Annotation de la continuité référentielle dans un corpus scolaire – premiers résultats

Martina Barletta

Annotation de la continuité référentielle dans un corpus scolaire – premiers résultats

Martina Barletta[1, 2]
(1) LIDILEM, Université Grenoble Alpes, 1086-1366 Avenue Centrale, 38400, Saint-Martin-d'Hères, France
(2) Dipartimento di Scienze Umane per l'Educazione "Riccardo Massa", Université Milan-Bicocca, 1 Piazza dell'Ateneo Nuovo, 20126, Milan, Italie
martina.barletta@univ-grenoble-alpes.fr

RÉSUMÉ

La recherche Scolinter s'intéresse à l'étude des compétences en écriture des élèves de l'école primaire en France, en Italie et en Espagne. Le corpus éponyme se présente comme un large corpus longitudinal d'écrits d'élèves comparables dans les trois langues (Ponton *et al.*, 2021). Il s'agit dans cette recherche de créer un outillage TAL applicable à ce type de corpus pour assister les chercheurs dans la description linguistique des phénomènes qui relèvent de la cohésion et de la cohérence textuelle, en particulier de la continuité référentielle. La première étape de cette recherche a consisté dans la conception d'un modèle et dans le choix d'un format d'annotation répondant à ces objectifs. Cette contribution fera tout d'abord un état des recherches sur l'annotation en anaphore, coréférence et continuité référentielle avant de présenter les spécificités du corpus Scolinter et de proposer des pistes méthodologiques pour la suite du travail.

ABSTRACT

Annotating referential continuity in a children's writing corpus - first results

The Scolinter research project investigates the writing proficiency of primary school students in France, Italy, and Spain. The eponymous corpus consists of a large longitudinal corpus of comparable children's writing in these three languages (Ponton *et al.*, 2021). The aim of this project is to create a NLP tool tailored to this corpus to assist researchers in the linguistic description of phenomena relating to cohesion and textual coherence, in particular referential continuity. To meet these goals, the first stage of this research consists in designing a model and choosing an annotation format meeting these goals. This paper reviews the research on anaphora, coreference and referential continuity, presents the specific features of the Scolinter corpus, suggesting methodological pathways for further work.

MOTS-CLÉS : corpus scolaires, TAL, continuité référentielle, annotation de corpus.

KEYWORDS: children's corpora, NLP, referential continuity, annotated corpora.

1 Linguistique de corpus, corpus scolaires et annotation de la continuité référentielle

La linguistique de corpus s'intéresse à l'identification de phénomènes linguistiques sur une grande quantité de données. Couplée aux outils du traitement automatique des langues (TAL), elle permet d'attester et vérifier de manière empirique des hypothèses quant au fonctionnement du langage, encore plus dans les cas où ces phénomènes concernent des exemples de langue non standard, comme

dans l'analyse d'écrits scolaires. Dans le traitement de ce type de textes éloignés de la norme, les outils méthodologiques de la linguistique de corpus associés à l'outillage du TAL se heurtent à ces difficultés spécifiques (Wolfarth, 2019). En effet, l'utilisation d'outils développés pour de la langue standard s'avère délicate sur des corpus aussi fautifs. Toutefois, ces traitements automatiques sont nécessaires car ils permettent de traiter de vastes corpus fournissant ainsi des analyses fondées sur la réalité langagière (Jacques, 2005) ; analyses qui peuvent ensuite nourrir la réflexion didactique et la formation des enseignants (Elalouf, 2011; Elalouf & Perrin, 2019). Ceci est d'autant plus nécessaire quand les phénomènes analysés ont été rarement décrits de manière empirique, comme dans le cas du développement de la cohérence et de la cohésion textuelle dans les textes d'élèves de l'école primaire.

1.1 Les corpus Scoledit et Scolinter

Le corpus Scoledit représente actuellement un des seuls corpus longitudinaux d'écrits scolaires entièrement transcrits et librement accessibles en ligne (Ponton *et al.*, 2021). Ce corpus rassemble des textes narratifs sollicités auprès d'élèves de différentes écoles en France par les chercheurs du projet. Son but est de suivre de manière longitudinale le développement des compétences d'écriture des mêmes élèves, suivis du CP au CM2. Néanmoins, sur les 4 300 textes recueillis (Wolfarth, 2019, p. 101), 1 820 constituent la partie véritablement longitudinale du corpus, incluant 5 écrits par élève (un pour chaque année scolaire, du CP au CM2). Les productions écrites ont été recueillies par les chercheurs dans 38 écoles de 4 académies françaises (Bordeaux, Clermont-Ferrand, Grenoble et Lyon) sur la base de deux consignes, composées d'images et proposées aux élèves lors du recueil. La consigne utilisée en classe de CP propose de raconter l'histoire d'un petit chat à partir de 4 vignettes de type bande-dessinée. La consigne adoptée pour les classes de CE1 à CM2 propose aux élèves de choisir un ou des personnages présentés sur les vignettes (un chat, un loup, un robot et une sorcière), puis d'écrire une histoire les mettant en scène. L'utilisation de cette même consigne sur les quatre niveaux facilite la comparaison entre les productions produites au fil des années (Wolfarth, 2019). S'appuyant sur la même méthodologie de conception du corpus Scoledit, le projet Scolinter (Ponton *et al.*, 2021) vise la constitution d'un corpus de textes d'élèves du primaire en France, en Italie et en Espagne avec un suivi longitudinal des mêmes cohortes d'élèves. Les objectifs de ce corpus sont d'étudier les compétences en littéracie des élèves à chaque niveau, ainsi que l'évolution de ces compétences dans les trois langues tout au long de l'école primaire (Ponton *et al.*, 2021), sur la base de textes produits à partir d'une même consigne. Le corpus Scolinter comprend en l'état actuel : les 1 820 textes qui forment la partie longitudinale du corpus Scoledit ainsi que 1 333 textes en italien déjà traités, et 813 textes en espagnol. Le corpus français est complet sur les 5 niveaux de primaire alors qu'une partie des textes des corpus italien et espagnol est encore en phase de transcription et de normalisation, notamment pour les niveaux 2, 3 et 5 en Italie (équivalents aux niveaux CE1, CE2 et CM2), et pour les niveaux 3, 4 et 5 en Espagne (équivalents aux niveaux CE2, CM1 et CM2). Actuellement, nous exploitons ce corpus en nous intéressant plus particulièrement aux aspects cohérence/cohésion des textes à travers l'étude des chaines de continuité référentielle (Garcia-Debanc *et al.*, 2021) portant sur les personnages des histoires produites par les élèves. Dans cet objectif, nous annotons les mentions et les chaines de continuité référentielles qui font référence aux personnages induits par la consigne ou bien insérés par l'élève dans l'intrigue de la narration.

