de la parole (p. ex. WAV2VEC2, Baevski *et al.* 2020) ? **(b)** Doit-on entraîner ces modèles à partir de zéro ou doit-on utiliser des approches de pré-entraînement continu (*continuous pre-training*, CPT, Nowakowski *et al.*, 2023; Gururangan *et al.*, 2020) ? **(c)** Quelle quantité de données d'entraînement est nécessaire pour affiner (*finetune*) les modèles sur une tâche de RAP ? Enfin, **(d)** est-il possible d'entraîner de tels modèles avec un budget limité ? (c'est-à-dire en utilisant un seul GPU et non 64 comme c'est le cas pour Baevski *et al.* 2020).

En outre, comme nous travaillons dans le contexte des langues créoles, nous visons également à explorer l'influence de la langue lexificatrice (comme un cas clair de langues apparentées) et explorons

1. Voir §§ 2 et 2.1 de (Bird, 2020) sur la notion de "*zero resource*" et la vision centrée "données" du traitement automatique des langues et de la parole.

2. Himmelmann (2018) rapporte que la transcription de 1 minute de parole peut prendre de 10 à 150 minutes, selon la langue, les connaissances du linguiste et le niveau de transcription (phonétique, phonologique, orthographique) et d'annotation annexe (morphologique, syntaxique, etc.).

(e) si l'approche CPT doit être effectuée sur des modèles SSL de la langue lexificatrice (p. ex. le français dans le cas du créole haïtien), ou si des modèles entraînés sur une langue non apparentée (p. ex. l'anglais dans le cas du créole haïtien) fonctionnent également ?

Travaux connexes. Le domaine du traitement de la parole pour les langues créoles par le biais de modèles neuronaux est relativement nouveau. Les seuls travaux de traitement de la parole pour ces langues sont ceux de (Breiter, 2014) pour le créole haïtien, ceux de (Macaire *et al.*, 2022) pour les créoles guadeloupéen et mauricien, et de (Gooda Sahib-Kaudeer *et al.*, 2019) pour le créole mauricien (avec un accent mis sur le domaine médical). Ainsi, le traitement de la parole pour les langues créoles — fussent-elles à base lexicale française, anglaise, portuguaises, etc. — reste largement inexploré.

Sans rapport direct avec le traitement de la parole pour les langues créoles — mais en rapport direct avec notre contexte méthodologique — Nowakowski *et al.* (2023) a exploré des approches de pré-entraînement continu, suivies d'une tâche d'affinage pour la reconnaissance vocale en ainu (langue native du nord du Japon) en utilisant d'anciennes données de terrain. Cependant, contrairement à l'objectif que nous nous fixons, ils n'entraînent pas leurs modèles avec un budget limité car (i) ils utilisent 4 GPU, (ii) utilisent le modèle XLSR-53 (Conneau *et al.*, 2021) qui est basé sur WAV2VEC2-LARGE et pré-entraîné sur 56k heures de données, et (iii) font un affinage multilingue par lequel le modèle de RAP n'est pas seulement entraîné sur la langue cible (ainu), mais sur plusieurs langues à la fois (anglais, japonais, en plus de l'ainu). Nous visons une approche plus stricte qui n'utilise que des données de terrain à toutes les étapes.

2 Données

ALH Nous avons utilisé le *Atlas Linguistique d'Haïti* (Fattier, 1998), constitué d'un ensemble de 499 enregistrements audio en créole haïtien collectés à Haïti entre 1978 et 1987 dans le but de créer un atlas linguistique. Les enregistrements ont été réalisés à l'origine sur des cassettes audio avec des magnétophones, puis numérisés dans les années 2010 par la Bibliothèque nationale de France. Chaque enregistrement dure en moyenne 45 minutes et consiste en un entretien dirigé entre un ou plusieurs enquêteurs qui demandent des mots ou des phrases à leurs collaborateurs Haïtiens. Ces enregistrements ont été numérisés et mis à disposition sur la plateforme COCOON (FLA and Fattier, 2015). [3] Bien que les enregistrements soient associés à des cahiers de terrain comportant des transcriptions manuscrites partielles (p. ex. transcription phonétique à l'échelle du mot), celles-ci n'ont pas été numérisées (ni alignées avec les enregistrements). Ainsi, ce corpus est entièrement constitué de parole brute.

Nous avons divisé l'ensemble de données (356,3 heures) en trois parties (train/val/test). Les données ont été réparties de manière à ce que l'ensemble de validation contienne au moins 5 heures de données et un minimum de 2 locuteurs inconnus, et l'ensemble de test au moins 5 heures de données et un minimum de 3 locuteurs inconnus. Nous avons obtenu la répartition suivante, qui répondait à nos contraintes : train = 345,6 heures ; val = 5,3 heures, 5 locuteurs inconnus ; et test = 5,4 heures, 8 locuteurs inconnus. [4]

CNCH Le *Corpus du créole haïtien du Nord* (*Corpus of Northern Haitian Creole*, Valdman *et al.*,

3. https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-8ea988d2-bf16-303d-81a0-0c55cc0
4. L'ensemble de test n'a pas été utilisé dans les expériences présentées dans ce document, mais le sera dans les premiers travaux futurs énumérés dans la section 5.

Modèle	Langue SSL	WER ↓	**UER ↓**	Entraînement	Décodage	Classement
SSL-ETRANGER+CPT+FT	FR	**36.8**	21.6	320mn	4-gram	**1**
SSL-NATIF+Ø+FT	HAT	37.4	**21.5**	360mn (max)	3-gram	5
SSL-ETRANGER+CPT+FT	EN	37.5	22.4	320mn	4-gram	6
SSL-ETRANGER+Ø+FT	FR	42.5	24.5	360mn (max)	3-gram	27
SSL-ETRANGER+Ø+FT	EN	50.4	29.0	320mn	3-gram	49

Modèle	Langue SSL	WER ↓	**UER ↓**	Entraînement	Décodage	Classement
SSL-ETRANGER+CPT+FT	FR	**38.2**	**17.1**	320mn	Viterbi	**1**
SSL-NATIF+Ø+FT	HAT	39.8	17.8	360mn (max)	Viterbi	3
SSL-ETRANGER+CPT+FT	EN	40.3	18.6	360mn (max)	Viterbi	6
SSL-ETRANGER+Ø+FT	FR	46.2	21.7	360mn (max)	Viterbi	12
SSL-ETRANGER+Ø+FT	EN	57.1	26.6	360mn (max)	Viterbi	38

TABLE 1 – Architecture qui donnent les meilleures performances en termes de WER (en haut) et de UER (en bas) pour chaque type de modèle affiné. *Classement* montre le rang des modèles de 1 (meilleur) à 200 (pire) lorsque le WER/UER est utilisé comme clé de tri.

2015) [5] comprend 10 entretiens enregistrés, menés au Cap-Haïtien (Nord d'Haïti) pour étudier les variations dialectales par rapport au haïtien standard. Ce corpus a été entièrement transcrit par le linguiste l'ayant récolté. Cependant, nous tenons à mentionner que les transcriptions utilisées sont non-standards et impressionistes, dans le sens où des variations orthographiques déviant de la norme sont utilisées pour retranscrire plus fidèlement la prononciation du locuteur : "*Powoprens*"/"*Potoprens*", Port-au-Prince ; "*eskeu*"/"*eske*", est-ce que ; "*deu*"/"*de*", deux ; etc.). Ces variations pourront donc influer (de manière non favorable) sur le taux d'erreur mot (WER) et caractère (UER).

Nous avons divisé l'ensemble des données (9 heures) en trois parties (train/val/test). Les données ont été réparties de manière à ce que l'ensemble de validation contienne au moins 1 heure de données et un minimum d'un locuteur inconnu, et l'ensemble test au moins 1 heure de données et un minimum d'un locuteur inconnu. Nous avons obtenu la répartition suivante, qui répondait à nos contraintes : train = 6,9 heures ; val = 1,1 heure, 1 locuteur inconnu ; test = 1,0 heure, 2 locuteurs inconnus.

Autre ensemble de données Nous tenons à souligner l'existence d'autres ensembles de données présentant de la parole en créole haïtien, que nous avons volontairement exclus car ils ne consistent pas en des données de terrain : l'ensemble de données Haïti-CMU librement accessible [6] qui contient de la parole lue ($\sim$ 20 heures), principalement des sections de la Bible, qui ne reflètent pas l'utilisation quotidienne de la langue ; et l'ensemble de données propriétaire Babel-IARPA comprenant 203 heures et étant uniquement consistué de "parole conversationelle téléphonique scénarisée" (Andrus *et al.*, 2017).

3 Expériences

Compte tenu de notre contrainte de budget limité, nous nous concentrons uniquement sur l'architecture WAV2VEC2-BASE, excluant ainsi l'entraînement d'un modèle basé sur WAV2VEC2-LARGE, ainsi

5. https://archive.org/details/interview-8-ujf-107-a-ujm-107-a
6. http://www.speech.cs.cmu.edu/haitian/

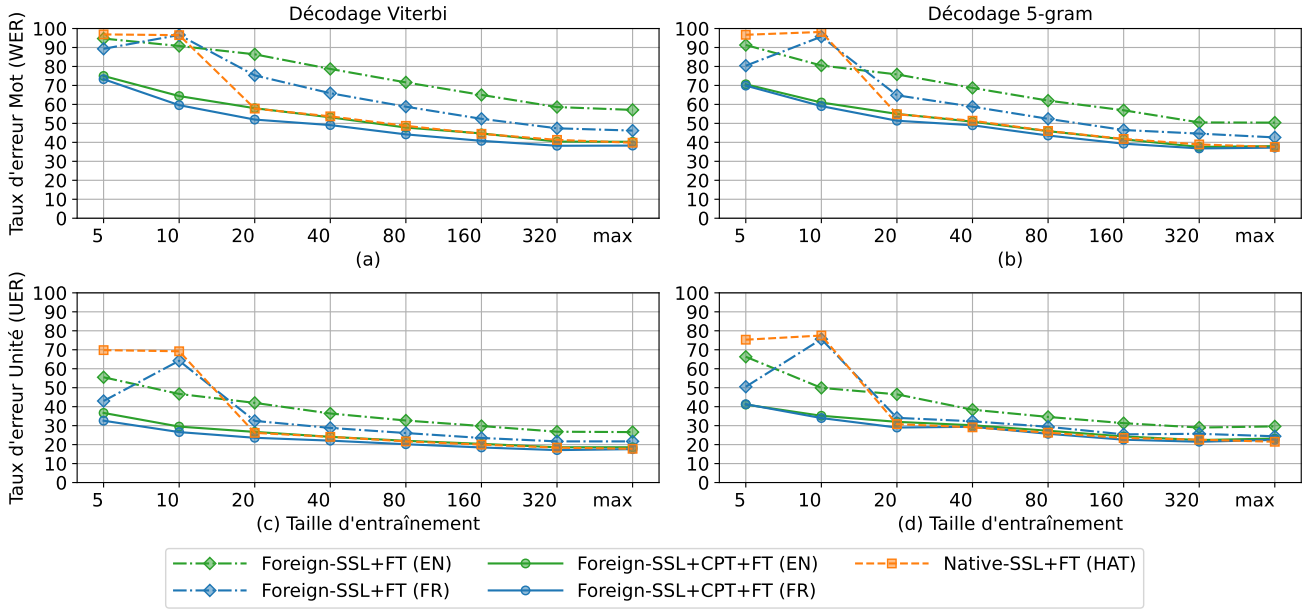


FIGURE 1 – (a, b) Taux d'erreur mot (WER) et (c, d) taux d'erreur unité (UER, au niveau des caractères) des modèles affinés sur une tâche de RAP avec décodage Viterbi (à gauche) et avec LM à 5-gram (à droite) en fonction de la quantité de données CNCH utilisées pour l'entraînement (en minutes, de 5 minutes à *max*, où $max = 6,9$ heures, ~ 360 minutes).

que l'affinage d'un modèle multilingue tel que XLSR-53 qui est basé sur l'architecture WAV2VEC2-LARGE.

Pré-entraînement SSL *natif* et *étranger*. Nous utilisons le corpus ALH pour entraîner nos modèles SSL. Un modèle de détection de l'activité vocale (Pyannote, Bredin *et al.*, 2020) a été utilisé pour isoler les sections correspondant à de la parole des bruits environnants, ce qui a permis d'obtenir 229h de sections parlées. Les segments résultants, plutôt courts ($\sim$ 2.3s), ont été fusionnés jusqu'à ce que les segments concaténés atteignent 19s en moyenne ($\sim 2.4\text{s} \pm 5.8\text{s}$). Les modèles WAV2VEC2 ont été entraînés sur un seul GPU [7] en accumulant le gradient pour 16 passes. Les modèles ont été entraînés jusqu'à convergence, définie soit comme le point d'intersection des courbes d'entraînement et de validation, soit comme le point où celles-ci sont restées stables pendant 10 000 passes.

Nous avons entraîné trois modèles. Le premier, entraîné à partir de zéro, que nous appelons désormais SSL-NATIF+Ø, puisque celui-ci n'a vu que du haïtien et ne se base pas sur un modèle existant dans le cadre d'un approche de pré-entrainement continu (+Ø). Les deux autres modèles ont été entrainés à partir des modèles existants dans le cadre d'une approche de pré-entraînement continu : l'un basé sur un modèle français (WAV2VEC2-FR-7K-BASE, pré-entraîné sur 7k heures en français, Parcollet *et al.* 2023), et l'autre à partir d'un modèle anglais (WAV2VEC2-BASE, Baevski *et al.* 2020) pré-entraîné sur Librispeech 960 (Panayotov *et al.*, 2015). Ces modèles sont appelés SSL-ETRANGER+CPT EN ou SSL-ETRANGER+CPT FR puisqu'ils ont été pré-entraînés sur une langue étrangère auparavant (soit de l'anglais, soit du français) et ont bénéficié d'une approche de pré-entrainement continu (+CPT) pendant laquelle ils ont été entraînés à modéliser de la parole en haïtien.

Affinage sur un tâche de RAP Nous avons affiné (+FT) les modèles pré-entraînés sur le corpus CNCH. 3 modèles de RAP ont été affinés à partir de modèles ayant vu du haitïen au pré-entraînement :

7. 32Gb Nvidia Tesla V100 ou 45Gb Nvidia A40 selon la disponibilité.

le modèle SSL-NATIF+Ø (appelé SSL-NATIF+Ø+FT après affinage), et les deux modèles SSL-ETRANGER+CPT basés sur de l'anglais ou du français (appelés SSL-ETRANGER+CPT+FT EN ou FR). En plus de ceux-ci, afin de comprendre la pertinence (ou non) d'un pré-entraînement continu sur des données de terrain, nous avons également affiné directement les modèles SSL-ETRANGER sans utiliser une approche CPT : SSL-ETRANGER+Ø+FT (EN ou FR). Ces modèles nous permettront ainsi de voir si le pré-entraînement sur des données de terrain permet de mieux transférer sur d'autres données de terrain dans une tâche de RAP ou non.

Afin de comprendre l'impact de la taille de l'entraînement sur les performances finales des modèles, nous utilisons différentes tailles d'entraînement : max (6.9 heures), 320, 160, 80, 40, 20, 10, et 5 minutes. Chaque taille d'entraînement inclut les tailles précédentes (par exemple, max $\supset$ ... $\supset$ 10 $\supset$ 5). Chaque modèle est affiné pour 20k passes. [8] Pour éviter le sur-entraînement, les paramètres ont été gelés pendant les 10k premières passes. Le meilleur modèle est sélectionné sur la base du WER le plus bas sur l'ensemble de validation. Le texte a été mis en minuscules et les diacritiques ont été supprimés (en raison d'une utilisation variable). Nous entraînons également des modèles de langue (LM) de 2 à 5 grammes sur les transcriptions pour chaque taille d'entraînement à l'aide de KenLM (Heafield, 2011), ce qui donne 32 LM différents (4 taille de n–gram $\times$ 8 taille de corpus d'affinage).

4 Résultats & Discussion

Nous avons utilisé l'outil SCTK [9] pour calculer le taux d'erreur mot (WER) et le taux d'erreur d'unité (UER, au niveau du caractère). Nous avons utilisé un décodage Viterbi standard ainsi qu'un réordonnancement a posteriori (*rescoring*) avec des LM de 2 à 5 grammes. Cela a permis d'obtenir 5 modèles $\times$ 8 tailles d'entraînement $\times$ (1 Viterbi + 4 ngram) décodages = 200 stratégies de décodage. Pour plus de clarté, seuls les rescores Viterbi et LM 5-grammes sont présentés dans la Fig. 1, et la meilleure configuration pour chacun des 5 types de modèles est présentée dans le Tab. 1.

Nos résultats montrent **(d)** qu'il est possible d'entraîner des modèles compétitifs avec un budget limité en utilisant un seul GPU et que **(a)** l'utilisation de données de terrain pour entraîner des modèles SSL de la parole est efficace. Bien que ces données soient intrinsèquement bruitées — par opposition aux livres audio ou aux discours radiodiffusés couramment utilisés pour entraîner les modèles SSL — le modèle haïtien SSL-NATIF+Ø que nous avons entraîné est resté très compétitif par rapport à d'autres approches. Ceci est particulièrement intéressant dans le cas des langues à faibles ressources, telles que la plupart des créoles à base française parlés dans les Caraïbes (haïtien, guadeloupéen, saint-lucien, etc.) ou en Amérique du Sud (guyanais). Cela signifie qu'il n'est pas nécessaire de collecter de nouvelles données, mais que les anciennes données enregistrées sur bande magnétique, une fois numérisées, peuvent être réutilisées à cette fin. Cela permettrait à de nombreuses langues du monde de disposer de modèles de traitement de la parole à la pointe de la technologie.

Quant à savoir **(b)** si nous devrions affiner les modèles SSL qui ont été pré-entraînés à partir de zéro ou les modèles pré-entraînés en utilisant une approche CPT, nos résultats montrent que les modèles entraînés dans une approche CPT montrent un léger avantage sur les modèles natifs entraînés à partir de zéro (-1.6 WER, et -0.7 UER, décodage de Viterbi, en utilisant l'UER le plus bas comme clef de tri). Cependant, nos résultats montrent que **(e)** cet avantage n'est vrai que lorsque le modèle utilisé

8. Compte tenu du peu de données dont nous disposons, les modèles convergent rapidement, restent stables et n'évoluent pas après 20k étapes, d'où cette valeur.

9. https://github.com/usnistgov/SCTK

pour le pre-entraînement continue est *celui de la langue lexificatrice* (ici, le français). Cet avantage semble disparaître lorsque ce n'est pas le cas, car le modèle affiné à partir d'une autre langue (ici, l'anglais) a généralement de moins bonnes performances qu'un modèle affiné à partir de la langue lexificatrice (+2.1 WER, +1.5 UER, *id.*) ou à partir de la langue cible (+0.5 WER, +0.8 UER, *id.*). Cependant, l'élément déterminant est l'utilisation de l'approche de pré-entraînement continue. Les modèles RAP directement affinés à partir des modèles SSL-ETRANGER+Ø+FT qui n'ont pas vu d'haïtien dans une approche CPT sont loin derrière (+8 WER, +4.6 UER pour les modèles basés sur le français, *id.*) ou très loin derrière (+18.9 WER, +9.5 UER pour les modèles basés sur l'anglais, *id.*) du meilleur modèle.

