Vers la traduction automatique des néologismes scientifiques

Paul Lerner, François Yvon

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.

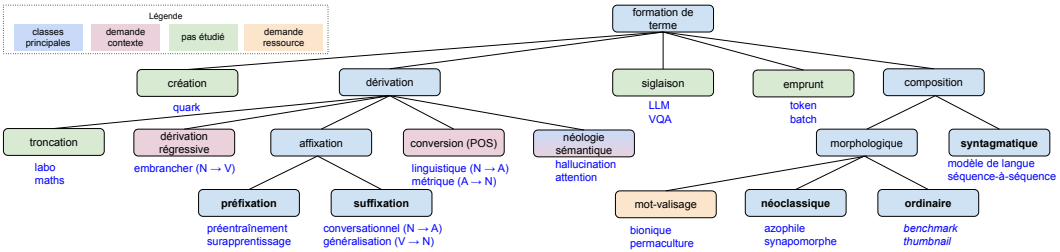


FIGURE 1 – Aperçu des procédés néologiques étudiés dans ce travail. D'après Daille (2017).

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 morpho-syntaxiques 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} :`

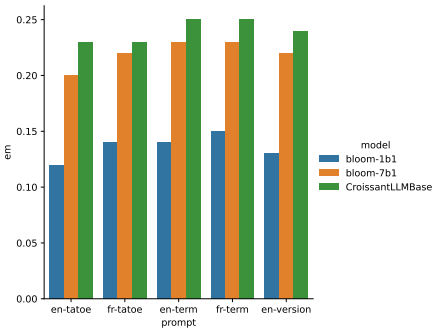
TABLE 1 – Les différentes amorces utilisées pour les modèles décodeurs.

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

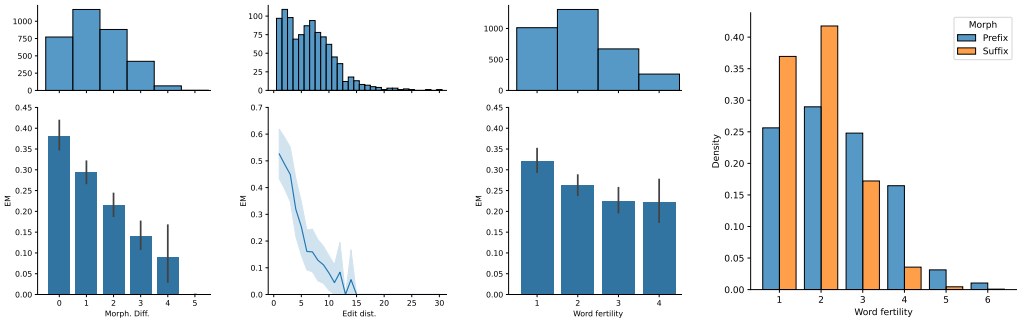
TABLE 2 – Pouvoir de prédiction (F1) des procédés morphosyntaxiques du terme français F selon les modèles de BLOOM-7,1G (et comparé aux procédés morphosyntaxiques du terme anglais A) sur le jeu de validation de FranceTerme.

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/



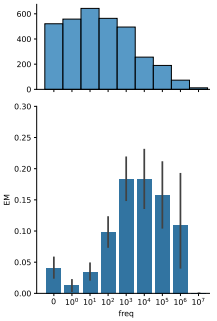
(a) (b) (c) (d)

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* → mycotoxine, *exocytosis* → exocytose, *iconomatic* → 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.