Le corpus Scolinter est composé des scans, des transcriptions ainsi que des normalisations des textes recueillis. Nous avons choisi d'annoter la version normalisée des textes des élèves, c'est-à-dire une version orthographiquement normée restant au plus près de la production initiale (Wolfarth, 2019). Elle permet à la fois des comparaisons avec la production de l'élève à différents niveaux (lexical, orthographique, morphologique...) et le recours aux outils TAL (Wolfarth *et al.*, 2018). Dans cette

normalisation, qui a été effectuée au format XML, des balises ont été introduites pour indiquer certains phénomènes spécifiques. Par exemple, une balise est proposée pour indiquer les tokens omis par l'enfant lors du processus d'écriture, là où le mot oublié influe sur la construction syntaxique de la phrase en question. Ceci constitue un obstacle pour notre travail, car ces omissions portent dans la plupart des cas sur des formes pronominales (51% des balises d'omission présentes du CE1 au CM2). L'absence de ces termes modifie l'analyse morphosyntaxique et en dépendance de la phrase, sur laquelle s'appuiera le processus d'annotation et d'analyse des chaines de continuité référentielle dans la suite de notre travail. Ces balises contiennent en l'état actuel seulement une proposition de catégorie grammaticale à laquelle le mot appartient, ce qui rend complexe la mise en œuvre d'une reconstruction automatisée de ces tokens dans les textes. Les textes ont été traités à l'aide du modèle transformeur bert-cased Flaubert (Devlin *et al.*, 2019; Le *et al.*, 2020) pour effectuer une substitution de ces balises qui indiquent une omission dans le texte avec le mot le plus probable selon le modèle.

2 Annotation de la continuité référentielle, des chaines de référence, de l'anaphore : état de l'art synthétique

La linguistique de corpus française s'intéresse, depuis des décennies, à l'étude et la description des phénomènes de construction de la textualité comme l'anaphore et les chaines de référence (Chastain, 1975; Corblin, 1985; Charolles, 1988; Corblin, 1995; Schnedecker, 1997, 2021). Bien que des corpus annotés en coréférence ou en anaphore existaient déjà avant la naissance de projets comme ANCOR (Muzerelle *et al.*, 2013) et DEMOCRAT (Landragin, 2016) en France, leurs caractéristiques ne les rendaient pas globalement représentatifs de la coréférence ou utilisables pour l'apprentissage profond (Grobol, 2020), soit parce que les annotations codent seulement certains types d'anaphore, comme dans le cas du corpus ARCADE (Tutin *et al.*, 2000), soit à cause de leur taille réduite, comme dans le cas du corpus DéDé (Gardent & Manuélian, 2005). En effet, même si ARCADE constitue un des premiers grands corpus annotés en anaphore pour le français avec son million de mots, seules les expressions anaphoriques et cataphoriques appartenant à des catégories fermées ont été retenues, peu importe leur antécédent. Les expressions anaphoriques et cataphoriques annotées sont : les pronoms personnels (à exception du pronom réfléchi), les pronoms et déterminants possessifs, les pronoms démonstratifs à l'exception des pronoms démonstratifs « neutres » (*ce, ça, cela, ceci*), les pronoms indéfinis et numéraux ; les adverbes anaphoriques comme *dedans, dessus* ; les expressions nominales anaphoriques ainsi que les « pointers » anaphoriques (*ce dernier, le premier*). Ces annotations encodent aussi la relation avec l'antécédent ainsi que la relation discursive et sémantique entre l'expression anaphorique et son antécédent (Tutin *et al.*, 2000). Cependant, dans ce corpus, les descriptions définies [1] ne sont pas annotées (Tutin *et al.*, 2000; Gardent & Manuélian, 2005).