En ce qui concerne **(c)** la quantité de données nécessaires pour affiner les modèles SSL sur une tâche RAP, nos résultats montrent une différence marquée entre trois groupes de modèles : (i) SSL-ETRANGER+CPT+FT très robuste à une quantité réduite de données d'entraînement, (ii) SSL-ETRANGER+Ø+FT peu robuste à une quantité réduite de données, et (iii) SSL-NATIF+Ø montrant des résultats intermédiaires. L'utilisation de 20 minutes de données comble l'écart entre (i) et (iii) alors que les modèles du groupe (ii) ont nécessité environ 4 fois cette quantité de données (80 minutes) pour atteindre des performances similaires. Nous supposons que les modèles du groupe (i) bénéficient du fait d'avoir vu plus de parole, car ils ont été pré-entraînés dans leur langue respective (français ou anglais), ont vu des données haïtiennes dans la phase CPT, et ont été affinés, ce qui pourrait expliquer pourquoi ils sont plus robustes que les autres modèles. Enfin, nous avons observé des résultats mitigés avec l'utilisation des LM pour le décodage. Alors qu'ils n'améliorent pas de manière significative (ni ne nuisent) aux modèles SSL-NATIF+Ø+FT ou SSL-ETRANGER+CPT+FT, ils améliorent de manière significative les scores WER du SSL-ETRANGER+Ø+FT (Fig. 1a et 1b) : par exemple, −10 WER avec un LM 5-gram pour un modèle EN WAV2VEC2 affiné avec 40 minutes de données. Par conséquent, lorsqu'aucune donnée audio pour faire du pré-entraînement continu n'est disponible, l'utilisation d'un LM est indispensable. Cependant, il semble que l'utilisation des LM, tout en améliorant les scores WER, se fait au détriment de UER plus élevés (Fig. 1c et 1d) ; ce qui indique que, bien qu'il y ait plus de mots transcrits avec précision, les autres sont moins bien transcrits, ce qui se traduit par des UER plus élevés.

5 Limitations and Travaux Futurs

Dans cet article, nous nous sommes concentrés sur l'exploration de la validité de l'utilisation des données de terrain pour pré-entraîner des modèles auto-supervisés. Nous avons affiné ces modèles sur un tâche de RAP (évaluation intrinsèque), mais nous avons laissé de côté l'étude des modèles et des représentations pré-entraînés eux-mêmes (évaluation intrinsèque). Dans nos travaux futurs, nous souhaitons utiliser une tâche ABX (Schatz *et al.*, 2013) pour comparer les représentations latentes et leur transfert au niveau des phonèmes. Cela nous aiderait à mieux comprendre les performances de nos modèles. Les données que nous utilisons pour le pré-entraînement continu ont été collectées il y a 40 ans, et la langue entre cette époque et aujourd'hui a changé (p. ex. mots tombés en désuétude, évolution de la phonologie, etc.). La question de la mesure de ce phénomène et de son impact reste donc ouverte. Enfin, nos résultats montrent que 350 heures d'enregistrements sur le terrain sont suffisantes pour pré-entraîner un modèle SSL natif et obtenir des résultats compétitifs lorsqu'ils sont affinés sur une tâche de RAP. Cependant, un tel trésor avec autant d'heures d'enregistrement n'existe pas pour toutes les langues : la question de la quantité minimale de données de terrain à utiliser reste ouverte.

6 Conclusion

Nous avons utilisé des données de terrain en haïtien, enregistrées sur bandes magnétiques il y a 40 ans, puis numérisées, pour entraîner un modèle SSL natif. Nous avons également utilisé une approche CPT sur des modèles SSL pré-entraînés de la langue lexificatrice (le français) et d'une langue non apparentée (l'anglais), que nous avons affinés sur un autre ensemble de données de terrain dans le cadre d'une tâche de RAP. Nous avons obtenu des résultats compétitifs et montré que le meilleur modèle est le modèle pré-entraîné de la langue lexificatrice avec CPT sur des enregistrements de terrain haïtiens, suivi par le modèle SSL natif. Par conséquent, lorsqu'aucun modèle de la langue lexificatrice n'est disponible, il est toujours utile d'entraîner un modèle natif à l'aide de données de terrain. Ceci est d'autant plus important qu'un modèle natif peut être une source de fierté pour la communauté des locuteurs, contrairement à un modèle dérivé de la langue lexificatrice, généralement celle de l'ancienne puissance colonisatrice. Par conséquent, l'approche consistant à mobiliser les données d'archive, comme préconisée par (Bird, 2020), est une voie prometteuse.

Références

ANDRUS T., BILLS A., CONNERS T., CRABB E. S., DUBINSKI E., FISCUS J. G., GILLIES B., HARPER M., HAZEN T. J., HEFRIGHT B., JARRETT A., LE H., RAY J., RYTTING A., SHEN W., SILBER R. & TZOUKERMANN E. (2017). Iarpa babel haitian creole language pack iarpa-babel201b-v0.2b. DOI : 10.35111/ENHB-6110.

BAEVSKI A., ZHOU H., MOHAMED A. & AULI M. (2020). Wav2vec 2.0 : A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA : Curran Associates Inc.

BIRD S. (2020). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 3504–3519, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.313.

BREDIN H., YIN R., CORIA J. M., GELLY G., KORSHUNOV P., LAVECHIN M., FUSTES D., TITEUX H., BOUAZIZ W. & GILL M.-P. (2020). pyannote.audio : neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain.

BREITER W. (2014). Rapid bootstrapping of haitian creole large vocabulary continuous speech recognition.

CONNEAU A., BAEVSKI A., COLLOBERT R., MOHAMED A. & AULI M. (2021). Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, p. 2426–2430. DOI : 10.21437/Interspeech.2021-329.

FATTIER D. (1998). *Contribution à l'étude de la genèse d'un créole : l'Atlas linguistique d'Haïti, cartes et commentaires, 6 vol.* Bibliographical record, Presses Universitaires du Septentrion, Villeneuve d'Ascq. Ph.D. Dissertation, Université de Provence.

FLA, FACULTÉ DE LINGUISTIQUE APPLIQUÉE DE L'UNIVERSITÉ D'ETAT D'HAÏTI (ANCIENNEMENT CENTRE DE LINGUISTIQUE APPLIQUÉE (CLA)) & FATTIER, DOMINIQUE (2015). Atlas linguistique d'Haïti. DOI : 10.34847/COCOON.8EA988D2-BF16-303D-81A0-0C55CC035240.

GOODA SAHIB-KAUDEER N., GOBIN-RAHIMBUX B., BAHSU B. S. & MAGHOO M. F. A. (2019). Automatic speech recognition for kreol morisien : A case study for the health domain. In A. A.

SALAH, A. KARPOV & R. POTAPOVA, Éds., *Speech and Computer*, p. 414–422, Cham : Springer International Publishing.

GURURANGAN S., MARASOVIĆ A., SWAYAMDIPTA S., LO K., BELTAGY I., DOWNEY D. & SMITH N. A. (2020). Don't stop pretraining : Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8342–8360, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.740.

HAZAEL-MASSIEUX M.-C. (2012). *Les Créoles à base française*. Gap, France : Editions Ophrys.

HEAFIELD K. (2011). KenLM : Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, p. 187–197, Edinburgh, Scotland : Association for Computational Linguistics.

HIMMELMANN N. P. (2018). *Meeting the transcription challenge*. University of Hawai'i Press.

MACAIRE C., SCHWAB D., LECOUTEUX B. & SCHANG E. (2022). Automatic speech recognition and query by example for creole languages documentation. In *Findings of the Association for Computational Linguistics : ACL 2022*, p. 2512–2520, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-acl.197.

NOWAKOWSKI K., PTASZYNSKI M., MURASAKI K. & NIEUWAŻNY J. (2023). Adaptation of a multilingual speech representation model for a new, underresourced language via multilingual fine-tuning and continued pretraining. *Science Talks*, **8**, 100249. DOI : https ://doi.org/10.1016/j.sctalk.2023.100249.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5206–5210. DOI : 10.1109/ICASSP.2015.7178964.

PARCOLLET T., NGUYEN H., EVAIN S., BOITO M. Z., PUPIER A., MDHAFFAR S., LE H., ALISAMIR S., TOMASHENKO N., DINARELLI M., ZHANG S., ALLAUZEN A., COAVOUX M., ESTEVE Y., ROUVIER M., GOULIAN J., LECOUTEUX B., PORTET F., ROSSATO S., RINGEVAL F., SCHWAB D. & BESACIER L. (2023). Lebenchmark 2.0 : a standardized, replicable and enhanced framework for self-supervised representations of french speech.

SCHATZ T., PEDDINTI V., BACH F., JANSEN A., HERMANSKY H. & DUPOUX E. (2013). Evaluating speech features with the minimal-pair ABX task : analysis of the classical MFC/PLP pipeline. In *Proc. Interspeech 2013*, p. 1781–1785. DOI : 10.21437/Interspeech.2013-441.

SIMONS G. F. & FENNIG C. D., Éds. (2023). *Ethnologue : Languages of the world*. Summer Institute of Linguistics, Academic Pub.

VALDMAN A., VILLENEUVE A.-J. & SIEGEL J. F. (2015). On the influence of the standard norm of haitian creole on the cap haïtien dialect : Evidence from sociolinguistic variation in the third person singular pronoun. *Journal of Pidgin and Creole Languages*, **30**(1), 1–43. DOI : 10.1075/jpcl.30.1.01val.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Utiliser l'explicabilité des modèles pour mettre en évidence les expressions genrées dans la parole

François Buet[1] Camille Guinaudeau[2]
Cyril Grouin[1] Sahar Ghannay[1] Shin'ichi Satoh[3]
(1) Université Paris-Saclay, CNRS, LISN, 91405 Orsay, France
(2) Université Paris-Saclay, CNRS, JFLI, 101-0003 Tokyo, Japon
(3) Nationat Institue of Informatics, 101-0003 Tokyo, Japon
prenom.nom@lisn.upsaclay.fr, satoh@nii.ac.jp

RÉSUMÉ

Dans de nombreux pays, des études ont souligné la sous-représentation des femmes dans les médias. Mais au-delà du déséquilibre quantitatif se pose la question de l'asymétrie qualitative des représentations des hommes et des femmes. Comment automatiser l'évaluation des contenus et des traits saillants spécifiques aux discours masculins et féminins ? Nous proposons dans cette étude d'exploiter les connaissances acquises par un modèle de classification entraîné à la détection du genre sur des transcriptions automatiques, afin de mettre en évidence des motifs distinctifs du discours masculin ou féminin. Notre approche est basée sur l'utilisation de méthodes développées pour l'intelligence artificielle explicable (IAX), afin de calculer des scores d'attribution au niveau des unités.

ABSTRACT

Using model explainability to highlight gendered expressions in speech

In many countries, studies have highlighted the under-representation of women in the media. But beyond quantitative imbalance is the question of the qualitative asymmetry of men's and women's representations. How to automate the evaluation of content and salient features specific to male and female discourse ? We propose in this study to leverage the knowledge acquired by a classification model trained for gender detection on automatic transcripts, in order to highlight patterns distinctive of male or female speech. Our approach is based on the use of methods developed for explainable artificial intelligence (XAI), to compute token-level attribution scores.

MOTS-CLÉS : Détection du genre, explicabilité, médias.

KEYWORDS : Gender Detection, Explainability, Media.

1 Introduction

La représentation du genre dans les médias est une question de société qui a été suivie à l'échelle mondiale (MediaWatch, 1995), afin de garantir l'égalité dans la participation à la vie publique. Dans de nombreux pays, les études ont souligné la moindre présence des femmes dans les organes d'information et les médias en général (GMMP, 2021). Parallèlement à l'analyse manuelle, le développement et la diffusion des chaînes d'outils et des ressources de traitement automatique des langues (TAL) ont permis d'accroître l'ampleur des observations (Ash *et al.*, 2022). Si les outils modernes permettent de détecter le genre du locuteur avec une certaine exactitude (Doukhan *et al.*, 2018), l'éva-

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 695–707.

luation qualitative et automatique du contenu des discours masculins et féminins reste une question ouverte. Notre but dans cette étude est d'exploiter les connaissances acquises par un modèle de classification entraîné pour la détection du genre à partir d'une transcription de parole, et de l'associer à l'IA explicable (IAX), afin de mettre en évidence des motifs lexicaux distinctifs du discours masculin ou féminin (Figure 1), et ainsi d'introduire une part d'automatisation dans l'analyse de la parole masculine et féminine dans les médias.

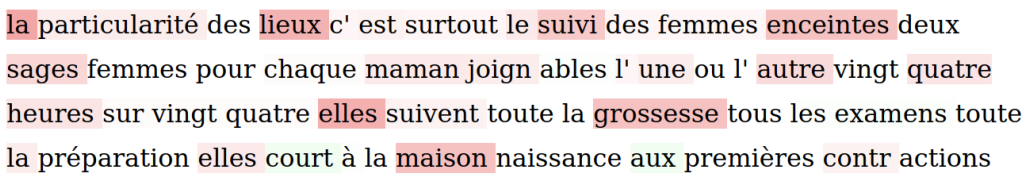


FIGURE 1 – Exemple de visualisation d'une explication issue de nos expériences. La teinte de rouge indique l'attribution à la classe féminine. L'énoncé (transcrit automatiquement) provient d'un journal radiophonique.

Les méthodes d'explicabilité sont conçues pour donner un aperçu des raisons qui sous-tendent les prédictions des modèles (Marcinkevics & Vogt, 2020). Elles peuvent s'avérer utiles soit aux utilisateurs de l'IA qui doivent s'assurer de leur confiance dans les prédictions du modèle, soit aux utilisateurs finaux qui veulent comprendre les décisions qui les concernent, soit aux développeurs et aux scientifiques des données qui doivent vérifier la robustesse de leur système. Nous proposons dans cette étude d'utiliser des techniques d'explicabilité répandues (Zeiler & Fergus, 2014 ; Ribeiro *et al.*, 2016 ; Sundararajan *et al.*, 2017) afin de calculer des explications locales pour la prédiction du genre du locuteur, sous la forme d'attributions au niveau des unités dans chaque segment de la transcription de parole. Il convient de souligner que notre problème principal n'est pas la classification du genre du locuteur (une tâche pour laquelle la modalité audio est plus adaptée), mais d'apporter une assistance pour l'analyse des discours masculins et féminins dans les médias. En masquant les informations acoustiques (par l'utilisation d'une transcription automatique pour entrée), nous cherchons à nous concentrer sur des indicateurs textuels permettant de discriminer les genres. Nos expériences se fondent sur des données provenant de programmes de télévision et de radio français et japonais. Nos principales contributions sont les suivantes : une première tentative (à notre connaissance) de détection du genre à partir de transcriptions automatiques de la parole (Section 3), et a fortiori l'application de l'IAX à cette détection (Section 4).

2 Représentation du genre dans les médias

2.1 Contexte et motivations

Les Nations unies ont reconnu l'importance de la participation et de la représentation des femmes dans les médias lors de la *Quatrième conférence mondiale sur les femmes* qui s'est tenue en septembre 1995 à Pékin (section J de la déclaration officielle) (UN, 1995). En conséquence, des initiatives telles que le *Global Media Monitoring Project* [1] (GMMP) ont été lancées pour mesurer l'état et l'évolution de la présence des femmes dans les sources d'information traditionnelles (journaux,

1. https://whomakesthenews.org/

Français	**Durée (h)**	**Exemples**	**Femmes %**
informations	14.8 / 1.5 / 1.8	5258 / 518 / 630	50 / 50 / 27
thématiques	27.0 / 2.9 / 3.5	9114 / 968 / 1168	50 / 50 / 27
téléréalité	12.1 / 0.8 / 1.0	6250 / 406 / 525	50 / 50 / 60
Japonais	**Durée (h)**	**Exemples**	**Femmes %**
train_5k	30,4	5k	50
train_100k	599	100k	50
Val / Test	9,1 / 8,4	1365 / 1184	50 / 50

TABLE 1 – Informations sur les ensembles de données. Pour la partie française, les valeurs des divisions sont rapportées dans l'ordre Train / Val / Test.

télévision, radio), ainsi que dans les médias numériques (sites web et tweets de la presse en ligne). Depuis 1995, les études successives du GMMP (une tous les cinq ans) ont révélé, entre autres, un déséquilibre entre les genres en ce qui concerne les sujets et les sources d'information (seulement 45 % de femmes en 2020), ainsi que pour les journalistes effectuant des reportages (40 % de femmes en 2020) (GMMP, 2021). De même, dans le contexte français, l'Arcom a noté dans son rapport annuel de veille 2022 (Arcom, 2023) que les femmes ne représentaient que 36 % du temps de parole global dans les émissions de télévision et de radio. Au-delà des analyses quantitatives, reste la question de la représentation [2] du genre, qui peut véhiculer et entretenir les stéréotypes et le sexisme. Ceux-ci existent, notamment, à travers la façon dont les hommes et les femmes parlent (p. ex., un style d'élocution caricatural), et à travers la préférence différenciée pour certains thèmes. Les exemples de ces phénomènes peuvent être relativement rares et subtils à détecter, ce qui demande la réalisation d'analyses fines à grande échelle.

Dans cet article, nous soutenons que l'utilisation de techniques de l'IA explicable peut aider à entreprendre une analyse qualitative fine sur des ensembles de données importants, en particulier dans le cas de la représentation des genres dans les médias audiovisuels.

2.2 Ensembles de données

Les données utilisées pour nos expériences sont des programmes de télévision et de radio diffusés en France et au Japon [3]. Cette dualité nous permet notamment de vérifier l'applicabilité de notre approche pour des langues éloignées, et nous donne une opportunité de chercher des points de comparaison entre des contextes culturels différents. Notons qu'en contrepartie de leur authenticité, ces données ne sont pas publiquement accessibles (elles ne peuvent être redistribuées que par les entreprises qui les ont produites et en détiennent les droits). Comme précédemment indiqué, nous ne traitons que les transcriptions préalablement engendrées par des systèmes de reconnaissance automatique de la parole (RAP), puisque nous nous limitons à l'étude de divergences lexicales dans le discours. La composition des ensembles de données français et japonais est résumée dans le tableau 1. Nous avons équilibré les classes de genre, sauf pour les ensembles de test en français, afin de ne pas affecter leur significativité statistique.

2. Nous entendons ici « représentation » au sens d'image renvoyée et non de présence ou de distribution.
3. Nous avons combiné plusieurs ensembles de données auxquels nous avions accès dans le cadre de contrats de projets de financement.

Données françaises Le corpus français est une combinaison de différents types d'émissions : (i) des programmes liés aux informations (des journaux télévisés locaux et nationaux, d'une durée de 20 à 30 minutes, ainsi que des matinales radio axées sur l'actualité, d'une durée de 2 à 4 heures), diffusés en 2021 par un ensemble de chaînes privées et publiques, (ii) des programmes radiophoniques thématiques (des magazines composés d'interviews, 50-60 min), diffusés en 2018 et centrés sur des thèmes tels que l'économie, le sport, la cuisine et les questions sociales, (iii) et des émissions de téléréalité (d'une durée de 45 minutes chacune) diffusées en 2021. La transcription automatique est réalisée avec une variante du système LIUM ASR décrit dans Tomashenko *et al.* (2016). L'outil InaSpeechSegmenter (Doukhan *et al.*, 2018) est utilisé pour effectuer la détection du genre et attribuer une étiquette (homme ou femme [4]) à chaque segment de parole reconnu par LIUM ASR. Ces étiquettes serviront de référence [5] lors de l'entraînement de nos classificateurs (Section 3). Certains programmes contiennent des publicités que nous avons manuellement retirées des ensembles de développement et de test.

Données japonaises Le corpus japonais comprend des transcriptions automatiques du programme de nouvelles télévisées *NHK News 7* du diffuseur public japonais NHK, retransmis entre 2001 et 2022. Ces programmes de 30 minutes ont été transcrits à l'aide de l'outil de RAP Whisper (Radford *et al.*, 2023). L'ensemble d'entraînement comprend des programmes d'information quotidiens de 2005 à 2022, tandis que les sections de test et de développement correspondent à des programmes diffusés en 2001, 2002 et 2004, respectivement. Les étiquettes de genre sont automatiquement associées à chaque énoncé avec l'outil InaSpeechSegmenter [6] (comme pour la partie française) et corrigées manuellement pour les ensembles de test et de développement, ainsi que pour 5000 exemples d'entraînement (train_5k).

3 Détection de genre fondée sur BERT

Pour effectuer la classification du genre du locuteur, nous avons opté pour une architecture fondée sur BERT (Devlin *et al.*, 2019), comme réalisé pour divers types de classification de textes (Sun *et al.*, 2019 ; González-Carvajal & Garrido-Merchán, 2020). BERT s'appuie sur le préentraînement d'un large modèle conçu pour être facilement affiné par la suite pour les tâches en aval, avec un minimum de modifications architecturales. Dans le cas de la classification de textes, cette modification consiste en une couche linéaire supplémentaire pour traiter la représentation agrégée de la séquence d'entrée (l'encodage de l'unité `[CLS]`).

3.1 Modèles

Modèles français Les modèles les plus couramment utilisés pour le français sont CamemBERT (Martin *et al.*, 2020) et FlauBERT (Le *et al.*, 2020). Comme nos données sont constituées de transcriptions automatiques de la parole, nous utilisons dans nos expériences des modèles FlauBERT-

4. L'attribution des étiquettes est basée sur le chevauchement des segments ; les énoncés qui chevauchent principalement les étiquettes de musique ou de bruit sont éliminés.

5. Doukhan *et al.* (2018) indiquent une F-mesure de détection du genre au niveau de la trame de 96,52 sur le corpus REPERE, qui contient des flux télévisés de chaînes françaises, similaires à nos données.

6. Une évaluation manuelle a posteriori sur 10 heures montre une exactitude de 94,98 % sur ces étiquettes automatiques.

Oral (Hervé *et al.*, 2022), basés sur FlauBERT, qui sont partiellement ou entièrement préentraînés sur des sorties de RAP (19 Go générés à partir d'émissions d'actualités françaises diffusées entre 2013 et 2020). Plus précisément : `FlauBERT-O-mixed` est un modèle préentraîné sur un mélange de données écrites (13 Go provenant de Wikipédia et d'articles de presse) et de transcriptions, et `FlauBERT-O-asr_nb` est un modèle préentraîné sur des données de transcriptions uniquement.

Modèles japonais Pour le japonais, nous utilisons les modèles bert-japanese [7], qui ont été préentraînés sur des articles Wikipédia (2,6 Go). Plus précisément, nous utilisons deux versions qui diffèrent par la méthode de segmentation : soit WordPiece (`bert-base-jp`), soit au niveau des caractères (`bert-base-jp-char`).

Enfin, nous utilisons le BERT multilingue original (Devlin *et al.*, 2019) (`mBERT`), qui a été préentraîné sur 100 langues correspondant aux versions de Wikipédia les plus importantes (47 Go au total), à la fois pour le français et le japonais [8].

3.2 Implémentation

Nous avons utilisé la bibliothèque Transformers de HuggingFace (Wolf *et al.*, 2020) pour mettre en uvre les modèles BERT. Nous avons suivi l'usage d'hyperparamètres largement acceptés : Adam en tant qu'optimiseur (β_1 = 0,9, β_2 = 0,999, ϵ = 1e-08), taux d'apprentissage = 2e-5, taille de batch = 32. L'entraînement est poursuivi jusqu'à ce qu'aucune amélioration de la fonction perte de validation ne soit constatée pendant 3 époques consécutives (le meilleur modèle est conservé). Les expériences ont été réalisées sur un seul GPU Tesla V100-SXM2. L'affinage d'un seul modèle prenait 3 à 6 minutes (environ 80 min en utilisant train_100k), tandis que l'inférence ne nécessitait approximativement qu'une minute. Les tailles des modèles sont les suivantes : `mBERT`, `CamemBERT`, `bert-base-jp`—110M de paramètres, `FlauBERT-O-mixed`, `FlauBERT-O-asr_nb`—138M, `bert-base-jp-char`—90M.

3.3 Résultats

Les résultats de la classification du genre du locuteur sont présentés dans le tableau 2. Comme certains ensembles de test ne sont pas équilibrés entre les classes masculine et féminine, nous utilisons P4 (Sitarz, 2023), la moyenne harmonique des scores F1 des deux classes (c.-à-d. la moyenne harmonique de la précision et du rappel mesurés pour chaque classe), comme principale mesure de classification binaire. Des intervalles de confiance ont été calculés selon l'approche de rééchantillonage *bootstrap* [9] (bootstraps = 5000, = 5, condition = instance d'émission).

En ce qui concerne la partie française, de façon générale, les modèles `FlauBERT-O` sont plus performants que `mBERT` et `CamemBERT`, ce qui montre l'avantage, dans notre cas, d'un préentraînement sur les sorties de RAP. Pour les programmes d'information, `FlauBERT-O-asr_nb` est le meilleur modèle. Il s'agit probablement d'une conséquence logique du fait que ce modèle a

7. https://github.com/cl-tohoku/bert-japanese/
8. Notons que la segmentation de `mBERT` est fondée sur WordPiece, comme pour `bert-base-jp`.
9. Ferrer et Riera,"Confidence Intervals for evaluation in machine learning." [Logiciel]. https://github.com/luferrer/ConfidenceIntervals

Modèle	P4	F1♂	F1♀	Ex.
Français (*informations*)				
mBERT	46,9 (31,1-50,9)	61,3	38,0	52,4
CamemBERT	49,9 (27,6-53,6)	56,5	44,7	51,3
FBO-mixed	51,0 (30,9-55,2)	58,4	45,2	52,7
FBO-asr_nb	55,0 (36,0-58,6)	58,9	51,6	55,6
Français (*thématiques*)				
mBERT	51,7 (30,7-63,7)	81,0	38,0	70,9
CamemBERT	55,3 (39,0-62,2)	76,2	43,4	66,4
FBO-mixed	58,8 (37,0-70,1)	82,8	45,6	73,9
FBO-asr_nb	57,9 (39,7-65,4)	78,0	46,1	68,8
Français (*téléréalité*)				
mBERT	60,5 (52,4-65,3)	55,6	66,3	61,7
CamemBERT	59,9 (52,5-65,4)	54,0	67,2	61,7
FBO-mixed	59,9 (52,9-65,3)	53,7	67,6	61,9
FBO-asr_nb	57,2 (49,9-62,3)	52,5	62,8	58,3

Modèle	P4	F1♂	F1♀	Ex.
Japonais (`train_5k`)				
mBERT	59,9 (56,9-62,7)	64,5	55,9	60,6
BB-jp	60,3 (57,5-63,0)	58,4	62,3	60,5
BB-jp-char	57,5 (54,6-60,2)	57,3	57,8	57,5
Japonais (`bert-base-jp`)				
train_5k	60,3 (57,5-63,0)	58,4	62,3	60,5
train_100k	61,8 (58,6-64,6)	70,5	54,9	64,4

TABLE 2 – Scores P4, F1 (pour les classes masculine-♂ et féminine-♀), et exactitude pour les différents modèles évalués sur l'ensemble de test. FBO : FlauBERT-O, BB : bert-base.

été entièrement préentraîné sur le même type de données. Pour les émissions thématiques, les modèles FlauBERT-O obtiennent les meilleurs résultats, alors que pour les émissions de téléréalité, FlauBERT-O-asr_nb obtient les plus mauvais scores (les autres modèles étant comparables). Encore une fois, nous pouvons supposer l'influence de la correspondance des domaines entre le préentraînement et le test (les transcriptions des émissions de télé-réalité sont assez bruitées, les gens parlant très spontanément dans ce type de programme, tandis que les informations télévisées et radiophoniques préparent une partie du discours à l'avance).

Concernant la partie japonaise, nous avons d'abord évalué des versions des modèles BERT affinées sur un ensemble de 5000 exemples vérifiés manuellement (train_5k). Les meilleurs résultats sont obtenus avec mBERT et bert-base-jp (qui utilisent tous deux la segmentation WordPiece). Nous avons ensuite comparé les performances de ce modèle en utilisant un plus grand jeu d'affinage (annoté automatiquement) : l'augmentation substantielle de la quantité de données n'entraîne qu'un léger gain.

Il apparaît que la plupart des classificateurs basés sur BERT peuvent, dans une mesure limitée, détecter le genre du locuteur sur la base d'une transcription d'énoncé (P4 $>$ 50). Ces résultats sont toutefois dans l'absolu plutôt modestes. Cela doit être mis en relation avec la difficulté intrinsèque de la tâche : les énoncés à classer ne sont composés que de 30-40 mots en moyenne dans la partie française, et seulement d'une vingtaine de mots dans la partie japonaise. Une part significative des exemples ne contient probablement pas d'indice clair. À titre de comparaison, les systèmes soumis à la tâche de profilage de genre PAN 2019 devaient prédire le genre de l'auteur à partir d'un ensemble de 100 tweets (pour l'anglais, la meilleure équipe a obtenu une exactitude de 84,17 %). Dans la section suivante, nous expliquons comment nous utilisons la confiance exprimée par le classificateur et les techniques d'explicabilité pour identifier des mots porteurs d'information sur le genre.

4 Analyse qualitative par les techniques d'explicabilité

4.1 Méthodes

Nous avons utilisé trois méthodes bien connues, représentatives de différents groupes de techniques de l'IAX, afin de calculer des valeurs de contribution aux classes masculine et féminine au niveau des mots dans les énoncés transcrits (comme illustré dans la Figure 1).

Occlusion Initialement proposée par Zeiler & Fergus (2014) dans le contexte du traitement des images, Occlusion est une approche *fondée sur les perturbations* qui, dans son principe, est l'une des techniques d'explicabilité les plus simples. Elle mesure l'effet sur la sortie du système de la suppression, ou du remplacement par une valeur neutre, d'une partie de l'entrée. Dans notre cas, il s'agit d'utiliser l'unité de masquage utilisée dans le préentraînement de BERT. Pour ce qui est de la quantification du changement causé par la perturbation, nous envisageons deux options : (a) tenir compte du changement (potentiel) de la classe prédite (ou *changement d'étiquette*), (b) tenir compte de la variation de la distribution de probabilité binaire (ou *changement de probabilité*). La première peut être considérée comme moins flexible, car toutes les perturbations n'impliquent pas un changement de la classe prédite. Pour s'assurer que l'effet de perturbation n'est pas contourné en raison de la répétition dans la séquence, nous masquons toutes les occurrences d'une unité en une seule fois. En outre, comme nous pensons que le changement d'étiquette ou de probabilité devrait être moins important dans le cas d'une séquence courte, nous pondérons chaque perturbation par le nombre d'unités uniques dans la séquence.

LIME Proposée par Ribeiro *et al.* (2016), LIME (*Local Interpretable Model-Agnostic Explanations*) est une technique *fondée sur la simplification* qui entraîne un modèle linéaire [10] de substitution à approximer, autour d'un exemple donné, la limite de décision locale du modèle complexe d'origine. Dans la classification des textes, LIME masque aléatoirement les unités de la séquence exemple. Elle ajuste alors un modèle linéaire pour faire correspondre la sortie (c.-à-d. la probabilité d'une certaine classe) du modèle complexe pour ces variantes masquées. Ce modèle linéaire prenant en entrée une représentation simplifiée, sous la forme d'un vecteur binaire indiquant la présence ou l'absence de chaque unité dans l'échantillon perturbé. En conséquence, les coefficients appris fournissent des valeurs au niveau des unités pour la contribution à la classe cible.

LIG Proposée par Mudrakarta *et al.* (2018), LIG (*Layer Integrated Gradients*) est une approche *fondée sur les gradients*, directement inspirée des gradients intégrés de Sundararajan *et al.* (2017). Intuitivement, on peut considérer que si le produit d'un modèle change considérablement en fonction de la variation d'une dimension d'entrée (c.-à-d. que le gradient de la sortie par rapport à la dimension d'entrée est élevé en valeur absolue), cela signifie que la valeur d'entrée de cette dimension particulière est importante pour la décision du modèle. Cependant, Sundararajan *et al.* (2017) remarquent qu'au lieu de la variation locale autour de la valeur d'entrée, il faudrait considérer la variation agrégée entre la valeur d'entrée et une valeur de base non informative représentant l'absence de la caractéristique (ce qui renvoie à l'idée de mesurer l'effet d'une perturbation). Mudrakarta *et al.*

10. Ce cas est le plus courant pour la mise en uvre de LIME, mais la description générale donnée par Ribeiro *et al.*, 2016 permet d'utiliser d'autres types de modèles interprétables (p. ex., des arbres de décision).

(2018) ont appliqué ce principe au traitement des textes, en définissant le neutre comme une séquence d'unités de remplissage, et en intégrant le gradient de la sortie du modèle sur les dimensions de la couche de plongement (qui correspond à un espace continu, par opposition aux unités de la séquence).

4.2 Implémentation

La mise en uvre des méthodes LIME et LIG a été réalisée à l'aide de la bibliothèque Captum (Kokhlikyan *et al.*, 2020). Pour LIME, nous calculons la mesure de proximité (notée π_x dans Ribeiro *et al.*, 2016) par le biais de la distance cosinus entre les encodages de `CLS` au sein des échantillons d'origine et perturbé, en utilisant le modèle *lasso linéaire* de scikit-learn (α = 0.001), et en échantillonnant 200 perturbations par exemple [11]. Les expériences ont été réalisées sur un seul GPU Tesla V100-SXM2, et sur un processeur Intel Cascade Lake 6248 (10 curs à 2,5 GHz). Les temps d'exécution des méthodes d'IAX, lors de l'analyse des prédictions de `FlauBERT-O-mixed` pour les émissions de téléréalité françaises de l'ensemble de test (525 exemples), sont les suivants : Occlusion—1 min, LIME—1 h, LIG—30 min.

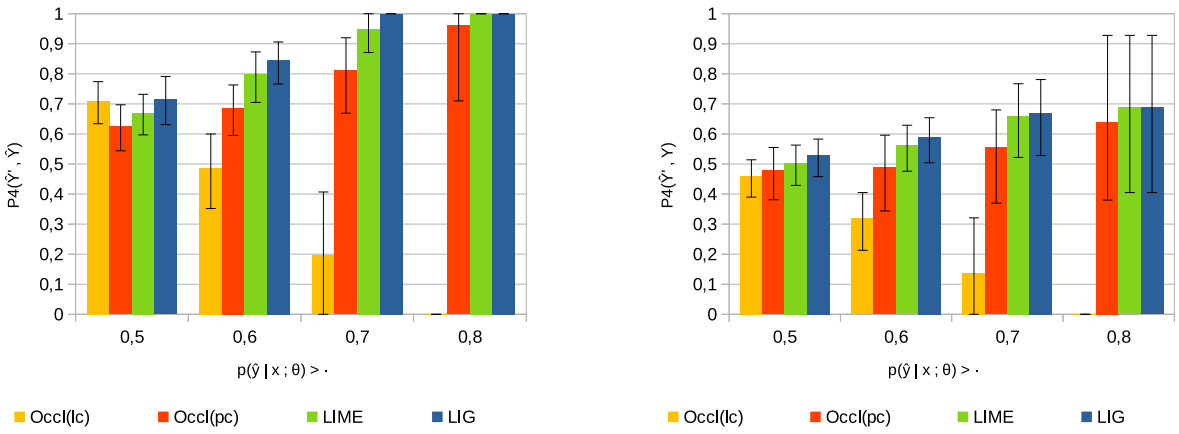
4.3 Identification des expressions genrées

Nous avons appliqué les trois méthodes d'explicabilité afin de produire des explications pour les prédictions de nos modèles sur les ensembles de tests. Nous avons choisi de calculer les scores d'attribution au niveau des unités en fonction de leur contribution à la classe cible masculine (les scores positifs indiquent une orientation masculine et les scores négatifs une orientation féminine). Afin de vérifier la **cohérence** des attributions avec, d'une part, les prédictions effectives du modèle, et d'autre part, les étiquettes de références, nous définissons une procédure de calcul de « pseudo-prédictions ». Plus précisément, soient θ un modèle de classification, ψ une méthode d'explication, x une séquence, et $(\psi(x, \theta, ♂))_t$ la valeur de contribution à la classe homme (♂) attribuée par ψ à la t-ième unité de x. Alors la pseudo-prédiction dérivée de l'explication $\psi(x, \theta, ♂)$ est : $\hat{y}' =$ ♂ si $\sum_t (\psi(x, \theta, ♂))_t > 0$ sinon ♀

Du calcul des pseudo-prédictions $\hat{Y}'$ nous inférons $\mathrm{P4}(\hat{Y}', \hat{Y})$, une évaluation vis-à-vis des prédictions effectives du modèle, ainsi que $\mathrm{P4}(\hat{Y}', Y)$, une évaluation vis-à-vis des références. La figure 2a compare les scores $\mathrm{P4}(\hat{Y}', \hat{Y})$ obtenus en fonction des techniques d'explicabilité employées, sur différents sous-ensembles d'exemples de test filtrés selon des seuils croissants de confiance du classificateur ($p(\hat{y}|x; \theta) > \cdot$). Nous pouvons notamment voir que les pseudo-prédictions sont d'autant plus proches des vraies prédictions quand la confiance du modèle est élevée (p. ex. pour LIG, à partir de $p(\hat{y}|x; \theta) > 0{,}7$, correspondant à 117 instances, $\hat{Y}' = \hat{Y}$). Nous observons également que LIG est la technique affichant la plus grande cohérence entre ses attributions et les prédictions. Notons que pour Occlusion(changement d'étiquette) $\mathrm{P4}(\hat{Y}', \hat{Y})$ diminue quand le niveau de filtrage augmente : de façon logique, puisque que la croissance de la confiance s'oppose naturellement au changement d'étiquette prédite pour l'échantillon perturbé. La figure 2b compare les scores $\mathrm{P4}(\hat{Y}', Y)$: les mêmes tendances se retrouvent dans ce cas, à ceci près que les scores sont globalement plus faibles (P4 ne dépasse pas $0{,}7$, même avec le seuil le plus élevé).