Le corpus DéDé, pour sa part, cible l'annotation des descriptions définies, « c'est-à-dire les expressions de la forme *le/la/les N* » pour en rendre possible l'annotation automatique (Gardent & Manuélian, 2005, p. 3). Ce corpus est composé d'articles du journal Le Monde et il comprend 48 360 mots annotés au niveau morphosyntaxique, dont 4 910 descriptions annotées selon quatre catégories différentes : description autonome, description coréférentielle, description contextuelle, description non référentielle. Sa méthodologie d'annotation prévoyait l'utilisation d'outils de prétraitement de l'annotation (annotation morphosyntaxique) pour faciliter cette tâche, réalisée par des linguistes expérimentés. Le schéma d'annotation utilisé était affiné par plusieurs itérations après une première phase d'annotation, dont l'objectif était de résoudre les possibles désaccords entre annotateurs et d'intégrer les modifications nécessaires (Gardent & Manuélian, 2005), selon une stratégie déjà utilisée dans divers travaux d'annotation (Brants, 2000; Erk *et al.*, 2003; Gardent & Manuélian, 2005).

1. À savoir, les syntagmes du type article défini + nom.

Ces projets ont contribué à ouvrir la voie aux réflexions sur l'annotation de ces phénomènes sur des corpus de grande taille, ciblés dans certains cas pour l'entrainement de modèles *machine learning*. Le corpus Annodis (Péry-Woodley *et al.*, 2011), le corpus ANCOR (Muzerelle *et al.*, 2013) pour l'oral spontané, et le corpus DEMOCRAT (Landragin, 2016) pour la langue écrite représentent les trois principaux corpus qui encodent l'anaphore et/ou la coréférence à différents niveaux en français. Nous allons par la suite présenter les objectifs et la composition de ces corpus, ainsi que les méthodes d'annotations adoptées et les phénomènes qu'ils encodent. Nous aborderons également le corpus RésolCo (Garcia-Debanc *et al.*, 2019, 2021) qui représente le seul corpus français d'écrits scolaires de taille moyenne annoté en continuité référentielle, ce qui le rend similaire en objectifs et en méthodologie au corpus Scolinter.

2.1 Annodis

Le corpus Annodis (Péry-Woodley *et al.*, 2011) qui précède le corpus DEMOCRAT en élaboration et publication, tout en ayant une taille comparable à ce dernier, ne représente pas un corpus annoté en coréférence car il encode des chaines topicales, donc des segments caractérisés par l'apport d'informations au sujet d'un seul et même référent (Federzoni *et al.*, 2020). Il reste cependant un exemple de corpus de grande taille où des structures discursives ont été annotées manuellement. Le corpus Annodis est composé de textes variés et sélectionnés selon plusieurs critères, comme « le genre, la longueur et le type d'organisation discursive » (Péry-Woodley *et al.*, 2011, p. 72), et « il est le résultat de deux types d'annotations manuelles », une annotation qui part des unités élémentaires du discours pour reconstruire les relations rhétoriques entre unités du texte, dans une « démarche ascendante » qui vise à « construire la structure complète d'un discours » , et une deuxième annotation qui s'intéresse plutôt à la « mise en texte » et vise l'annotation sélective de structures discursives multi-échelles dont les structures énumératives et les chaines topicales (Péry-Woodley *et al.*, 2011, p. 72). La méthodologie et les questionnements en annotation qui ont marqué ce projet ont constitué les jalons du travail d'annotation fait sur le corpus RésolCo (Garcia-Debanc *et al.*, 2019, 2021), ce dernier étant lui-même inspiré des travaux effectués sur le corpus DEMOCRAT (Landragin, 2022).

2.2 ANCOR

Le corpus ANCOR représente « le premier corpus d'oral spontané d'envergure annoté en coréférence » et en anaphore pour la langue française et distribué librement (Muzerelle *et al.*, 2013). Il est composé de plusieurs corpus de parole spontanée transcrite (Accueil_USB, OTG et ESLO) et il compte 418 000 mots. Son objectif était de répondre au manque d'un corpus francophone en libre accès et « de taille suffisante pour entrainer un système de résolution de la coréférence efficace » (Muzerelle *et al.*, 2013, p. 2), à un moment où d'autres langues majoritaires étaient déjà dotées de tels corpus. Cependant, ce corpus était représentatif du français parlé conversationnel, alors qu'un corpus de taille similaire annoté pour la langue écrite était encore absent du panorama francophone. L'annotation a été réalisée de manière déportée sur ce corpus par deux annotateurs en quatre phases : repérage des entités nommées par un annotateur, consensus entre annotateurs par rapport à cette première annotation, repérage et caractérisation des relations anaphoriques par un annotateur et révision finale des relations caractérisées par un superviseur. Cette démarche a été adoptée pour éviter une « surcharge cognitive » des codeurs et pour favoriser l'accord interannotateur (Muzerelle *et al.*, 2013, p. 558).

2.3 DEMOCRAT

Le corpus DEMOCRAT répond à l'inexistence de grands corpus annotés représentatifs de l'écrit. Il constitue « le premier corpus de grande taille librement disponible pour le français écrit » (Landragin, 2021, p. 12) : 560 000 mots dont 198 000 expressions référentielles annotées et 20 000 chaines