Afin de fournir une vue d'ensemble, nous avons calculé la moyenne des scores associés à chaque

11. Notre code est disponible à cette adresse : https://github.com/Cyossnarf/XSpeakerGender



(a) (b)

FIGURE 2 – Évaluation de la cohérence des explications. Le modèle analysé est `FlauBERT-O-mixed`, appliqué sur les émissions françaises de téléréalité (les exemples sont filtrés selon la confiance du classificateur). Les intervalles de confiance ont été calculés selon l'approche *bootstrap* (bootstraps = 5000, = 5, condition = instance de programme).

occurrence d'unité pour produire un lexique genré agrégé. La figure 3 présente des exemples de lexiques correspondant aux scores calculés les plus élevés (♂) et les plus bas (♀) parmi les vocabulaires [12]. Pour les données japonaises, certains thèmes sont davantage associés aux hommes (politique) et aux femmes (météo), reflétant une tendance existant effectivement dans les programmes. Pour la partie française, nous avons pu noter un usage plus marqué des pronoms personnels par les femmes dans les émissions de téléréalité. Cela coïncide avec les observations d'études antérieures (Pennebaker, 2011 ; Kocher & Savoy, 2016).

	Français (*téléréalité*), `FlauBERT-O-mixed`, LIME
♂	`soirée`, `famille`, `temps`, `soir`, `aime`, `amour`, `vie`, `faut`, `demain`, `chez`
♀	`lui`, `clairement`, `cas`, `euh`, `cela`, `tu`, `son`, `avoir`, `moins`, `avait`
	Japonais (*informations*, `train_100k`), LIG
♂	市場 (marché), 国会 (régime), まし (meilleur), 選挙 (élection), 政府 (gouvernement), 議員 (membre du parlement), 党 (parti), 側 (côté), ます (masu), です (est), ね (hé)
♀	雪 (neige), 朝 (matin), 北海道 (Hokkaido), 晴れ (ensoleillé), 東北 (Tohoku), 雨 (pluie), 夜 (nuit), そう (oui), にかけて (dessus), 日 (jour)

FIGURE 3 – Lexique pour les classes masculine-♂ et féminine-♀, extrait par le biais des techniques d'explicabilité (automatiquement en français dans le cas du japonais).

12. Nous n'avons conservé que les unités qui apparaissaient 9 fois ou plus, afin de calculer des valeurs moyennes fiables.

5 Travaux connexes

Ces dernières années, l'IAX a été appliquée à une variété de sous-domaines afin de fournir une assistance dans la prise de décision par un agent : par exemple pour la détection de fausses nouvelles (Yang *et al.*, 2019), l'intervention d'un instructeur dans un MOOC (Alrajhi *et al.*, 2022), ou encore la détection de discours haineux (Kim *et al.*, 2022). La détection du genre fait partie du domaine plus vaste du profilage des auteurs, qui vise à déduire des traits sociaux ou de personnalité sur la base des messages produits par une personne (Stajner & Yenikent, 2020). Sánchez *et al.* (2022) mentionnent trois contextes en particulier dans lesquels elle peut être utilisée : la linguistique judiciaire—afin d'identifier les auteurs de violence et de harcèlement sur internet, le marketing—afin de mener des stratégies publicitaires personnalisées, et la sociolinguistique—afin de mettre en relation des motifs linguistiques avec des variables sociales comme le genre. Les algorithmes utilisés pour le profilage de genre à partir de textes vont des approches classiques, telles que les SVM et la classification naïve bayésienne (Burger *et al.*, 2011), aux approches modernes basées sur les neurones (Bartle & Zheng, 2015). Depuis 2013, l'association PAN [13], dans le cadre du *Conference and Labs of the Evaluation Forum* (CLEF), a organisé une série de tâches partagées de profilage des auteurs, parmi lesquelles la détection du genre a été abordée à plusieurs reprises. Enfin, plusieurs études ont lié la détection du genre avec les analyses stylistiques (Savoy, 2022). Par exemple, en analysant des tweets, Ikae & Savoy (2022) constatent que certains termes ou catégories de mots—tels que les articles, les pronoms personnels, les négations, les émotions, les nombres, la ponctuation, les émojis—peuvent être davantage liés à un genre qu'à l'autre.

6 Conclusion

Cette étude présente notre méthodologie de détection du genre à partir de transcriptions automatiques de la parole. Nous avons expliqué comment utiliser les techniques développées pour l'explicabilité des modèles de façon à mettre en évidence des informations lexicales genrées, en prenant comme cas d'application des programmes télévisés et radiophoniques en français et en japonais. Nos expériences montrent que les classificateurs basés sur BERT peuvent, dans une mesure limitée, prédire le genre du locuteur sur la base d'un énoncé. Par conséquent, l'IAX devrait être associée à l'exploitation de la confiance de la prédiction du classificateur afin de potentiellement localiser des indicateurs spécifiques au genre. Les orientations de recherche future pourraient inclure : l'utilisation des méthodes d'IAX afin de produire des attributions plus complexes, et la récupération des annotations sur le genre de l'interlocuteur pour analyser les différences selon ce critère.

Remerciements

Ce travail a été financé par l'ANR (projet *Gender Equality Monitor* – ANR-19-CE38-0012). Il a en outre bénéficié dun accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2023-AD011014209 attribuée par GENCI.

13. https://pan.webis.de/

Références

ALRAJHI L., PEREIRA F. D., CRISTEA A. I. & ALJOHANI T. (2022). A good classifier is not enough : A XAI approach for urgent instructor-intervention models in moocs. In M. M. T. RODRIGO, N. MATSUDA, A. I. CRISTEA & V. DIMITROVA, Éds., *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium - 23rd International Conference, AIED 2022, Durham, UK, July 27-31, 2022, Proceedings, Part II*, volume 13356 de *Lecture Notes in Computer Science*, p. 424–427 : Springer. DOI : 10.1007/978-3-031-11647-6_84.

ARCOM (2023). Women representation on television and radio. Available online : https://www.arcom.fr/sites/default/files/2023-06/Representation_des_femmes_a_la_television_et_a_%20la_radio-Rapport_sur_exercice_2022-Arcom.pdf. In French. Last accessed : 16/10/2023.

ASH E., DURANTE R., GREBENSHCHIKOVA M. & SCHWARZ C. (2022). *Visual Representation and Stereotypes in News Media*. CESifo Working Paper Series 9686, CESifo.

BARTLE A. & ZHENG J. (2015). Gender classification with deep learning. *Stanfordcs, 224d Course Project Report*, p. 1–7.

BURGER J. D., HENDERSON J., KIM G. & ZARRELLA G. (2011). Discriminating gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, p. 1301–1309, Edinburgh, Scotland, UK. : Association for Computational Linguistics.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DOUKHAN D., CARRIVE J., VALLET F., LARCHER A. & MEIGNIER S. (2018). AN OPEN-SOURCE SPEAKER GENDER DETECTION FRAMEWORK FOR MONITORING GENDER EQUALITY. In *IEEE International Conference on Acoustic Speech and Signal Processing*, Calgary, Canada. HAL : hal-01927560.

GMMP (2021). 6th global media monitoring project highlight of findings. Available online : https://whomakesthenews.org/wp-content/uploads/2021/08/GMMP-2020.Highlights_FINAL.pdf. Last accessed : 16/10/2023.

GONZÁLEZ-CARVAJAL S. & GARRIDO-MERCHÁN E. C. (2020). Comparing BERT against traditional machine learning text classification. *CoRR*, **abs/2005.13012**.

HERVÉ N., PELLOIN V., FAVRE B., DARY F., LAURENT A., MEIGNIER S. & BESACIER L. (2022). Using ASR-Generated Text for Spoken Language Modeling. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, p. 17–25, virtual+Dublin, France : Association for Computational Linguistics. DOI : 10.18653/v1/2022.bigscience-1.2, HAL : hal-03770460.

IKAE C. & SAVOY J. (2022). Gender identification on twitter. *J. Assoc. Inf. Sci. Technol.*, **73**(1), 58–69. DOI : 10.1002/asi.24541.

KIM J., LEE B. & SOHN K. (2022). Why is it hate speech ? masked rationale prediction for explainable hate speech detection. In N. CALZOLARI, C. HUANG, H. KIM, J. PUSTEJOVSKY, L. WANNER, K. CHOI, P. RYU, H. CHEN, L. DONATELLI, H. JI, S. KUROHASHI, P. PAGGIO, N. XUE, S. KIM, Y. HAHM, Z. HE, T. K. LEE, E. SANTUS, F. BOND & S. NA, Éds., *Proceedings*

of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022, p. 6644–6655 : International Committee on Computational Linguistics.

KOCHER M. & SAVOY J. (2016). Unine at CLEF 2016 : Author profiling. In K. BALOG, L. CAPPELLATO, N. FERRO & C. MACDONALD, Éds., *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016*, volume 1609 de *CEUR Workshop Proceedings*, p. 903–911 : CEUR-WS.org.

KOKHLIKYAN N., MIGLANI V., MARTIN M., WANG E., ALSALLAKH B., REYNOLDS J., MELNIKOV A., KLIUSHKINA N., ARAYA C., YAN S. & REBLITZ-RICHARDSON O. (2020). Captum : A unified and generic model interpretability library for pytorch. *CoRR*, **abs/2009.07896**.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBE B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised Language Model Pre-training for French. In *LREC*, Marseille, France. HAL : hal-02890258.

MARCINKEVICS R. & VOGT J. E. (2020). Interpretability and explainability : A machine learning zoo mini-tour. *CoRR*, **abs/2012.01805**.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., VILLEMONTE DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a Tasty French Language Model. In *ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics*, Seattle / Virtual, United States. DOI : 10.18653/v1/2020.acl-main.645, HAL : hal-02889805.

MEDIAWATCH (1995). Global media monitoring project : Women's participation in the news.

MUDRAKARTA P. K., TALY A., SUNDARARAJAN M. & DHAMDHERE K. (2018). Did the model understand the question ? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1896–1906, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1176.

PENNEBAKER J. W. (2011). Your use of pronouns reveals your personality. *Harvard business review*, **89**(12), 32–33.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*.

RIBEIRO M. T., SINGH S. & GUESTRIN C. (2016). "why should I trust you ?" : Explaining the predictions of any classifier. In B. KRISHNAPURAM, M. SHAH, A. J. SMOLA, C. C. AGGARWAL, D. SHEN & R. RASTOGI, Éds., *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, p. 1135–1144 : ACM. DOI : 10.1145/2939672.2939778.

SÁNCHEZ D. M., MORENO A. & JIMÉNEZ-LÓPEZ M. D. (2022). Machine learning methods for automatic gender detection. *Int. J. Artif. Intell. Tools*, **31**(3), 2241002 :1–2241002 :8. DOI : 10.1142/S0218213022410020.

SAVOY J. (2022). Stylometric analysis of characters in Shakespeares plays. *Digital Scholarship in the Humanities*, **38**(3), 1238–1246. DOI : 10.1093/llc/fqac092.

SITARZ M. (2023). Extending F1 metric, probabilistic approach. *Adv. Artif. Intell. Mach. Learn.*, **3**(2), 1025–1038. DOI : 10.54364/aaiml.2023.1161.

STAJNER S. & YENIKENT S. (2020). A survey of automatic personality detection from texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 6284–6295, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.553.

SUN C., QIU X., XU Y. & HUANG X. (2019). How to fine-tune BERT for text classification ? In M. SUN, X. HUANG, H. JI, Z. LIU & Y. LIU, Éds., *Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings*, volume 11856 de *Lecture Notes in Computer Science*, p. 194–206 : Springer. DOI : 10.1007/978-3-030-32381-3_16.

SUNDARARAJAN M., TALY A. & YAN Q. (2017). Axiomatic attribution for deep networks. In D. PRECUP & Y. W. TEH, Éds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 de *Proceedings of Machine Learning Research*, p. 3319–3328 : PMLR.

TOMASHENKO N., VYTHELINGUM K., ROUSSEAU A. & ESTÈVE Y. (2016). LIUM ASR systems for the 2016 Multi-Genre Broadcast Arabic Challenge. In *IEEE Workshop on Spoken Language Technology*, San Diego, CA, USA, United States. DOI : 10.1109/SLT.2016.7846278, HAL : hal-01433188.

UN U. N. (1995). Beijing declaration and platform for action. Available online : https://www.un.org/womenwatch/daw/beijing/pdf/BDPfA%20E.pdf. Last accessed : 16/10/2023.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., SCAO T. L., GUGGER S., DRAME M., LHOEST Q. & RUSH A. M. (2020). Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics.

YANG F., PENTYALA S. K., MOHSENI S., DU M., YUAN H., LINDER R., RAGAN E. D., JI S. & HU X. B. (2019). Xfake : Explainable fake news detector with visualizations. In *The World Wide Web Conference*, WWW '19, p. 3600‑3604, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3308558.3314119.

ZEILER M. D. & FERGUS R. (2014). Visualizing and understanding convolutional networks. In D. FLEET, T. PAJDLA, B. SCHIELE & T. TUYTELAARS, Éds., *Computer Vision – ECCV 2014*, p. 818–833, Cham : Springer International Publishing.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Vers une pédagogie inclusive : une classification multimodale des illustrations de manuels scolaires pour des environnements d'apprentissage adaptés

Saumya Yadav[1], Élise Lincker[2], Caroline Huron[3], Stéphanie Martin[3], Camille Guinaudeau[4, 5], Shin'ichi Satoh[5] and Jainendra Shukla[1]
(1) HMI Lab, IIIT-Delhi, India
(2) Cedric, CNAM, Paris, France
(3) Le Cartable Fantastique, Paris, France
(4) Japanese French Laboratory for Informatics, CNRS, Japan
(5) National Institute of Informatics, Tokyo, Japan
saumya@iiitd.ac.in

RÉSUMÉ

Afin de favoriser une éducation inclusive, des systèmes automatiques capables d'adapter les manuels scolaires pour les rendre accessibles aux enfants en situation de handicap sont nécessaires. Dans ce contexte, nous proposons de classifier les images associées aux exercices selon trois classes (*Essentielle*, *Informative* et *Inutile*) afin de décider de leur intégration ou non dans la version accessible du manuel pour les enfants malvoyants. Sur un ensemble de données composé de 652 paires (texte, image), nous utilisons des approches monomodales et multimodales à l'état de l'art et montrons que les approches fondées sur le texte obtiennent les meilleurs résultats. Le modèle CamemBERT atteint ainsi une exactitude de 85,25 % lorsqu'il est combiné avec des stratégies de gestion de données déséquilibrées. Pour mieux comprendre la relation entre le texte et l'image dans les exercices des manuels, nous effectuons également une analyse qualitative des résultats obtenus avec et sans la modalité image et utilisons la méthode LIME pour expliquer la décision de nos modèles.

ABSTRACT

Towards Inclusive Pedagogy : A multimodal Classification of Textbook Illustrations for Adaptive Learning Environments

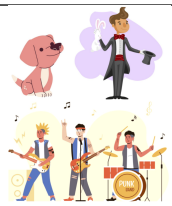
To foster inclusive education, automatic systems that can adapt textbooks to make them accessible to children with disabilities are necessary. In this context, we propose a task to classify the images associated with the exercises according to three classes (*Essential*, *Informative*, and *Useless*) to decide whether to integrate them into the accessible version of the textbook for visually impaired children. On a dataset composed of 652 (text, image) pairs, we use state-of-the-art monomodal and multimodal approaches and show that text-based approaches achieve better results. The CamemBERT model achieves an accuracy of 85.25% when combined with unbalanced data management strategies. To better understand the relationship between text and image in textbooks' exercises, we also perform a qualitative analysis of the results obtained with and without the image modality and use the LIME method to explain the decision of our models.

MOTS-CLÉS : Classification multimodale · Éducation inclusive · Explicabilité des modèles.

KEYWORDS: Multimodal Classification · Inclusive Education · Model explainability.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 708–719.

TABLE 1 – Exemples de classification de paires (texte, image)

Classe	Essentielle	Informative	Inutile
Images			
Texte	Écris le son commun aux 3 objets représentés par les dessins.	Texte : À la préhistoire, les hommes dessinaient des peintures rupestres sur les murs de leur caverne - Q : Trouve le verbe. À quel temps est il conjugué ?	Recopie les phrases si tu reconnais le verbe “aller”. (a) Je quitte la maison à la même heure tous les matins. (b) Samedi, je me suis baladé dans le parc.

1 Introduction

Le droit à l’éducation est universel, transcendant les limitations imposées par des handicaps physiques ou cognitifs. Cependant, les ressources éducatives conventionnelles, en particulier les manuels scolaires, ne sont pas intrinsèquement conçues pour répondre aux besoins divers des apprenants, surtout ceux en situation de handicap. Cette disparité dans l’accessibilité des matériaux éducatifs entrave considérablement le processus d’apprentissage pour les enfants ayant des besoins spéciaux, amplifiant ainsi l’écart éducatif.

L’évolution des techniques d’intelligence artificielle a ouvert de nouvelles voies pour l’apprentissage adapté, pourtant l’intégration de ces technologies dans le soutien des enfants en situation de handicap reste peu explorée. Spécifiquement, les apprenants malvoyants rencontrent des barrières découlant de la nature visuelle des matériaux éducatifs standards, qui sont principalement basés sur le texte et l’image. Cela limite leur capacité à accéder à l’information et affecte leur engagement et leur motivation. Des associations commencent à produire des manuels numériques adaptés aux enfants en situation de handicaps, en effectuant les transformations à la main. Malheureusement, étant donné la grande diversité des collections et le renouvellement des programmes d’enseignement, ces adaptations artisanales ne permettent pas de répondre aux besoins. Dans ce contexte, l’utilisation d’approches automatiques est indispensable pour rendre accessible les matériaux éducatifs au plus grand nombre.

L’automatisation de l’adaptation de manuels scolaires a été peu étudiée. Lincker *et al.* (2023b) a été pionnier dans la classification des exercices des manuels en fonction de leurs objectifs d’apprentissage, facilitant l’adaptation automatique des manuels pour les enfants ayant des problèmes de coordination motrice (dyspraxie). Cette adaptation atténue le besoin d’écriture manuelle et rationalise les interactions avec les manuels tout en préservant l’intégrité éducative. Cependant, les auteurs se concentrent sur la mise en page et le contenu textuel des manuels, ce qui ne correspond pas aux besoins des élèves malvoyants. Afin de combler cette lacune, notre travail propose donc un nouveau cadre de classification des images accompagnant les exercices des manuels afin de déterminer leur caractère facultatif ou non. La classification de ces images dans trois classes (*Essentielle*, *Informative* et *Inutile*), voir exemples dans le Tableau 1, est cruciale pour l’adaptation des manuels, guidant les décisions sur l’inclusion d’une image dans l’interface utilisateur (avec du texte alternatif généré) ou son omission pour un document plus clair et plus accessible, adapté à la consommation auditive via un assistant vocal. Pour cela, nous annotons un ensemble de données de plus de 600 paires (texte, image) avec ces 3 catégories et utilisons des algorithmes de classification monomodaux et multimodaux à l’état de l’art. Notre motivation pour combiner les deux modalités, image et texte repose sur l’intuition qu’une

image *Inutile* présente un chevauchement sémantique significatif avec le texte (l'image ne fournit pas d'informations supplémentaires), tandis qu'une image *Essentielle* sera sémantiquement très différente du texte de l'exercice qu'elle illustre. Nous comparons également cette approche multimodale avec des méthodes monomodales pour analyser l'impact des différentes modalités indépendamment. Les principales contributions de ce travail sont triples : (1) une comparaison des approches multi et monomodales pour la classification (texte, image) ; (2) une comparaison qualitative des résultats pour mieux comprendre l'impact de chaque modalité ; (3) une analyse des fonctionnalités utilisées par le modèle à travers la méthode d'explicabilité LIME (Ribeiro *et al.*, 2016).

2 État de l'art

Le travail présenté dans cet article est lié à différents domaines : le Traitement Automatique des Langues appliqué aux manuels scolaires et la similarité texte-image.

La recherche appliquée aux manuels scolaires est relativement rare dans le domaine du traitement automatique des langues. La plupart des études existantes se concentrent soit sur l'analyse du contenu linguistique des manuels ((Green, 2019; Lucy *et al.*, 2020)), sur la génération de questions à partir de ceux-ci ((Ch & Saha, 2022; Gerald *et al.*, 2022)) ou la création de ressources lexicales qui pourraient être utilisées pour la classification et la représentation des manuels (Manulex (Lété *et al.*, 2004), ReSyf (Billami *et al.*, 2018) ou Alector (Gala *et al.*, 2020)). Semblable à notre objectif d'adapter les manuels pour les enfants en situation de handicap, des études récentes se sont concentrées sur la modélisation et l'extraction de contenu à partir de manuels (Lincker *et al.*, 2023b) ou la classification d'exercices basée sur leur objectif éducatif (Lincker *et al.*, 2023a). Cependant, ces travaux se concentrent uniquement sur l'analyse de la mise en page et du texte et non sur les images présentes dans le manuel.

L'analyse de la similarité ou de la nature de la relation entre un texte et une image associée est souvent fondée sur des *transformers* vision-langage pré-entraînés qui se basent typiquement sur des ensembles de données de légendes d'images tels que MS COCO (Lin *et al.*, 2014) ou Flickr30k (Young *et al.*, 2014) pour évaluer leurs modèles (par exemple, (Rao *et al.*, 2022) sur les tâches de recherche d'information ou (Huang *et al.*, 2019) sur les tâches de légendage d'images). Ces ensembles de données comprennent des images complexes représentant de multiples objets dans des arrière-plans riches. Malgré la richesse des domaines visuels dans ces ensembles de données, leurs légendes tendent à être des descriptions en une seule phrase, alors que dans notre ensemble de données, les relations entre le texte et l'image sont plus variées, où l'image et le texte peuvent être soit redondants, soit complémentaires. L'analyse comparative des modalités image et texte a été révolutionnée par l'introduction du modèle Contrastive Language–Image Pre-training (CLIP) (Radford *et al.*, 2021) qui apprend les concepts visuels à partir des descriptions textuelles, facilitant une association plus nuancée entre le texte et les images que les modèles traditionnels. Ce modèle offre la capacité d'estimer la similarité entre un texte et une image, largement utilisée dans la recherche d'informations cross-modale ou les système de questions-réponses visuels (*Visual Question Answering* (VQA)). Plus étroitement liés à notre travail, deux articles récents analysent la relation entre le texte et l'image dans le contexte de la recherche d'information image-texte (Qu *et al.*, 2021) et de la classification (Otto *et al.*, 2020). Otto *et al.* (2020) présentent un cadre de classification analysant les relations sémantiques entre les images et les descriptions textuelles. Ils définissent huit classes, s'appuyant sur trois concepts : l'Information Mutuelle Cross-Modale, la Corrélation Sémantique, et le Statut (qui décrivent la relation hiérarchique entre le texte et l'image). L'étude implique la création automatique d'ensemble de

données à partir de MSCOCO, VIST (Malakan *et al.*, 2023) et ImageNET (Deng *et al.*, 2009) et se base sur deux classifieurs d'apprentissage profond, un classique et un en cascade, pour évaluer la difficulté de la tâche. Bien que ces ensembles de données soient publiquement disponibles, ils diffèrent significativement de notre objectif. Les parties textuelles associées aux images dans ces ensembles de données consistant principalement en une légende d'image d'une seule phrase (MSCOCO) ou une étiquette d'un seul mot (ImageNet). Le texte dans nos exercices, destiné à poser une question, peut en effet être constitué de plusieurs phrases et servir un but légèrement différent.

3 Données

Les images des activités et leçons de manuels scolaires jouent différents rôles, et reconnaître l'importance de ces éléments visuels est crucial pour l'adaptation automatique des manuels. Dans ce travail, les exercices avec images ont été extraits de trois manuels scolaires français pour le primaire. Pour ce faire, chaque manuel au format PDF est converti en fichier XML au format ALTO en utilisant les outils pdfalto [1] et MuPDF [2]. Cette approche permet l'extraction des mots dans une représentation structurée et organisée du contenu tout en fournissant des informations sur la mise en page et le style de police. Suivant la méthode employée par Lincker *et al.* (2023b), les mots extraits sont ensuite regroupés en segments de texte, qui à leur tour sont regroupés en blocs d'activités en fonction de la mise en page, du style de police et des caractéristiques d'espacement. Les images sont associées aux blocs selon leur position sur la page. Finalement, deux experts ont manuellement annoté les images avec leur texte respectif en trois classes différentes :

- `Images Essentielles` : Ces images sont indispensables pour comprendre ou résoudre une activité.
- `Images Informatives` : Elles contribuent à la compréhension du texte et fournissent des informations supplémentaires sans être essentielles pour résoudre l'exercice (ajout d'indices pour résoudre l'exercice ou explication sur un concept inconnu des élèves).
- `Images Inutiles` : Elles peuvent être exclues lors de l'adaptation pour les enfants en situation de handicap afin de simplifier l'interface adaptée.

Pour simplifier au maximum l'interface adaptée pour les enfants malvoyants, il est essentiel d'exclure les images de la classe *Inutile* des manuels adaptés. Un exemple de la classification des images avec leur texte respectif est présenté dans le Tableau 1. Dans la classe *Essentielle*, l'image est obligatoire pour résoudre l'exercice, tandis que le but de l'image de la classe *Informative* est de donner des informations supplémentaires à l'élève, qui peut ne pas savoir de qu'est exactement une peinture rupestre. Enfin, dans le dernier exemple, l'image associée au texte « Copie les phrases si tu... » n'est pas utile pour résoudre l'exercice et n'a qu'un but décoratif.

Trois manuels scolaires français ont été utilisés dans notre étude ; deux provenant du même éditeur et un troisième venant d'un éditeur différent. Pour maintenir une évaluation rigoureuse, nous avons combiné les manuels du même éditeur et les avons partitionnés en ensembles d'entraînement et de validation en utilisant un ratio de 80/20. L'ensemble de test est, quant à lui, constitué à partir des paires (texte, image) extraites du manuel d'un éditeur différent afin de garantir que les modèles sont évalués sur des données non vues lors de l'entraînement. Le Tableau 2 présente le nombre de paires (texte, image) ainsi que le nombre de mots correspondants pour chaque classe des ensembles d'entraînement / validation et de test. Le Tableau 3 illustre, quand à lui, les 5 mots les plus fréquents dans chacune des 3 classes annotées, révélant des motifs distincts dans les occurrences de mots à

1. https://github.com/kermitt2/pdfalto
2. https://github.com/ArtifexSoftware/mupdf

TABLE 2 – Nombre de paires (texte, image) et nombre de mots différents dans chaque classe

	Essentielle		Informative		Inutile	
	# mots	# paires	# mots	# paires	# mots	# paires
Entraînement+Validation	16.0	258	21.5	96	11.2	58
Test	13.5	131	32.4	75	11.2	42

TABLE 3 – Mots les plus fréquents dans chaque classe

Test			Entraînement+Validation		
Essentielle	Informative	Inutile	Essentielle	Informative	Inutile
nom : 29	mot : 77	texte : 14	écris : 160	mot : 93	verbe : 21
dessin : 27	texte : 46	verbe : 11	dessin : 103	texte : 47	texte : 20
écris : 24	verbe : 41	phrase : 10	mot : 101	verbe : 42	phrase : 16
trouver : 23	observe : 34	recopie : 8	nom : 89	observe : 24	mot : 15
donne : 17	phrase : 25	combine : 7	utilise : 75	écris : 23	écris : 15

travers les catégories. En effet, nous pouvons remarquer que les mots « nom », « dessin » et « écris » sont assez caractéristiques de la classe *Essentielle*, alors que les classe *Informative* et *Inutile* tendent à partager davantage les mots qui apparaissent les plus fréquemment (« texte », « verbe » ou « phrase »).

4 Approches

Pour la classification des illustrations, nous avons utilisé différentes modalités. Pour cela, nous avons utilisé des approches multimodales, qui considèrent le texte et l'image simultanément ainsi que des méthodes monomodales qui se fondent uniquement sur le texte ou sur l'image.

4.1 Approche multi-modale

CLIP (Radford *et al.*, 2021) est une architecture permettant d'apprendre des représentations visuelles à partir d'une faible supervision textuelle. Nos données textuelles étant extraites de livres en français, nous avons tout d'abord traduit les textes en anglais en utilisant la bibliothèque de traduction hors ligne open source Argos Translate [3], qui repose sur OpenNMT (Klein *et al.*, 2017), pour répondre à l'exigence d'entrée de texte en anglais de CLIP. Nous avons utilisé la variante RN101 de CLIP car elle donnait de meilleurs résultats sur notre ensemble de données de validation.

Dans cette première approche multi-modale, nous avons calculé la similarité entre les images et les textes à partir du modèle CLIP afin d'analyser la relation entre l'image et le texte, à partir de l'équation suivante :

$$CLIPScore(I,T) = max(100 * cos(E_I, E_T), 0) \quad (1)$$

où E_I et E_T correspondent aux plongements de l'image et du texte, respectivement. Nos données textuelles pouvant dépasser la longueur par défaut de 77 *tokens* définie par CLIP, nous avons utilisé deux stratégies différentes pour gérer cette limitation : la troncature et la segmentation. Dans le

3. https://github.com/argosopentech/argos-translate

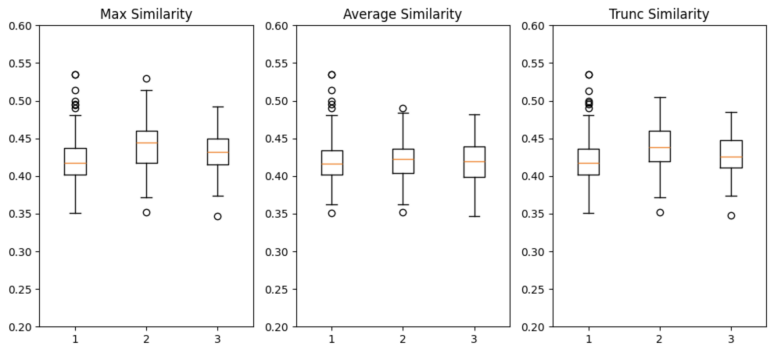


FIGURE 1 – Similarité texte-image calculée à partir du modèle CLIP pour chaque classe (1 : *Essentielle*, 2 : *Informative*, 3 : *Inutile*)

premier cas, les textes sont tronqués à la longueur par défaut alors qu'ils sont divisés en plusieurs segments dans le second cas (la similarité globale entre le texte et l'image correspond alors à la moyenne ou à la similarité maximale de tous les segments textuels). La Figure 1 présente les valeurs de similarité obtenues avec ces deux stratégies. Nous pouvons constater que, contrairement à notre intuition initiale, il n'existe que peu de différences de similarité entre texte et image en fonction des 3 classes. Par conséquent, nous définissons une deuxième approche multimodale, consistant à combiner les plongements des images obtenues grâce à la même variante RN101 du modèle CLIP et les plongements des textes, calculés grâce aux modèles de langue présentés dans la section suivante, pour les fournir à un Perceptron Multi-Couches (*Multi-Layers Perceptron* - MLP).

4.2 Approches mono-modales

Fondées sur le texte Pour l'encodage des plongements basés sur le texte dans l'approche monomodale, nous avons utilisé le modèle de langue BERT (Kenton & Toutanova, 2019), appliqué sur les données traduites en anglais, pour une comparaison plus directe avec les résultats fournis par CLIP, et le modèle de langue CamemBERT (Martin *et al.*, 2020) appliqué directement sur les exercices en français. Nous avons utilisé la même approche de classification que celle utilisée dans le cadre multi-modal. Les textes des exercices ont d'abord été tokenisés avant d'être fournis aux modèles `bert-base-uncased`[4] pour l'anglais et `camembert-base`[4] pour le français, permettant l'extraction de plongements contextualisées à partir de la dernière couche cachée de l'architecture BERT et CamemBERT. Un pooling moyen adaptatif a été appliqué à travers la dimension de la longueur de la séquence afin d'obtenir une représentation de taille fixe, encapsulant les informations essentielles du texte d'entrée. Les plongements textuels obtenus à travers ces modèles sont ensuite soumis à un MLP pour la classification.

Finalement, pour améliorer davantage notre représentation sémantique des exercices, nous utilisons le modèle de langage CamemBERT affiné sur des données éducatives : leçons et activités de quatre manuels (deux manuels de la collection utilisée pour l'entraînement, excluant les exercices utilisés pour construire notre jeu de données, et deux autres manuels non vus), 1293 Fantastiques Exercices fournis par l'association Le Cartable Fantastique[5], et les 79 textes de lecture originaux d'Alector. Ce modèle de langue *CamemBERT-education*, proposé dans Lincker *et al.* (2023a), est utilisé de façon similaire à CamemBERT pour calculer les plongements des textes des exercices.

4. L'utilisation de modèles plus large sur notre ensemble de validation a fourni de moins bons résultats.
5. https://www.cartablefantastique.fr/

Fondées sur l'image Pour l'extraction des plongements des images des exercices, nous avons utilisé les modèles ResNET (He *et al.*, 2016), VGG16 (Simonyan & Zisserman, 2015) et Inception-v3 (Szegedy *et al.*, 2016), qui sont connus pour leur efficacité sur les tâches de reconnaissance d'image. L'étape initiale consiste à charger un modèle `ResNet-50` (ou `VGG16` ou `Inception-v3`) pré-entraîné puis de personnaliser la dernière couche du modèle afin de produire des plongements de taille 512, identiques à ceux du modèle CLIP. Toutes les couches, à l'exception de la dernière couche modifiée, sont gelées pour préserver les connaissances encodées dans les couches antérieures. Par ailleurs, les images ont été préalablement redimensionnées pour garantir la compatibilité avec les attentes d'entrée du modèle ResNET.

Finalement, suivant la même méthode que pour le modèle CLIP ou la classification monomodale basée sur le texte, les plongements extraits sont fournis à un MLP pour notre tâche de classification.

4.3 Gestion des données déséquilibrées

Les données annotées dans le cadre de ce travail sont fortement déséquilibrées, comme le montre le Tableau 2. Le déséquilibre des classes pouvant affecter la performance du modèle, en favorisant particulièrement la classe majoritaire, nous avons utilisé deux stratégies pour gérer ce problème de déséquilibre, conjointement ou séparément :

— `Stratégie de pondération des classes` **:** Cette stratégie consiste à attribuer différents poids aux classes afin d'accorder plus d'importance à la classe minoritaire. Les pondérations sont calculées à l'aide de la fonction `compute_class_weight` de la bibliothèque Python `scikit-learn`, avec la stratégie *balanced* qui ajuste dynamiquement le poids des classes en fonction de leur distribution dans les données d'entraînement, fournissant des poids plus élevés aux classes sous-représentées.

— `Génération de données` **:** Dans cette stratégie, l'ensemble d'entraînement initial est complété par 150 instances de la classe *Inutile*. Ces nouvelles instances correspondent à des exercices sans image extraits de nos manuels d'entraînement auxquels nous avons associé des images aléatoires provenant de la classe *Inutile* au sein des mêmes manuels. Par la suite, nous avons fusionné ces données augmentées avec les données d'entraînement originales et suivi la même procédure d'extraction de plongements et de passage à travers le MLP que nous l'avions fait pour les modèles précédents.

5 Expérimentations

5.1 Configuration

Le modèle CLIP propose plusieurs variantes, et nous avons sélectionné RN101 pour sa performance supérieure sur la partie validation de notre jeu de données, avec un plongement de dimension 512. Nous avons ensuite extrait les plongements des données monomodales, en maintenant la même dimension et la même procédure qu'avec le modèle CLIP. Nous avons utilisé diverses techniques pour fusionner les modalités, incluant l'extraction du maximum et du minimum de deux plongements et leur addition. Les résultats optimaux, sur notre ensemble de validation, ont été obtenus en concaténant les deux plongements. Ainsi, nous les avons concaténé en un vecteur de taille 1024 que nous avons fourni en entrée du MLP. L'architecture MLP utilisée se compose d'une couche d'entrée, de deux couches cachées avec activation ReLU, et d'une couche de sortie pour les tâches de classification.

TABLE 4 – Résultat de classification avec les modèles mono et multi-modaux

Modèles	Modalité	Exactitude
Classe majoritaire	-	0.7267
BERT	texte	0.8156
CamemBERT	texte	0.8361
CamemBERT-educational	texte	0.80
ResNET	image	0.5246
Inception-v3	image	0.5041
VGG16	image	0.50
BERT+ResNET	texte+image	0.7582
CamemBERT+ResNET	texte+image	0.8033
CLIP	texte+image	0.8074

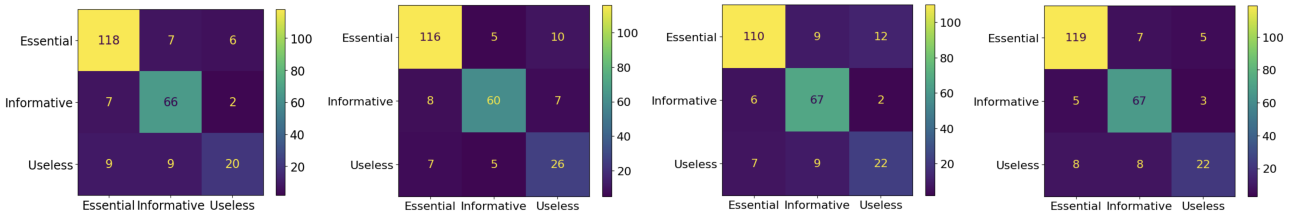
TABLE 5 – Résultats de classification avec CamemBERT et les stratégies de gestion de données déséquilibrées.