de référence. Il est constitué de 58 textes appartenant à plusieurs genres, du roman aux articles journalistiques en passant par des textes littéraires historiques. L'objectif initial de ce projet était de constituer un corpus diachronique de textes, écrits entre le 12e et le 21e siècle, relevant de genres textuels variés, « et d'en autoriser des exploitations par des outils de traitement automatique des langues (TAL), plus précisément par des outils faisant appel à de l'apprentissage profond » (Landragin, 2022, p. 50). Ce corpus, sur lequel sont intervenues plus de 40 personnes, a fait l'objet de plusieurs expérimentations d'annotations pour vérifier la faisabilité du guide proposé. Il a également fait l'objet de séances d'annotation chronométrées pour calculer l'effort nécessaire pour réaliser l'annotation du corpus dans son entièreté. Après l'étape d'annotation manuelle, un script a été utilisé sur l'ensemble du corpus pour obtenir automatiquement la construction des chaines de référence à partir des annotations des différentes expressions référentielles (Landragin, 2021). Les choix méthodologiques fondamentaux du projet DEMOCRAT, qui le démarquent de projets existants similaires, sont le fait d'avoir annoté les chaines tout au long des textes inclus dans le corpus, ainsi que le fait d'avoir allié un travail d'annotation automatique à une génération automatique des chaines (Landragin, 2021, p. 20).

2.4 RésolCo

Dans le domaine des corpus d'écrits scolaires, dans le contexte du projet E-Calm, le corpus RésolCo (Garcia-Debanc *et al.*, 2017, 2021) a abordé l'annotation de la continuité référentielle sur des écrits de niveaux scolaires variés. Composé d'environ 400 textes, il a été récolté dans des classes de niveaux différents, du CE2 à l'université (Garcia-Debanc *et al.*, 2021). En s'appuyant sur l'expérience d'annotation faite lors de la conception d'Annodis ainsi que sur le guide et certains points méthodologiques du projet DEMOCRAT, ce corpus se donne le double objectif de (1) constituer une ressource annotée en continuité référentielle et de (2) élaborer une cartographie des formes linguistiques qui manifestent les compétences textuelles et discursives en cours de développement (Garcia-Debanc *et al.*, 2021). Les textes qui constituent ce corpus annoté ont été produits à partir de la même consigne, qui impose aux élèves la résolution de problèmes de cohésion textuelle (Garcia-Debanc & Bonnemaison, 2014; Garcia-Debanc & Bras, 2016). Cette consigne intègre trois phrases contenant des pronoms personnels (« ils » et « elle ») et des syntagmes nominaux introduits par un déterminant démonstratif (« cette maison », « ce grand bruit », « cette aventure »). La tâche impose aux élèves d'insérer ces trois phrases dans un texte narratif fictionnel (Garcia-Debanc *et al.*, 2021). Le phénomène annoté est défini « continuité coréférentielle » car les mentions annotées ne sont pas seulement celles qui représentent la coréférence stricte. Ces annotations sont aussi circonscrites « aux seuls référents présents dans l'ensemble des textes du corpus ; autrement dit, aux référents provoqués par la consigne RésolCo » (Garcia-Debanc *et al.*, 2021, p. 104). Le choix d'une annotation de ce type permet à l'annotateur de se concentrer sur les référents qui jouent un rôle de premier plan dans le récit, et permet aussi de créer des annotations comparables entre textes au même niveau ou tout au long du corpus.

Corpus	Année	Taille	Mentions annotées	Genre
ANCOR	2013	487 000 tokens	116 000 mentions, 51 000 relations anaphoriques	parole spontanée transcrite
DEMOCRAT	2019	58 textes 560 000 tokens,	198 000 mentions, 20 000 chaines de référence	écrits narratifs et autres genres variés
RésolCo	2021	385 textes, 72 873 tokens	12 261 mentions	écrits scolaires

TABLE 1 – Résumé des caractéristiques des corpus analysés

3 Annotation en continuité référentielle du corpus Scolinter

Bien que ces différents corpus s'intéressent tous aux phénomènes de cohérence et de cohésion textuelles, ils présentent des spécificités, à la fois liées aux genres textuels annotés et aux objectifs que les annotations contribuent à réaliser. Afin de mettre en place une annotation répondant au mieux à nos propres objectifs de recherche, nous avons comparé différents choix opérés, notamment au niveau méthodologique, dans les projets connexes, pour ensuite effectuer les choix les plus pertinents par rapport à nos objectifs d'annotation. Ce travail a abouti à la création d'une première version de notre guide d'annotation que nous avons ensuite testé sur notre corpus.

Afin de vérifier l'applicabilité du même guide sur deux des langues présentes dans le corpus, nous avons annoté 15 textes en français et 15 textes en italien. Cette première itération de test a été réalisée par plusieurs annotateurs experts et a comportée plusieurs sessions d'adjudications sur la base desquelles nous avons ultérieurement clarifié les descriptions des expressions linguistiques à annoter dans le guide. Ces premières annotations ont été utilisées ensuite pour produire les exemples montrés dans le guide. Certains des choix que nous avons faits à cette étape sont techniques et méthodologiques à la fois, comme par exemple le choix d'annoter les référents multiples à travers une superposition des étiquettes des référents indiqués, dans le tentative de résoudre partiellement le problème de l'annotation de l'anaphore discontinue.
Une deuxième itération d'application du guide a été effectuée sur des échantillons réduits de textes d'élèves du CE2, annotés par 22 annotateurs experts (étudiants en master de sciences du langage), à hauteur de 14 textes par binôme, ce qui nous a permis de vérifier la stabilité des lignes directrices décrites dans le guide. Les annotations obtenues par chaque binôme ont été analysées du point de vue qualitatif, selon les observations faites par les différents binômes dans leurs rapports finaux. Certaines des observations issues de ces annotations nous ont permis d'affiner ultérieurement les critères d'annotations, et d'apporter davantage d'exemples tirés du corpus dans le guide même, notamment en ce qui concerne la délimitation de certains types de mentions constituées par syntagmes nominaux. Cette itération a été cruciale pour la définition dans le guide du statut de l'annotation des pronoms dans le discours direct, et pour la première définition d'entité à annoter dans les textes en tant qu'entité animée et actante dans l'univers du texte analysé.
Suite à ces deux itérations de test du guide, nous avons mené une campagne d'annotation sur une partie restreinte du corpus français, sélectionnée pour être le plus représentative possible de la variété attestée dans le corpus longitudinal. Grâce à cette campagne nous avons obtenu un corpus de référence (corpus gold) qui nous a permis d'ébaucher une description des caractéristiques des chaines présentes dans des textes d'école primaire.