Class Weight	Data Augmentation	Exactitude
✗	✗	0.8361
✓	✗	0.8156
✗	✓	0.8279
✓	✓	**0.8525**

5.2 Résultats

Le Tableau 4 présente les résultats obtenus sur le jeu de données de test, pour les approches de classification mono-modales et multi-modales. Nous pouvons constater que les modèles basés sur le texte surpassent leurs homologues basés sur l'image. En effet, la classification basée sur l'image donne des résultats inférieurs à la classe majoritaire pour les trois modèles (ResNET, VGG16 et Inception-v3) suggérant que l'image seule n'est pas suffisante pour décider si une image est nécessaire, informative ou inutile dans le contexte d'un exercice. Pour la classification basée sur le texte, la meilleure performance est obtenue avec le modèle de langue CamemBERT sans affinage des données éducatives. CamemBERT-education étant supposé avoir une meilleure représentation sémantique des données éducatives en français (il offre de meilleurs résultats sur une tâche de classification des exercices en fonction de leurs objectifs pédagogiques (Lincker *et al.*, 2023a)), sa faible performance sur notre tâche de classification suggère que les plongements sémantiques de la partie textuelle des exercices ne constituent pas un facteur décisif dans le processus de classification des images.

Par ailleurs, le Tableau 4 montre que les approches multimodales n'améliorent pas les performances par rapport à la classification basée sur le texte seulement. Dans ce cas, les meilleurs résultats sont obtenus grâce au modèle CLIP, légèrement au-dessus de ceux obtenus avec la concaténation des plongements dérivés du texte et de l'image, en utilisant respectivement CamemBERT et ResNet. Ces faibles résultats s'expliquent par le fait que, comme montré précédemment sur la Figure 1, les valeurs de similarités entre le texte et l'image calculées avec le modèle CLIP ont des valeurs presque similaires pour les trois classes, contrairement à notre intuition qui était que les images inutiles avaient une similarité sémantique plus élevée avec le texte (redondance) tandis que les images nécessaires avaient une similarité sémantique plus faible (complémentarité).



(a) (b) (c) (d)

FIGURE 2 – Matrices de confusion pour le modèle CamemBERT seul (a), avec une augmentation de données (b), avec une pondération des classes (c) et avec une combinaison des deux stratégies (d) (axe des x = classe de référence, axe des y = classe prédite)

TABLE 6 – Comparaison des résultats obtenus avec les approches mono- et multi-modales. ✓(resp. ✗) correspond à la classification correcte (resp. incorrecte) de la paire (texte, image)

	Exercice 1	Exercice 2	Exercice 3
Texte	Résouds ce rébus.	Choisis les bons adjectifs pour décrire la princesse.	Quel type d'art est-ce ?
Image			
Monomodal (texte)	✗	✓	✓
Monomodal (image)	✓	✓	✗
Multimodal	✗	✗	✗

Finalement, le Tableau 5 présente les résultats obtenus avec les différentes stratégies utilisées pour traiter notre problème de déséquilibre des données. Les meilleurs résultats sont obtenus lorsque les stratégies de pondération des classes et d'augmentation des données sont utilisées conjointement, atteignant une exactitude de 85,25%. D'un point de vue qualitatif (cf. Figure 2), l'augmentation des données tend à classer plus d'exemples dans la classe *Inutile*, à la fois correctement et incorrectement, (2b), tandis que la stratégie de pondération tend à améliorer le nombre d'instances correctement classées pour les classes sous-représentées (*Inutile* et *Informative*) aux dépends de la classe *Essentielle*, (2c). Enfin, combiner les deux stratégies améliore globalement le nombre d'instances correctement classifiées pour toutes les classes, (2d).

5.3 Analyse des résultats

Le Tableau 6 présente la comparaison des résultats obtenus avec des modèles mono-modaux (basés sur le texte ou sur l'image) ou multimodaux sur 3 exercices de la classe *Essentielle*. Sur ces exercices, le modèle basé sur le texte étiquette incorrectement l'exercice 1, ayant pour objectif l'analyse d'image (lecture d'un rébus), tandis que le modèle basé sur l'image fournit de bons résultats pour les exercices 1 et 2, c'est-à-dire dans des cas d'analyse et de description d'images. Cependant, la modalité visuelle ne permet pas de classifier correctement l'exercice 3, que ce soit dans une approche mono- ou multi-modale, alors que la modalité textuelle (sous la forme d'une question ouverte) est suffisante pour la classification. Finalement, l'approche multimodale fournit de mauvais résultats pour ces 3 exemples, mettant en évidence les défis dans l'intégration efficace des modalités texte et image.

Afin d'améliorer l'interprétabilité de nos différents modèles, nous utilisons la méthode LIME (pour Explications Interprétables Locales Agnostiques au Modèle - *Local Interpretable Model-agnostic*

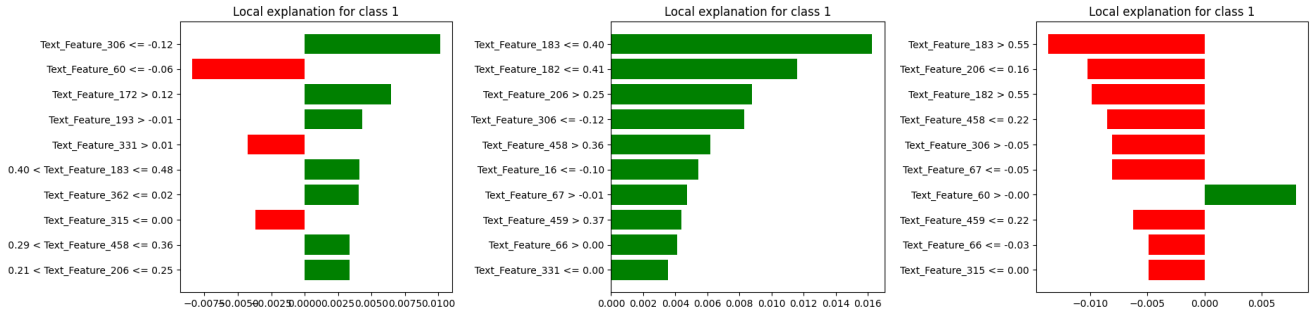


FIGURE 3 – Explications fournies par LIME pour 3 instances aléatoires des classes (a) Inutile (b) Informative et (c) Essentielle

Explanations) (Ribeiro *et al.*, 2016) qui fournit des explications localisées pour des prédictions individuelles et améliore la transparence. La Figure 3 présente le résultat de LIME pour trois instances aléatoires de trois classes différentes, où l'axe des y représente les 10 meilleures caractéristiques extraites, c'est-à-dire, les plongement produits par CamemBERT. Pour la classe *Inutile*, la prédiction semble être influencée par un équilibre délicat de contributions positives et négatives de diverses caractéristiques, suggérant que la frontière de décision pour la classe *Inutile* est nuancée, sans aucune caractéristique dominante orientant la prédiction. Au contraire, les classes *Informative* et *Essentielle* présentent des contributions de caractéristiques positives plus prononcées indiquant des influences plus fortes sur la décision du modèle. Par ailleurs, certaines caractéristiques, telles que les caractéristiques 183 ou 206, présentent des impacts variables à travers les classes indiquant leur pertinence contextuelle dans la différenciation entre les classes *Essentielle*, *Informative* et *Inutile*.

6 Conclusion et perspectives

Notre étude aborde le besoin impératif d'une éducation inclusive en proposant un système automatique de classification des images associées aux exercices de manuels scolaires en 3 classes (*Essentielle*, *Informative* et *Inutile*). Nous avons utilisé un ensemble de données composé de 652 paires (texte, image) et avons exploré les approches monomodales et multimodales pour la classification. Étonnamment, les méthodes monomodales basées sur le texte ont surpassé leurs homologues multimodales. Notre analyse des résultats montre que l'aspect sémantique n'est probalement pas le plus important pour la classification des images, et que des éléments surfaciques du texte de l'exercice jouent une part importante dans la classification. Finalement, nous avons utilisé la méthode d'explication LIME pour obtenir des aperçus du processus de prise de décision de nos modèles et ainsi montrer que les classes *Informative* et *Essentielle* étaient fortement caractérisées contrairement à la classe *Inutile*. Malgré ces résultats prometteurs, notre étude présente certaines limitations, la principale étant la quantité relativement faible de nos données, qui ne sont par ailleurs pas partageables avec la communauté, pour des raisons de propriété intellectuelle, ce qui entrave la reproductibilité de nos expériences. Cette limitation souligne la nécessité de disposer de jeux de données plus larges et disponibles publiquement. À l'avenir, nous prévoyons d'élargir notre ensemble de données en (1) extrayant plus de manuels scolaires (2) incorporant des données de Otto *et al.* (2020) afin d'améliorer la représentation de la classe *Informative*.

Remerciements Ce travail a été financé par le *NII International Internship Program* et le projet MALIN (MAnuels scoLaires INclusifs / ANR-21-CE38-0014).

Références

BILLAMI M. B., FRANÇOIS T. & GALA N. (2018). ReSyf : a French lexicon with ranked synonyms. In *International Conference on Computational Linguistics*.

CH D. R. & SAHA S. K. (2022). Generation of multiple-choice questions from textbook contents of school-level subjects. *IEEE Transactions on Learning Technologies*.

DENG J., DONG W., SOCHER R. *et al.* (2009). Imagenet : A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*.

GALA N., TACK A., JAVOUREY-DREVET L. *et al.* (2020). Alector : A parallel corpus of simplified french texts with alignments of misreadings by poor and dyslexic readers. In *12th Language Resources and Evaluation for Language Technologies*.

GERALD T., ETTAYEB S., LE H. Q., VILNAT A., PAROUBEK P. & ILLOUZ G. (2022). An annotated corpus for abstractive question generation and extractive answer for education. In *Conférence sur le Traitement Automatique des Langues Naturelles*.

GREEN C. (2019). A multilevel description of textbook linguistic complexity across disciplines : Leveraging NLP to support disciplinary literacy. *Linguistics and Education*.

HE K., ZHANG X., REN S. & SUN J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*.

HUANG L., WANG W., CHEN J. & WEI X.-Y. (2019). Attention on attention for image captioning. In *IEEE/CVF international conference on computer vision*.

KENTON J. D. M.-W. C. & TOUTANOVA L. K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, p. 4171–4186.

KLEIN G., KIM Y., DENG Y. *et al.* (2017). OpenNMT : Open-source toolkit for neural machine translation. In *Association for Computational Linguistics - System Demonstrations*.

LÉTÉ B., SPRENGER-CHAROLLES L. & COLÉ P. (2004). MANULEX : A grade-level lexical database from french elementary school readers. *Behavior Research Methods, Instruments, & Computers*.

LIN T.-Y., MAIRE M., BELONGIE S. *et al.* (2014). Microsoft coco : Common objects in context. In *European Conference in Computer Vision*.

LINCKER É., GUINAUDEAU C., PONS O. *et al.* (2023a). Noisy and unbalanced multimodal document classification : Textbook exercises as a use case. In *20th International Conference on Content-based Multimedia Indexing*.

LINCKER E., PONS O., GUINAUDEAU C., BARBET I., DUPIRE J., HUDELOT C., MOUSSEAU V. & HURON C. (2023b). Layout-and activity-based textbook modeling for automatic pdf textbook extraction. In *Intelligent Textbooks 2023*.

LUCY L., DEMSZKY D., BROMLEY P. & JURAFSKY D. (2020). Content analysis of textbooks via natural language processing : Findings on gender, race, and ethnicity in texas US history textbooks. *AERA Open*.

MALAKAN Z. M., ANWAR S., HASSAN G. M. & MIAN A. (2023). Sequential vision to language as story : A storytelling dataset and benchmarking. *IEEE Access*.

MARTIN L., MULLER B., SUÁREZ P. J. O. *et al.* (2020). Camembert : a tasty french language model. In *Annual Meeting of the Association for Computational Linguistics*.

OTTO C., SPRINGSTEIN M., ANAND A. & EWERTH R. (2020). Characterization and classification of semantic image-text relations. *International Journal of Multimedia Information Retrieval*.

QU L., LIU M., WU J., GAO Z. & NIE L. (2021). Dynamic modality interaction modeling for image-text retrieval. In *44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

RADFORD A., KIM J. W., HALLACY C. *et al.* (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*.

RAO J., WANG F., DING L. *et al.* (2022). Where does the performance improvement come from ? -a reproducibility concern about image-text retrieval. In *ACM SIGIR Conference on Research and Development in Information Retrieval*.

RIBEIRO M. T., SINGH S. & GUESTRIN C. (2016). "why should I trust you ?" : Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining*.

SIMONYAN K. & ZISSERMAN A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)* : Computational and Biological Learning Society.

SZEGEDY C., VANHOUCKE V., IOFFE S., SHLENS J. & WOJNA Z. (2016). Rethinking the inception architecture for computer vision. In *IEEE conference on computer vision and pattern recognition*.

YOUNG P., LAI A., HODOSH M. & HOCKENMAIER J. (2014). From image descriptions to visual denotations : New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

astroECR : enrichissement d'un corpus astrophysique en entités nommées, coréférences et relations sémantiques

Atilla Kaan Alkan[1,2], Felix Grezes[3], Cyril Grouin[1],
Fabian Schüssler[2], Pierre Zweigenbaum[1]
(1) Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique, 91405, Orsay, France.
(2) IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France.
(3) Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
{atilla.alkan, cyril.grouin, pz}@lisn.upsaclay.fr,
fabian.schussler@cea.fr, felix.grezes@cfa.harvard.edu

RÉSUMÉ

Le manque de ressources annotées constitue un défi majeur pour le traitement automatique de la langue en astrophysique. Afin de combler cette lacune, nous présentons astroECR, une extension du corpus TDAC (Time-Domain Astrophysics Corpus). Notre corpus, constitué de 300 rapports d'observation en anglais, étend le schéma d'annotation initial de TDAC en introduisant cinq classes d'entités nommées supplémentaires spécifiques à l'astrophysique. Nous avons enrichi les annotations en incluant les coréférences, les relations sémantiques entre les objets célestes et leurs propriétés physiques, ainsi qu'en normalisant les noms d'objets célestes via des bases de données astronomiques. L'utilité de notre corpus est démontrée en fournissant des scores de référence à travers quatre tâches : la reconnaissance d'entités nommées, la résolution de coréférences, la détection de relations, et la normalisation des noms d'objets célestes. Nous mettons à disposition le corpus ainsi que son guide d'annotation, les codes sources, et les modèles associés.

ABSTRACT

astroECR : an Enriched Corpus for Astrophysical Entities, Coreferences, and Relations

The lack of annotated resources poses a significant challenge for natural language processing in astrophysics. To address this gap, we introduce astroECR, an extension of the TDAC (Time-Domain Astrophysics Corpus). This corpus, comprised of 300 observation reports in English, expands the initial annotation scheme of TDAC by introducing five additional named entity classes specific to astrophysics. We enhanced annotations to include coreferences, semantic relations between celestial objects and their physical properties, and normalization of celestial object names using astronomical databases. We demonstrate our corpus's utility by providing baseline scores across four tasks : named entity recognition, coreference resolution, relation detection, and normalization of celestial object names. We provide the corpus, annotation guide, source code, and associated models to the community.

MOTS-CLÉS : Annotation de corpus, Extraction d'information, Astrophysique.

KEYWORDS: Corpus Annotation, Information Extraction, Astrophysics.

31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position, pages 720–733.

1 Introduction

Ces dernières années, le besoin de développer des systèmes d'analyse et d'extraction d'information en astrophysique a engendré une multiplication des travaux en Traitement Automatique des Langues (TAL). Les récents modèles de langue tels que astroBERT (Grezes *et al.*, 2021) et astroLLaMa (Nguyen *et al.*, 2023) sont utilisés non seulement pour identifier dans la littérature des entités spécifiques au domaine (Grezes *et al.*, 2022), mais également pour l'extraction d'information essentielles telles que les coordonnées célestes ou les propriétés physiques mesurées lors de l'observation d'objets célestes (Sotnikov & Chaikova, 2023). Néanmoins, un défi notable persiste dans la disponibilité des ressources et la diversité des annotations. En effet, la plupart des corpus existants (Becker *et al.*, 2005; Hachey *et al.*, 2005; Murphy *et al.*, 2006) ne sont pas accessibles et servent uniquement à la détection d'entités nommées. Parmi les corpus disponibles, on compte celui de la campagne d'évaluation DEAL (Grezes *et al.*, 2022) et un second plus restreint, TDAC (Alkan *et al.*, 2022), axé sur l'observation des phénomènes transitoires tels que les explosions de supernovae et les sursauts gamma. Les deux corpus partagent les mêmes classes d'entités et se limitent à une annotation en entités nommées, laissant ainsi un manque dans la diversité des annotations. Or, une extraction d'information plus complète nécessite des corpus avec des annotations comprenant les coréférences, pour identifier toutes les mentions se référant à une même entité, ou encore les relations entre les différentes paires d'entités. Par exemple, comme illustré dans la figure 1, lorsqu'il y a mention de plusieurs objets célestes dans le texte, se limiter à annoter uniquement les entités nommées ne permet pas d'établir de manière précise les liens entre les différentes propriétés physiques et les objets célestes correspondants.

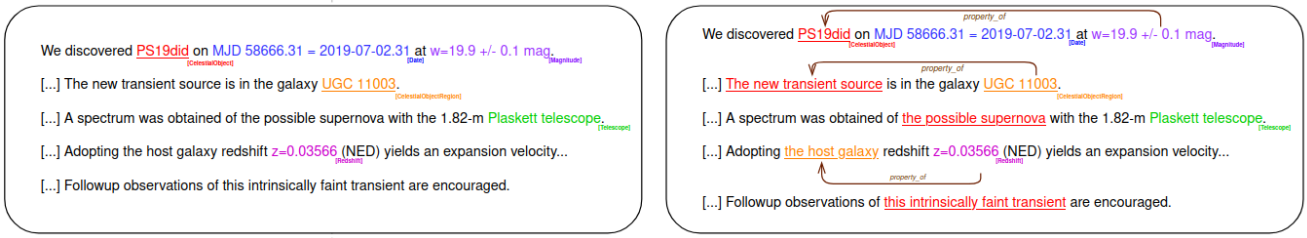


FIGURE 1 – Extrait d'un rapport d'observation. A gauche, un exemple d'annotation en entités nommées uniquement, et à droite, l'annotation des entités nommées avec en plus l'annotation des mentions de coréférences et des relations sémantiques entre les objets célestes (mentions de type CelestialObject) et leurs propriétés physiques.