3.1 Le guide d'annotation

Le guide conçu pour l'annotation de la continuité référentielle s'inspire du travail effectué sur le corpus RésolCo, dans lequel sont annotées les chaines provoquées par les entités imposées par la consigne. En ce qui concerne Scoledit, nous avons décidé d'annoter les mentions (y compris singletons et anaphores) et les chaines relatives aux personnages présents dans la consigne ou ajoutés par l'élève. Dans le jeu d'étiquettes conçu, nous incluons les quatre personnages de la consigne (*cat*, *witch*, *robot*, *wolf*) ainsi que ces personnages « externes » qui interviennent de manière active dans l'intrigue du texte annoté à travers l'étiquette *extN*. Le *N* est rajouté à chaque nouvelle apparition d'un personnage. Dans le cas où seraient présents plusieurs personnages appartenant à la même catégorie, par exemple deux chats, les annotateurs rajoutent un chiffre en fin d'étiquette (*cat* et *cat1* si deux chats sont présents dans l'histoire, *cat2* si un troisième intervient et ainsi de suite). Concernant les passages à annoter, nous avons suivi les principes décrits dans le guide RésolCo, qui s'inspire lui-même du guide

d'annotation utilisé par le projet DEMOCRAT. Nous annotons : les syntagmes nominaux, les noms propres, les pronoms personnels, corrélés, objets et relatifs, les déterminants possessifs, l'anaphore zéro dans des phrases coordonnées ou dans les verbes à l'impératif en français et dans les phrases où le sujet n'est pas explicite en italien et en espagnol. Notre annotation tout comme celle de RésolCo ne relève pas de l'annotation de la coréférence stricte mais plutôt de la « continuité référentielle » au sens large, y compris les relations anaphoriques et les singletons présents dans les textes selon les critères définis auparavant. Nous n'annotons pas toutes les entités présentes dans les textes mais nous sélectionnons ces entités que nous considérons animées et récurrentes dans les textes pour permettre de comparer les résultats de manière longitudinale puis contrastive entre les trois langues du corpus.

3.2 Constitution et annotation du corpus de référence

Les textes à annoter ont été sélectionnés depuis le corpus longitudinal français, qui contient des textes des mêmes 337 élèves du CP au CM2. Le critère de sélection choisi est celui de la représentativité en termes de distribution par longueur des textes dans le corpus longitudinal sur chaque niveau scolaire. Après avoir effectué les traitement décrit dans 1.1 sur l'intégralité du corpus, nous avons réalisé des statistiques quant au nombre de tokens par texte. De cette manière, nous avons obtenu une image de la distribution des textes par nombre de tokens sur chaque niveau, pour pouvoir ensuite reproduire cette même distribution sur notre corpus de référence. Nous avons décidé d'exclure du corpus de référence les textes entre 1 et 20 tokens soit 21 textes au total, car nous les avons considérés trop courts pour contenir suffisamment d'informations quant à la cohérence textuelle. Nous avons partitionné le corpus par niveau et selon le nombre de tokens par texte (tranches 20-50 tokens, 51-100 tokens, 101-150 tokens etc.). Dans chaque tranche et pour chaque niveau, nous avons sélectionné aléatoirement le même pourcentage de textes que sur le corpus global pour conserver la même distribution. Nous avons décidé de rajouter davantage de textes en CE1 car la majorité des textes dans ce niveau sont assez courts et en CM2 pour enrichir l'échantillon de textes plus longues et donc susceptibles de contenir des chaines plus longues et complexes. Le corpus de référence est composé au final de 111 textes pour 16 838 tokens. Le corpus ainsi obtenu a été préalablement tokenisé automatiquement grâce à la librairie spacy-conll [2] de Spacy [3] et enregistré dans des fichiers au format CoNLL-U. Ces fichiers sont ensuite importés sur la plateforme INCEpTION (Klie *et al.*, 2018). Les textes ont été enfin annotés à l'aide de cette plateforme par deux annotateurs experts. Le schéma d'annotation s'appuie sur le layer d'annotation de la coréférence par défaut existant dans INCEpTION mais adapté à notre travail en proposant certaines étiquettes spécifiques comme *cat, witch, wolf*, et *robot*, qui permettent de suivre les personnages de la consigne et des étiquettes *ext1, ext2 et ext3* pour les autres personnages. Nous avons laissé les annotateurs libres de rajouter, si besoin, de nouvelles étiquettes sur la base de lignes directrices du guide (par exemple *ext4, wolf2*, etc.). Les analyses que nous présentons par la suite ont été effectuées à partir de l'export fourni par INCEpTION au format WebAanno TSV v3.3 grâce à des codes Python ciblés sur l'analyse des mentions et des chaines de référence annotées dans ce corpus.

4 Observation et résultats

Les annotations obtenues lors de cette première campagne sont en cours d'adjudication. Pour l'heure, l'adjudication a été opérée sur les textes présentant un écart important entre le nombre d'entités annotées et/ou le nombre de mentions annotées. Ceci a permis de faire ressortir les principales difficultés dans l'utilisation du guide d'annotation. Le processus d'adjudication en cours porte surtout sur des éléments tels que les personnages à annoter dans les textes (voir 4.2), sur l'annotation de

2. Disponible au lien suivant https://pypi.org/project/spacy-conll/. Consulté le 10/02/2024
3. Disponible au lien suivant https://spacy.io/. Consulté le 10/02/2024.

l'anaphore zéro (voir 4.3), ainsi que sur les limites des mentions constituées par des syntagmes nominaux, qui va nécessiter davantage d'explications surtout par rapport à la non inclusion des prépositions dans les syntagmes nominaux. Les résultats même partiels de l'adjudication nous permettent de faire ressortir certaines caractéristiques globales du corpus et des chaines annotées comme le nombre de maillons par texte, le nombre de référents annotés, la longueur des chaines, la présence des personnages issus de la consigne, etc. Ces premières mesures seront à confirmer une fois la phase d'adjudication achevée.

De nos annotations, nous avons exclu les singletons dans les calculs ici effectués. Nous avons retenu pour ces statistiques les chaines composées au moins de deux maillons, en raison de la longueur des textes présents dans notre corpus : certains textes sont assez courts (de 20 à 50 tokens) et on retrouve des chaines limitées à deux maillons par personnage. Ces chaines de deux maillons représentent 13% des chaines annotées dans nos textes. Les singletons concernent habituellement des personnages secondaires et non pas les quatre personnages de la consigne, par conséquent cette exclusion n'a pas eu d'impact sur les statistiques relatives à la présence des personnages de la consigne dans les textes.

4.1 Caractérisation du corpus de référence

Pour chaque texte du corpus de référence, nous avons calculé : le nombre de tokens, le nombre de maillons annotés (à l'exclusion des singletons), le nombre des chaines (à partir de deux maillons), la densité référentielle [4] et la longueur des chaines annotées. Nous avons pu observer que le nombre de tokens moyen par texte augmente avec le niveau scolaire, en même temps que le nombre d'entités présentes dans les textes et de maillons annotés. Comme remarqué par Landragin *et al.* (2024), en général, les textes narratifs présentent une densité référentielle plus importante que les textes appartenant à d'autres genres textuels [5], et cela confirme la densité calculée sur notre corpus (18,59% en moyenne sur tout le corpus de référence).

Cependant, une première comparaison quantitative entre les versions des deux annotateurs nous a permis d'observer un certain désaccord entre annotateurs, notamment en ce qui concerne le nombre de mentions ainsi que le nombre d'entités annotés dans les textes. Comme décrit dans 3.2, nous avons utilisé pour nos annotations le layer d'annotation de la coréférence proposé par défaut par INCEpTION. Même si la plateforme propose habituellement des fonctionnalités pour faciliter les étapes d'adjudication et du calcul de l'accord interannotateur, celles-ci n'ont pas été implémentées sur le niveau par défaut d'annotation de la coréférence. A terme, nous prévoyons de mettre en place des méthodes plus efficaces pour le calcul de l'accord interannotateur, ainsi que la création de notre propre layer d'annotation sur INCEpTION qui puisse nous permettre de mesurer cet accord interannotateur directement depuis la plateforme. Nous avons néanmoins pu observer des différences quant au nombre d'entités et de maillons annotés. Entre annotateurs, nous avons pu remarquer un écart type moyen de 2,06 sur le nombre de mentions annotées par texte et un écart type moyen de 0,55 sur le nombre d'entités annotées par texte. Parmi les différences observées dans une première étude qualitative, deux sont particulièrement saillantes, et en lien avec les observations quantitatives mentionnées. La première concerne l'annotation d'entités « externes » aux quatre personnages présents dans la consigne et la deuxième porte sur l'annotation de l'anaphore ou sujet zéro. Le tableau 2 résume les statistiques descriptives pour chaque niveau scolaire représenté dans le corpus de référence.

4. Nous définissons ici la densité référentielle comme le nombre des maillons divisé par le nombre de tokens.

5. Dans le corpus DEMOCRAT, la densité référentielle des textes narratifs du 16e siècle se situe à plus de 20,55%. (Landragin *et al.*, 2024)

Niveau	Nb textes	Nb tokens	Nb moyen tokens par texte	Maillons	Nb moyen de référents par texte	Densité référentielle	Longueur moyenne des chaines
CE1	32	2 388	74,63	461	2,22	19,30%	6,54
CE2	25	3 475	139	651	2,56	18,73%	10,68
CM1	25	4 541	168,19	868	3,11	19,11%	11,01
CM2	27	5 979	221,44	1 066	3,48	17,83%	11,6
Corpus	111	16 383	150,81	3 046	2,84	18,59%	9,96

TABLE 2 – Résumé des statistiques sur les chaines annotées dans le corpus de référence, inspiré de Landragin *et al.* (2024)