Afin de combler cette lacune et faciliter un spectre plus large de recherches, notre travail vise à créer un corpus annoté englobant l'annotation d'entités nommées, de mentions de coréférences, de relations astrophysiques entre les corps célestes et leurs propriétés physiques, ainsi qu'en fournissant une normalisation des noms d'objets célestes. Nous avons pour objectif de fournir à la communauté astrophysique et TAL une ressource permettant le développement de modèles d'extraction d'information. Pour ce faire, nous avons étendu la première version de notre corpus existant TDAC, afin de construire astroECR, un nouveau corpus en astrophysique plus riche en annotations.

Les principales contributions de ce travail sont les suivantes :

— Nous avons augmenté la taille du corpus TDAC, en passant de 75 à 300 documents annotés. Cette augmentation comprend un ensemble plus complet d'annotations couvrant davantage d'entités nommées astrophysiques, la normalisation des noms d'objets célestes (liage référentiel), des annotations de coréférence et des relations astrophysiques. À notre connaissance, il s'agit de la première et unique ressource de ce genre dans le domaine ;

— Nous avons repris les catégories d'entités nommées du corpus TDAC et y avons ajouté cinq catégories d'entités nommées supplémentaires ;
— Nous démontrons l'utilité de ce corpus en réalisant des expériences sur quatre tâches d'extraction d'information, pour lesquelles nous avons développé des modèles et fourni de premiers scores. En perspective, ces modèles faciliteront l'annotation automatisée de documents supplémentaires ;
— Nous mettons notre corpus, notre guide d'annotation, le code associé et les modèles à la disposition de la communauté de recherche via notre dépôt GitHub [1].

2 Travaux connexes

Ressources pour la détection d'entités nommées La reconnaissance d'entités nommées implique l'identification de mentions d'entités, telles que des personnes, lieux ou organisations (Grishman & Sundheim, 1996). Il s'agit d'une tâche utile en recherche d'information (Yadav & Bethard, 2018; Banerjee *et al.*, 2019) et également pour les systèmes de question-réponse (Mollá Aliod *et al.*, 2006). En astrophysique, Becker *et al.* (2005); Hachey *et al.* (2005) ont créé l'Astronomy Bootstrapping Corpus (ABC), composé de 209 résumés d'articles scientifiques radioastronomiques en anglais pour la détection d'entités nommées spécifiques au domaine telles que les noms d'instruments astronomiques, les objets célestes, leurs types, et leurs caractéristiques spectrales. Cependant, ce corpus n'est pas accessible. Murphy *et al.* (2006) ont également élaboré un corpus de 7840 phrases issues d'articles scientifiques en anglais, définissant 43 types d'entités nommées, incluant des catégories caractérisant les objets célestes : leurs coordonnées et propriétés physiques (fréquence, luminosité). À notre connaissance, ce corpus n'est pas accessible non plus. Plus récemment, le corpus de la campagne d'évaluation DEAL (Grezes *et al.*, 2022) a été rendu public, devenant l'un des premiers corpus accessibles en astrophysique [2]. Il comprend des extraits de texte intégral et des sections de remerciements provenant d'articles d'astrophysique en anglais, annotés spécifiquement pour la campagne, avec 31 catégories d'entités nommées. Il est divisé en trois sous-ensembles : entraînement (1753 documents), développement (1366 documents) et test (2505 documents). Dans un précédent article, nous avions introduit TDAC (Alkan *et al.*, 2022), le seul corpus annoté en entités nommées construit à partir de rapports d'observation astronomique en anglais, se concentrant sur un vocabulaire spécifique à l'astronomie (l'étude des phénomènes transitoires). Accessible [3], il se compose de 75 documents, dont 25 circulaires du réseau GCN (Barthelmy *et al.*, 1995) de la NASA, 25 télégrammes astronomiques (Rutledge, 1998) et 25 AstroNotes issus du Transient Name Server (Gal-Yam, 2021).

Ressources pour la résolution des coréférences La résolution de coréférences est une tâche visant à identifier toutes les mentions dans un texte se référant à une même entité (Jurafsky & Martin, 2023; Zheng *et al.*, 2011). Comparée à la détection d'entités nommées, la tâche de résolution de coréférences dans les documents en astrophysique a reçu moins d'attention. Kim & Webber (2006) se sont penchés sur la résolution d'anaphores dans la littérature astrophysique. Les auteurs se sont uniquement concentrés sur la classification automatique du pronom "*they*" dans les articles, en distinguant ceux qui se réfèrent à des recherches citées et ceux qui ne le font pas. Leur système repose sur un classificateur d'entropie maximale avec des caractéristiques basées sur la distance

1. https://github.com/AtillaKaanAlkan/astroECR
2. https://huggingface.co/datasets/adsabs/WIESP2022-NER/
3. https://github.com/AtillaKaanAlkan/TDAC

entre les citations précédentes et les types de verbes associés au pronom en question. Brack *et al.* (2021) ont construit le corpus STM pour la résolution de coréférences. Le corpus se compose de dix disciplines scientifiques (dont onze résumés annotés en astrophysique). Les auteurs ont comparés plusieurs approches existantes dans la littérature, avec notamment l'utilisation de modèles de type BERT pour la résolution des coréférences (Joshi *et al.*, 2019), mais également en s'inspirant de la méthode proposée par Luan *et al.* (2018) via l'utilisation de représentations de mots ELMo (Peters *et al.*, 2018).

Synthèse Les corpus annotés pour le domaine astrophysique sont limités. Ces ressources se concentrent principalement sur des corpus orientés reconnaissance d'entités nommées, limitant leur usage pour des tâches plus larges en TAL. De plus, les documents concernés sont principalement des articles, restreignant la variété des sources de données que les chercheurs peuvent exploiter. Pour combler cette lacune, nous avons basé notre travail sur le corpus existant TDAC, pour construire un corpus plus riche et l'étendre à des tâches de TAL non traitées telles que la résolution de coréférences, la détection de relations astrophysiques et la normalisation des noms d'objets célestes.

3 Annotation du corpus

Dans cette section, nous décrivons le processus d'annotation des entités nommées (3.1), la normalisation des noms d'objets célestes (3.2), les coréférences (3.3) et les annotations des relations astrophysiques (3.4). Nous avons utilisé BRAT (Stenetorp *et al.*, 2012) comme outil d'annotation.

3.1 Extension des classes d'entités nommées et annotation

Nous avons adopté le guide d'annotation de TDAC en proposant une extension du schéma d'annotation avec cinq catégories d'entités supplémentaires jugées essentielles par les astronomes.

- **`Date :`** Dates et expressions temporelles se référant à une date de détection ou à la durée d'une observation. Exemple : *We report the discovery of a probable nova in M31 on a co-added 990-s R-band CCD frame taken under poor conditions on* ***2019 Mar. 12.791 UT*** $_{[Date]}$ *with the 0.65-m telescope at Ondrejov.*
- **`Reference :`** Références vers d'autres rapports d'observation, utiles pour répérer et regrouper tous les rapports concernant un même objet céleste. Exemple : *In comparison to the optical region (ref : the SALT spectrum in* ***ATel #3289*** $_{[Reference]}$*), few strong NI lines are expected in the JHK bands.*
- **`Magnitude :`** Equations et valeurs numériques qui caractérisent la luminosité des corps célestes (propriété utile pour les astronomes afin de déterminer la visibilité des objets). Exemple : *As reported to CBAT, this nearby-M31 object was discovered by Koichi Itagaki at* ***16.5 mag.*** $_{[Magnitude]}$
- **`Flux :`** Valeur numérique caractérisant l'énergie d'un corps céleste. Exemple : *The flux values ranged from* ***1.01 +/-0.06 E+11 cgs*** $_{[Flux]}$ *to* ***1.71 +/- 0.04 E+11 cgs.*** $_{[Flux]}$
- **`Redshift :`** Equations et valeurs numériques caractérisant la distance d'un corps céleste par rapport à un observateur. Exemple : *The host KUG 0180+227 is an E+A galaxy at* ***z=0.022.*** $_{[Redshift]}$

3.2 Normalisation des mentions de type CelestialObject

La normalisation consiste à désambiguïser les mentions d'entités en les reliant à leur entrée respective dans des bases de connaissances (Sevgili *et al.*, 2020). Nous normalisons les noms d'objets célestes du corpus, tels que les supernova, les sursauts gamma, et les galaxies, à leurs entrées spécifiques dans les catalogues astronomiques. En égard aux différentes conventions de dénomination des objets célestes en astronomie, cette normalisation est essentielle pour l'intégration de données provenant de divers articles et rapports d'observation. Par exemple, la galaxie d'Andromède [4] a au moins 39 désignations, chacune devant être correctement associée. Nous utilisons pour cela trois catalogues astronomiques complémentaires : SIMBAD (Wenger *et al.*, 2000), NED (Mazzarella *et al.*, 2001), et TNS (Gal-Yam, 2021).

3.3 Périmètre d'annotation des coréférences

Notre guide d'annotation détaillé est accessible via notre dépôt GitHub. Ici, nous donnons un aperçu général de nos choix d'annotation des coréférences. Nous avons également annoté les cas où un objet céleste est désigné par un autre de ses noms dans le texte. Nous avons exclu les expressions mathématiques, les quantités numériques et d'autres relations coréférentielles non associées à un objet céleste de notre schéma d'annotation. Pour clarifier cette distinction, examinons les exemples suivants :

- Coréférences annotées (une même couleur marque les éléments d'une chaîne de coréférence) :
  - *We discovered **PS19did** on MJD 58666.31 = 2019-07-02.31, at w=19.9 +/- 0.1 [...] **The new transient source** is in the galaxy **UGC 11003** [...] Adopting **the host galaxy** redshift z=0.03566 (NED) yields an expansion velocity [...] Followup observations of **this intrinsically faint transient** are encouraged.*
  - *We report on the discovery and follow-up of a very bright and highly magnified microlensing event **Gaia19bld**. [...] **It** has been detected and announced by the Gaia Science Alerts program.*
  - *We report on the NIR brightening of the intermediate redshift quasar **PKS0735+17** (z=0.424), also known as **CGRaBSJ04738+1742**.*
- Coréférences exclues du processus d'annotation :
  - *Analysis of **the data** is ongoing. We remind the community that all **Swift data** are public, and encourage **their** use.*
  - ***The observations** continued until 2019-04-26 20 :15 UT, when **they** were aborted to begin followup of.*
  - *The estimated AB magnitude is **17.6**. **This magnitude** is not corrected for the host galaxy contribution.*

4. http://simbad.cds.unistra.fr/simbad/sim-id?Ident=Andromeda+Galaxy&NbIdent=1&Radius=2&Radius.unit=arcmin&submit=submit+id

3.4 Annotation des relations astrophysiques entre les mentions de type CelestialObject et leurs propriétés physiques

La détection de relation vise à établir les liens entre paires d'entités (Bassignana & Plank, 2022). Dans le cadre de notre étude, nous définissons un unique type de relation, qui relie un objet céleste à ses propriétés physiques. Dans le cas où plusieurs objets célestes sont mentionnés dans le texte, notre objectif est donc de pouvoir associer les propriétés physiques mentionnées à l'objet correspondant. Dans notre schéma d'annotation donc, seules les mentions de type `CelestialObject` sont reliées aux mentions d'entités décrivant des attributs physiques tels que `CelestialRegion` (coordonnées dans le ciel), `Flux` (énergie du corps par unité de temps), `Magnitude` (intensité lumineuse de l'objet céleste) etc.

4 Statistiques du corpus annoté

Dans cette section, nous décrivons les principales caractéristiques du corpus annoté résultant (statistiques globales et accord inter-annotateurs), et nous fournissons des tableaux comparatifs entre le corpus TDAC et notre nouveau corpus annoté astroECR.

4.1 Statistiques globales

Paramètres	**TDAC**		**astroECR**	
	Entraînement	**Test**	**Entraînement**	**Test**
# documents	59	16	210	90
# tokens	15374	3638	43481	10578
# tokens annotés	4338	1014	17392	3173
# mentions de coréférence	-	-	412	101
# chaînes de coréférence	-	-	257	65
Long. moyenne des chaînes	-	-	3,5 (+/- 2,26)	3,4 (+/- 1,61)
# relations intra-phrases	-	-	490	143
# relations inter-phrases	-	-	154	26
# total de relations	-	-	644	169

TABLE 1 – Statistiques de comparaison entre les corpus TDAC et astroECR.

La Figure 2 illustre la distribution des mentions d'entités nommées entre les corpus TDAC et astroECR. À l'issue de l'annotation, les types d'entités les plus représentées sont les classes spécifiques au domaine astrophysique telles que : les noms d'objets célestes (`CelestialObject`), les mentions de type `Magnitude` ou encore des mentions relatives à des outils d'observations tels que les télescopes (`Telescope`) et instruments (`Instrument`). Nous remarquons que les classes telles que `Software`, `Grant`, `Collaboration` ou encore `Archive` sont moins fréquentes dans les rapports d'observations.

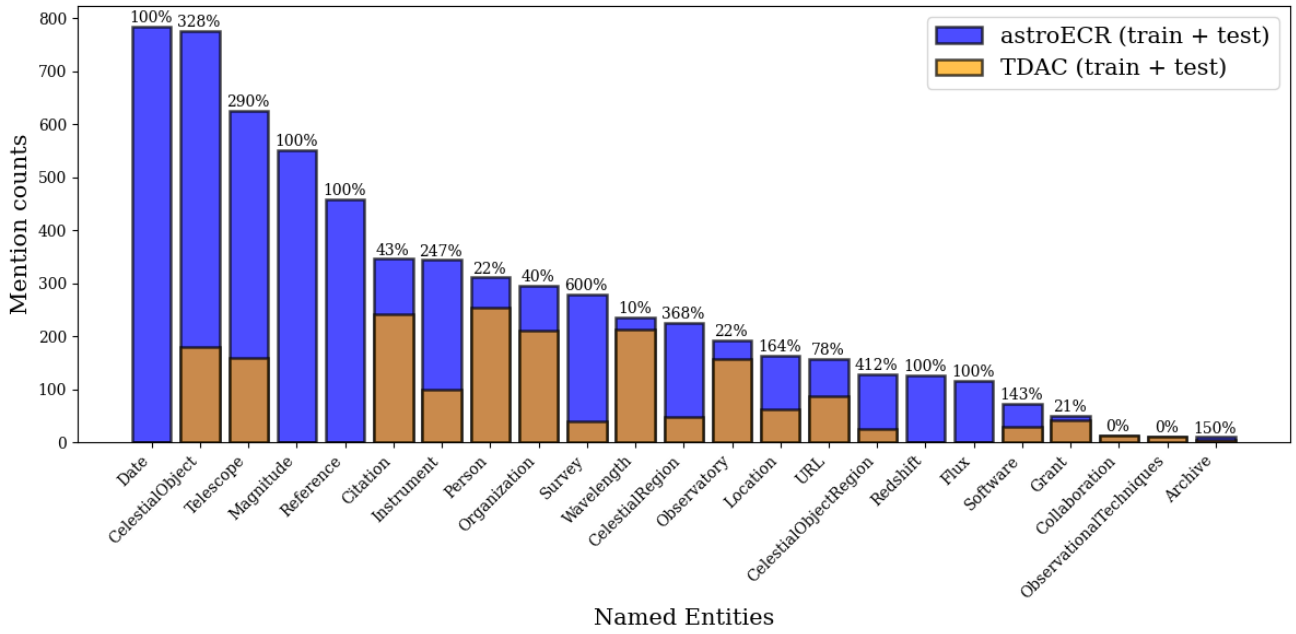


FIGURE 2 – Distribution des mentions d'entités nommées dans les corpus TDAC (en jaune) et astroECR (en bleu).

4.2 Accord inter-annotateurs et consensus

Nous avons impliqué un expert en astrophysique et un expert en TAL pour annoter un sous-ensemble du corpus (30 documents, soit 6499 unités lexicales). Les erreurs d'annotation ont été identifiées et corrigées au cours d'une phase de consensus, permettant la création du jeu de référence. Nous avons ensuite comparé les annotations des deux annotateurs avec le jeu de référence produit en utilisant les scores de précision, de rappel et de F-mesure, conformément à la méthodologie de Galibert *et al.* (2012). Les résultats du tableau 2 montrent que l'expert en astrophysique a obtenu une F-mesure plus élevée (0,94) que l'expert en TAL (0,91) par rapport au consensus (en évaluation souple), autorisant la poursuite de l'annotation par l'expert en astrophysique seul sur les 270 documents restants.

Tâche	**Annotateurs**	**Stricte**			**Souple**		
		P	**R**	**F1**	**P**	**R**	**F1**
Entités nommées	Astro vs. TAL	0,65	0,59	0,62	0,84	0,92	0.88
	Astro vs. consensus	0,83	0,86	0,84	0,93	0,96	0,94
	TAL vs. consensus	0,73	0,69	0,71	0,94	0,89	0,91
Coréférences	Astro vs. TAL	0,77	0,88	0,82	0,78	0,89	0,83
	Astro vs. consensus	0,97	1,00	0,98	0,97	1,00	0,98
	TAL vs. consensus	0,74	0,89	0,81	0,75	0,90	0,82

TABLE 2 – Accord inter-annotateurs pour l'annotation des entités nommées et des mentions de coréférences entre les deux annotateurs, et comparaison avec le consensus. L'annotateur astrophysicien est dénommé "*Astro*", et l'expert en TAL est dénommé "*TAL*". Les métriques utilisées sont la Précision (P), le Rappel (R) et la F-mesure (F1). Deux modes d'évaluation : stricte et souple. En évaluation stricte, une entité annotée est considérée comme vrai positive si le type d'entité et les frontières sont correctement annotées. En évaluation souple, les frontières d'annotation ne sont pas pénalisées.

5 Expériences

5.1 Configurations expérimentales

— **Reconnaissance d'entités nommées** : Nous avons ajouté et entraîné une tête de classification aux modèles astroBERT (Grezes *et al.*, 2021) et SciBERT (Beltagy *et al.*, 2019) sur le corpus astroECR$_{\text{train}}$, puis les avons évalués sur l'ensemble de test astroECR$_{\text{test}}$. L'entraînement a été effectué sur 20 époques, avec un taux d'apprentissage $\alpha = 2,10^{-5}$, et une taille de lot d'entraînement de 4. L'entraînement a été réitéré 5 fois avec des amorces différentes.

— **Normalisation des mentions de type `CelestialObject`** : Notre système interroge d'abord la base de données SIMBAD (Wenger *et al.*, 2000) avec des requêtes ADQL [5], extrayant l' identifiant unique de chaque objet céleste, sa désignation canonique, ainsi qu'une liste de toute ses désignations. Si une source n'est pas identifiée dans la base SIMBAD, la requête s'étent à la base NED (Mazzarella *et al.*, 2001) et, si nécessaire, à la base TNS (Gal-Yam, 2021).