4.2 Ambiguïté du guide : l'identification des personnages « externes » à la consigne

Bien que la consigne de la tâche d'écriture cherche à imposer l'utilisation d'un ou deux personnages déterminés, cela n'a pas toujours été respecté par les enfants, et a donné lieu à un corpus riche en représentations de personnages différents. Dans notre guide, nous avons ciblé l'annotation de cette richesse, mais la définition donnée dans le guide de « personnage à annoter » s'est révélée floue par rapport à la réalité des textes auxquels nous avons été confrontés lors du processus d'annotation : si la notion de personnage animé tient à la confrontation avec la réalité des textes à annoter, la définition plus large donnée dans le guide d'« entités animées ou qui effectuent des actions utiles afin de suivre l'intrigue de la narration du texte »[6] n'était pas suffisamment claire et délimitée, ce qui a entrainé un désaccord sur le nombre d'entités annotées dans les textes.[7] Par exemple, dans le texte présenté dans la Figure 1, si on se limite à l'annotation des entités *cat* (présent dans la première partie du texte, qu'on ne reporte pas ici) et à la « protagoniste » *ext1*, on perd la présence des différents maillons qui indiquent les autres référents qui participent aux événements décrits dans le texte.

6. Le guide est actuellement disponible sur demande et pour les annotateurs participant au projet. Il sera publié lors de la publication du corpus.

7. Ce désaccord porte sur environ 50% des textes du corpus de référence.

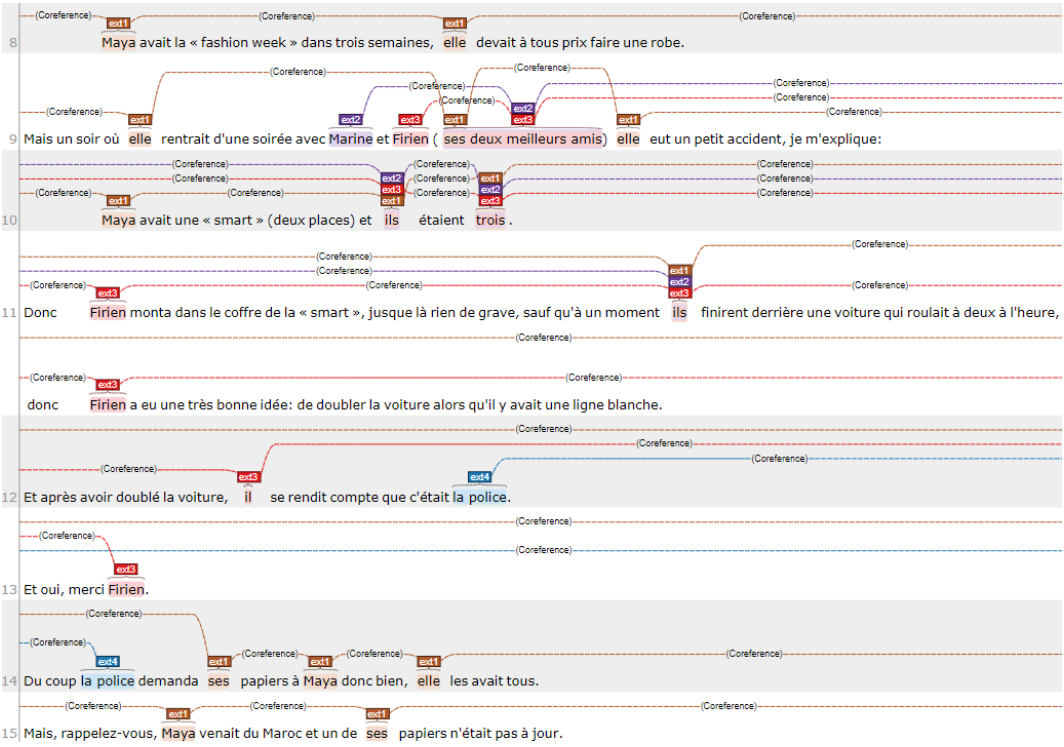

FIGURE 1 – Exemple de présence de personnages externes à la consigne dans les phrases 9 à 15 du texte NORM-EC-CM2-2018-14-D1-S1325. Annotation effectuée sur la plateforme INCEpTION.

Ces observations nous ont permis de faire évoluer la définition d'entité à annoter dans le texte à « entité animée dans le texte donné », en incluant des entités habituellement non animées mais qui prennent la forme de personnage animé dans le texte.

4.3 Anaphore zéro

Un autre point de difficulté que l'on a pu observer lors d'une analyse qualitative des annotations a porté sur l'annotation de l'anaphore zéro, ou sujet zéro. L'annotation de ce phénomène nous semble encore plus importante dans une perspective comparative car les deux autres langues du corpus, l'italien et l'espagnol, sont des langues où le sujet n'est pas obligatoirement exprimé à travers une marque pronominale, ce qui rend le phénomène de l'anaphore zéro très fréquent dans nos textes selon une première étude informelle. Toutefois, dans notre corpus de référence, ce phénomène linguistique n'est pas toujours annoté de manière cohérente. Si sa présence est toujours observable de manière fréquente dans les textes italiens et espagnols du corpus, son annotation pourrait être plus compliquée sur la partie française du corpus. Landragin *et al.* (2024) observent la différence dans l'annotation du sujet zéro entre français ancien et français contemporain, où les annotateurs du français moderne ont tendance à oublier ce phénomène, alors que sa présence et sa fréquence en ancien français « empêche tout oubli » (Landragin *et al.*, 2024, p. 15). De manière similaire entre le français et l'italien, notre hypothèse, suite aux tests effectués en parallèle sur des textes dans les deux langues, est que la

fréquence du phénomène en italien (ou dans la pratique d'annotation des annotateurs italophones) rend la présence de ce phénomène évidente et son annotation presque automatique, alors que cela ne l'est pas pour un annotateur francophone. Dans le texte d'un élève de CE1 (Figure 2), le verbe qui marque une anaphore zéro dans une phrase coordonnée (« voulait ») n'a pas été annoté par l'un des annotateurs. Cette difficulté rencontrée par les annotateurs sera abordée dans le guide, en fournissant davantage d'exemples de possibles formes verbales à annoter dans les textes en français car ils font effectivement partie des mentions à annoter.