— **Résolution des coréférences** : Nous avons évalué F-coref (Otmazgin *et al.*, 2022), un outil de résolution des coréférences basé sur l'architecture LingMess (Otmazgin *et al.*, 2023). Nous avons choisi F-coref en raison de sa facilité d'utilisation via sa bibliothèque Python *fastcoref* [6]. Nous avons procédé à une première évaluation du modèle sans entraînement en comparant ses prédictions avec nos annotations. Ensuite, nous avons entraîné le modèle sur 50 époques en utilisant astroECR$_{\text{train}}$ et l'avons évalué sur astroECR$_{\text{test}}$. Chaque expérience a été répétée cinq fois avec différentes des amorces aléatoires.

— **Détection de relations** : Nous avons entraîné un réseau de neurones de type biLSTM, que nous avons affiné sur l'ensemble d'entraînement pendant 20 époques avec un taux d'apprentissage $\alpha = 10^{-3}$, et une taille de lot d'entraînement fixée à 128. Nous avons évalué les performances du système sur astroECR$_{\text{train}}$.

5.2 Résultats sur le corpus de test astroECR$_{\text{test}}$

Dans cette section, nous présentons et analysons les résultats obtenus sur astroECR$_{\text{test}}$, l'ensemble de test d'astroECR, à l'issue de l'entraînement sur astroECR$_{\text{train}}$, son corpus d'entraînement.

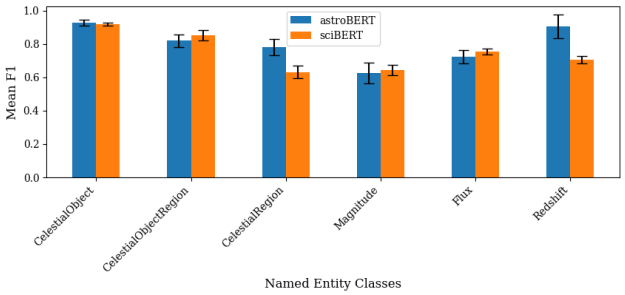
Détection d'entités nommées Le tableau 3 met en évidence la supériorité d'astroBERT par rapport à SciBERT en matière de rappel et F-mesure. En effet, une différence de 6 points sur la F-mesure globale (0,76 pour SciBERT, contre 0,82 avec astroBERT) est constatée. Une analyse plus détaillée des performances pour certaines classes spécifiques au domaine (figure 3a) montre qu'astroBERT est particulièrement plus performant dans la reconnaissance d'entités spécifiques à l'astrophysique, notamment le repérage des coordonnées célestes (`CelestialRegion`), caractérisée par une diversité de formes, ou encore pour la classe `Redshift`, de nature équationnelle. Ces résultats peuvent être attribués au fait qu'astroBERT est un modèle de langue pré-entraîné sur des textes spécifiques au domaine, ce qui renforce sa capacité à mieux repérer ces types d'entités.

5. Le language ADQL est basé sur le langage SQL, avec quelques extensions pour prendre en charge des requêtes spécifiques à l'astronomie, notamment pour des requêtes sur les coordonnées célestes.

6. https://pypi.org/project/fastcoref/

Modèle	Précision	Rappel	F-mesure
SciBERT	0,84 (0,01)	0,70 (0,01)	0,76 (0,01)
astroBERT	0,83 (0,01)	0,81 (0,01)	0,82 (0,01)

TABLE 3 – Performance moyenne (avec écart-type) des systèmes de REN fondés sur SciBERT et astroBERT. Les modèles ont subi cinq entraînements distincts avec diverses amorces sur l'ensemble d'entraînement d'astroECR, puis ont été évalués sur le jeu de test d'astroECR. Les métriques utilisées sont la Précision, le Rappel, et la F-mesure.



(a) F-mesure moyenne (avec écart-type) de classes d'entités nommées spécifiques au domaine.

Résolution des coréférences, normalisation des noms d'objets célestes, et détection de relations Les résultats du tableau 4 montrent que le système de base F-coref a une précision très faible (0.09) et un rappel élevé (0.26), entraînant un faible score F1 (0.13). Le modèle manque de connaissances spécifiques au domaine pour résoudre avec précision les coréférences dans ce contexte. Cependant, en affinant le modèle (astroFastCoref), il a pu apprendre des motifs spécifiques à l'astrophysique, le rendant plus efficace dans la résolution des coréférences liées aux objets célestes en atteignant un score F1 CoNLL de 0.53.

Modèle	CoNLL		
	Précision	Rappel	F1
F-coref	0,09 (0)	0,26 (0)	0,13 (0)
astroFastCoref	0,67 (0,01)	0,44 (0,01)	0,53 (0,01)

TABLE 4 – Précision moyenne, rappel et F-mesure (avec écart-type) du système F-coref évalué sur l'ensemble de test de notre corpus avec et sans affinage. Chaque expérience a été exécutée cinq fois (sur 50 époques lors de l'affinage) avec différentes amorces aléatoires.

Catalogue	Précision
SIMBAD	60,39
SIMBAD + NED	71,28
SIMBAD + NED + TNS	80,19

TABLE 5 – Précision d'un système de normalisation des noms d'objets célestes à l'aide de bases de données astronomiques.

Précision	Rappel	F1
0,77	0,80	0,79

TABLE 6 – Performance d'un système biLSTM de détection de relation entre un objet céleste et une propriété physique.

Le Tableau 6 présente les performances de notre système de détection de relations. La F-mesure de 0,79 suggère une performance satisfaisante dans l'identification de relations entre les objets célestes et les propriétés physiques.

5.3 Analyse des gains obtenus sur $\text{TDAC}_{\text{test}}$

Dans cette section, nous analysons l'intérêt du corpus d'entraînement astroECR pour la détection d'entités nommées sur le corpus TDAC. La Figure 3 illustre l'évolution des performances sur $\text{TDAC}_{\text{test}}$, le jeu de test du corpus TDAC (Alkan *et al.*, 2022), en fonction de différents corpus d'entraînement utilisés. Pour cela, nous comparons l'ensemble d'entraînement de référence $\text{TDAC}_{\text{train}}$, avec les deux ensembles d'entraînement $\text{DEAL}_{\text{train}}$ et $\text{astroECR}_{\text{train}}$. Pour ces deux ensembles, nous faisons varier la taille du corpus d'entraînement par incrément de 25%. Les résultats obtenus montrent clairement l'intérêt d'astroECR, permettant d'améliorer la F-mesure globale moyenne d'environ 4 points. Dans le cadre de notre travail, l'entraînement d'un système sur le corpus DEAL ne permet pas l'amélioration de la F-mesure. Ceci peut s'expliquer par la nature différente des documents du corpus DEAL (articles scientifiques), qui possèdent des propriétés différentes des rapports d'observation.

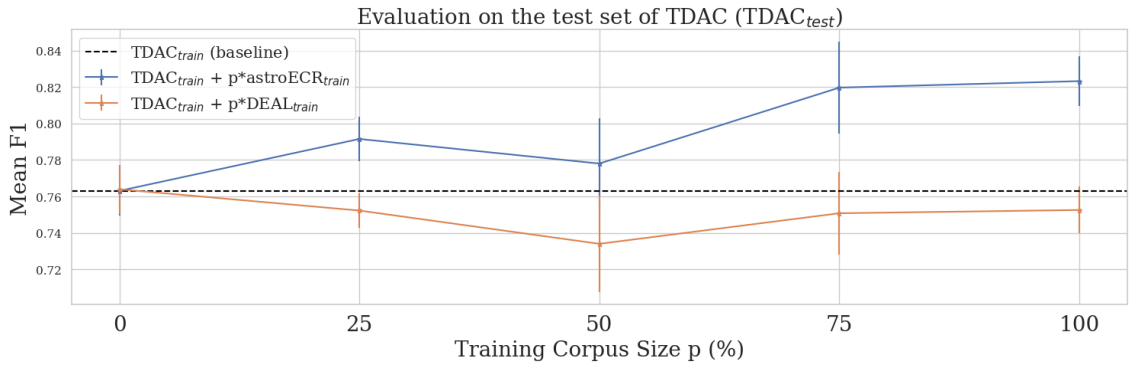


FIGURE 3 – Evaluation sur $\text{TDAC}_{\text{test}}$ d'un système de détection d'entités nommées à base d'un modèle astroBERT pour la détection d'entités nommées en fonction de la taille du corpus d'entraînement.

Classe	**$\text{TDAC}_{\text{train}}$**				**100% $\text{astroECR}_{\text{train}}$**				$\Delta F1$ (%)
	N	P	R	F1	N	P	R	F1	
`CelestialObject`	130	0,88	0,94	0,90	519	0,94	1,0	0,97	+ 7,7
`CelestialRegion`	20	0,31	0,23	0,26	149	0,64	1,0	0,78	+ 200
`Observatory`	60	0,54	0,58	0,64	101	0,80	0,67	0,72	+ 12,49
`Database`	36	0,75	0,81	0,78	79	0,77	0,90	0,83	+ 6,4

TABLE 7 – Comparaison des gains obtenus par classe sur le jeu de test $\text{TDAC}_{\text{test}}$ en fonction du corpus d'entraînement utilisé. Les métriques utilisées sont la Précision (P), Rappel (R) et la F-mesure (F1). N correspond au nombre de mentions d'entités de la classe dans le corpus d'entraînement.

Le Tableau 7 présente en détail les performances de certaines classes essentielles du domaine, notamment le repérage des noms d'objets célestes (`CelestialObject`), des coordonnées dans le ciel (`CelestialRegion`), des installations astronomiques impliquées dans l'observation (`Observatory`), ainsi que des bases de données astronomiques (`Database`). La plupart des classes bénéficient d'une amélioration des performances. Toutefois, la classe `CelestialRegion` est celle qui a le plus tiré profit de l'enrichissement du corpus. En effet, nous constatons une nette amélioration de la F-mesure (passant de 0,26 à 0,78). Cette progression significative s'explique par une augmentation marquée du rappel et une hausse plus modérée de la précision.

6 Conclusion et perspectives

Nous avons cherché à remédier au manque de données annotées, en élargissant le corpus TDAC (Alkan *et al.*, 2022) de 75 à 300 documents annotés. Notre corpus devient ainsi une ressource unique dans le domaine, proposant des annotations plus riches en entités nommées astrophysiques, dont cinq catégories nouvellement définies. Nous avons également annoté les coréférences et les relations entre objets célestes et leurs propriétés physiques, tout en normalisant les noms d'objets célestes à l'aide de bases de données astronomiques. À travers des expérimentations sur le corpus test d'astroECR, nous avons développé des modèles et fourni des scores de référence, soulignant l'utilité de notre ressource pour l'annotation automatisée de futurs documents. Nous avons démontré que l'augmentation de la taille du corpus améliore notablement la détection d'entités nommées sur le corpus de test de TDAC. Notre objectif est de mettre à disposition de la communauté TAL une ressource propice à des études complémentaires, telles que la résolution de coréférences scientifiques ou la détection de relations. Notre corpus peut également enrichir d'autres corpus spécialisés tels que ceux proposés par Chaimongkol *et al.* (2014) et Brack *et al.* (2021). À l'avenir, les modèles développés pourraient être utilisés à des fins d'extraction d'information comme suggéré par Sotnikov & Chaikova (2023). Nous mettons à disposition notre corpus, guide d'annotation, code, et modèles pour les deux communautés.

Références

ALKAN A. K., GROUIN C., SCHUSSLER F. & ZWEIGENBAUM P. (2022). TDAC, the first corpus in time-domain astrophysics : Analysis and first experiments on named entity recognition. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, p. 131–139, Online : Association for Computational Linguistics.

BANERJEE P. S., CHAKRABORTY B., TRIPATHI D., GUPTA H. & KUMAR S. S. (2019). A information retrieval based on question and answering and ner for unstructured information without using sql. *Wirel. Pers. Commun.*, **108**(3), 1909–1931. DOI : 10.1007/s11277-019-06501-z.

BARTHELMY S. D., BUTTERWORTH P. S., CLINE T. L., GEHRELS N., FISHMAN G. J., KOUVELIOTOU C. & MEEGAN C. A. (1995). BACODINE, the real-time BATSE gamma-ray burst coordinates distribution network. *Astrophysics and Space Science*, **231**, 235–238.

BASSIGNANA E. & PLANK B. (2022). What do you mean by relation extraction ? a survey on datasets and study on scientific relation classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop*, p. 67–83, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-srw.7.

BECKER M., HACHEY B., ALEX B. & GROVER C. (2005). Optimising selective sampling for bootstrapping named entity recognition. In *In Proceedings of the ICML Workshop on Learning with Multiple Views*, p. 5–11.

BELTAGY I., LO K. & COHAN A. (2019). Scibert : A pretrained language model for scientific text. In *EMNLP* : Association for Computational Linguistics.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

BRACK A., MÜLLER D. U., HOPPE A. & EWERTH R. (2021). Coreference resolution in research papers from multiple domains. *CoRR*, **abs/2101.00884**.

CHAIMONGKOL P., AIZAWA A. & TATEISI Y. (2014). Corpus for coreference resolution on scientific papers. In *Proceedings of the Ninth International Conference on Language Resources*

and Evaluation (LREC'14), p. 3187–3190, Reykjavik, Iceland : European Language Resources Association (ELRA).

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

GAL-YAM A. (2021). The TNS alert system. *Bulletin of the AAS*, **53**(1). https ://baas.aas.org/pub/2021n1i423p05.

GALIBERT O., ROSSET S., GROUIN C., ZWEIGENBAUM P. & QUINTARD L. (2012). Extended named entities annotation on OCRed documents : from corpus constitution to evaluation campaign. In *International Conference on Language Resources and Evaluation*, Istanbul, Turkey. HAL : hal-01831254.

GREZES F., BLANCO-CUARESMA S., ACCOMAZZI A., KURTZ M. J., SHAPURIAN G., HENNEKEN E. A., GRANT C. S., THOMPSON D. M., CHYLA R., MCDONALD S., HOSTETLER T. W., TEMPLETON M. R., LOCKHART K. E., MARTINOVIC N., CHEN S., TANNER C. & PROTOPAPAS P. (2021). Building astrobert, a language model for astronomy & astrophysics. *CoRR*, **abs/2112.00590**.

GREZES F., BLANCO-CUARESMA S., ALLEN T. & GHOSAL T. (2022). Overview of the first shared task on detecting entities in the astrophysics literature (DEAL). In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, p. 1–7, Online : Association for Computational Linguistics.

GRISHMAN R. & SUNDHEIM B. (1996). Message Understanding Conference- 6 : A brief history. In *COLING 1996 Volume 1 : The 16th International Conference on Computational Linguistics*.

HACHEY B., ALEX B. & BECKER M. (2005). Investigating the effects of selective sampling on the annotation task. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, p. 144–151, Ann Arbor, Michigan : Association for Computational Linguistics.

JOSHI M., LEVY O., ZETTLEMOYER L. & WELD D. (2019). BERT for coreference resolution : Baselines and analysis. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 5803–5808, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1588.

JURAFSKY D. & MARTIN J. H. (2023). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. USA, 3rd édition.

KIM Y. & WEBBER B. (2006). Implicit reference to citations : a study of astronomy. *ERPANET*.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LUAN Y., HE L., OSTENDORF M. & HAJISHIRZI H. (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In E. RILOFF, D. CHIANG, J. HOCKENMAIER & J. TSUJII, Éds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 3219–3232, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1360.

MAZZARELLA J. M., MADORE B. F. & HELOU G. (2001). Capabilities of the NASA/IPAC extragalactic database in the era of a global virtual observatory. In J.-L. STARCK & F. D. MURTAGH, Éds., *SPIE Proceedings* : SPIE. DOI : 10.1117/12.447177.

MOLLÁ ALIOD D., VAN ZAANEN M. & SMITH D. (2006). Named entity recognition for question answering. In L. CAVEDON & I. ZUKERMAN, Éds., *Proceedings of the Australasian Language Technology Workshop, ALTA 2006, Sydney, Australia, November 30-December 1, 2006*, p. 51–58 : Australasian Language Technology Association.

MURPHY T., MCINTOSH T. & CURRAN J. R. (2006). Named entity recognition for astronomy literature. In *Proceedings of the Australasian Language Technology Workshop 2006*, p. 59–66, Sydney, Australia.

NGUYEN T. D., TING Y.-S., CIUCĂ I., O'NEILL C., SUN Z.-C., JABŁOŃSKA M., KRUK S., PERKOWSKI E., MILLER J., LI J., PEEK J., IYER K., RÓŻAŃSKI T., KHETARPAL P., ZAMAN S., BRODRICK D., MÉNDEZ S. J. R., BUI T., GOODMAN A., ACCOMAZZI A., NAIMAN J., CRANNEY J., SCHAWINSKI K. & UNIVERSETBD (2023). Astrollama : Towards specialized foundation models in astronomy.

OTMAZGIN S., CATTAN A. & GOLDBERG Y. (2022). F-coref : Fast, accurate and easy to use coreference resolution. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing : System Demonstrations*, p. 48–56, Taipei, Taiwan : Association for Computational Linguistics.

OTMAZGIN S., CATTAN A. & GOLDBERG Y. (2023). LingMess : Linguistically informed multi expert scorers for coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 2752–2760, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.202.

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In M. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 2227–2237, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1202.

RUTLEDGE R. E. (1998). The Astronomer's Telegram : A Web-based Short-Notice Publication System for the Professional Astronomical Community. *Publications of the Astronomical Society of the Pacific*, **110**(748), 754–756. DOI : 10.1086/316184.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SEVGILI Ö., SHELMANOV A., ARKHIPOV M. Y., PANCHENKO A. & BIEMANN C. (2020). Neural entity linking : A survey of models based on deep learning. *CoRR*, **abs/2006.00575**.

SOTNIKOV V. & CHAIKOVA A. (2023). Language models for multimessenger astronomy. *Galaxies*, **11**(3). DOI : 10.3390/galaxies11030063.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). brat : a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 102–107, Avignon, France : Association for Computational Linguistics.

WENGER M., OCHSENBEIN F., EGRET D., DUBOIS P., BONNAREL F., BORDE S., GENOVA F., JASNIEWICZ G., LALOË S., LESTEVEN S. & MONIER R. (2000). The SIMBAD astronomical database. *Astronomy and Astrophysics Supplement Series*, **143**(1), 9–22. DOI : 10.1051/aas :2000332.

YADAV V. & BETHARD S. (2018). A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 2145–2158, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

ZHENG J., CHAPMAN W. W., CROWLEY R. S. & SAVOVA G. K. (2011). Coreference resolution : A review of general methodologies and applications in the clinical domain. *Journal of Biomedical Informatics*, **44**(6), 1113–1122. DOI : https ://doi.org/10.1016/j.jbi.2011.08.006.