FIGURE 2 – Exemple d'anaphore zéro dans la phrase 10 du texte NORM-EC-CE1-2015-130-D1-S1125. Annotation effectuée sur la plateforme INCEpTION.

5 Conclusion

Dans cette contribution, nous avons présenté les différentes étapes ayant mené à la constitution d'un corpus de référence annoté en continuité référentielle issu des textes du corpus Scolinter. Nous avons effectué plusieurs itérations d'annotation, suivies par des redéfinition et clarification du guide d'annotation. Celles-ci ont portées sur le plan technique, comme l'annotation empilée des référents multiples, ainsi que sur le plan méthodologique, comme l'annotation des pronoms dans le discours direct, ou l'éclaircissement des définitions d'anaphore zéro et de personnage à annoter dans les textes. Nous avons pu mener une campagne d'annotation conduite par deux annotateurs experts sur 111 textes, issus des productions d'élèves des niveaux scolaires du CE1 au CM2 du corpus français. Le corpus de référence a été sélectionné selon les critères de représentativité du corpus longitudinal en termes de distribution des textes par longueur sur chaque niveau scolaire retenu pour nos analyses. Cette première campagne d'annotation nous a permis d'effectuer une première description des anaphores et des chaines annotées dans notre corpus. Cette description semble confirmer que les textes présents dans le corpus correspondent à certains critères propres aux textes narratifs de scripteurs confirmés, ce qui confirme que le genre textuel influence profondément la construction des chaines de continuité référentielle. Les divergences que l'on a rencontrées dans les annotations nous ont permis d'identifier les « points faibles » de notre guide. Ceci nous permettra d'en proposer une nouvelle version, qui établira de manière plus claire les critères d'annotation des entités qui ne sont pas représentées dans la consigne et qui représente davantage d'exemples permettant aux annotateurs d'identifier de manière plus précise les occurrences d'anaphore zéro dans les textes du corpus français, et qui constituera la base de départ pour la rédaction d'un guide d'annotation spécifique pour la langue italienne.

Remerciements

Je remercie mon co-encadrant de thèse Claude Ponton pour les relectures de cette contribution ainsi que mes co-encadrantes de thèse Catherine Brissaud et Federica Da Milano pour leur aide.

Références

BRANTS T. (2000). Inter-annotator Agreement for a German Newspaper Corpus. In M. GAVRILIDOU, G. CARAYANNIS, S. MARKANTONATOU, S. PIPERIDIS & G. STAINHAUER, Éds., *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece : European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2000/pdf/333.pdf.

CHAROLLES M. (1988). Les plans d'organisation textuelle : périodes, chaînes, portées et séquences. *Pratiques*, **57**(1), 3–13. DOI : 10.3406/prati.1988.1468.

CHASTAIN C. (1975). Reference and Context. *Language, mind, and knowledge*, **7**, 194–269.

CORBLIN F. (1985). Les chaînes de référence : analyse linguistique et traitement automatique. *Intellectica. Revue de l'Association pour la Recherche Cognitive*, **1**(1), 123–143. DOI : 10.3406/intel.1985.851.

CORBLIN F. (1995). *Les formes de reprise dans le discours. Anaphores et chaînes de référence*. Presses Universitaires de Rennes. HAL : ijn_00550962.

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.

ELALOUF M.-L. (2011). Constitution de corpus scolaires et universitaires : vers un changement d'échelle ? *Pratiques. Linguistique, littérature, didactique*, (149-150), 56–70. DOI : 10.4000/pratiques.1702.

ELALOUF M.-L. & PERRIN S. (2019). Entre recherche et formation, quels usages des corpus de textes scolaires ? In *Écrire et faire écrire dans l'enseignement postobligatoire Enjeux, modèles et pratiques innovantes*, p. 197–212. Presses universitaires du Septentrion. DOI : https ://doi.org/10.4000/books.septentrion.77013.

ERK K., KOWALSKI A., PADÓ S. & PINKAL M. (2003). Towards a Resource for Lexical Semantics : A Large German Corpus with Extensive Semantic Annotation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, p. 537–544, Sapporo, Japan : Association for Computational Linguistics. DOI : 10.3115/1075096.1075164.

FEDERZONI S., HO-DAC L.-M. & REBEYROLLE J. (2020). Les chaînes topicales dans la ressource ANNODIS. *SHS Web of Conferences, Congrès Mondial de Linguistique Française CMLF 2020*, **78**, 11005. DOI : 10.1051/shsconf/20207811005, HAL : hal-02890989.

GARCIA-DEBANC C. & BONNEMAISON K. (2014). La gestion de la cohésion textuelle par des élèves de 11-12 ans : réussites et difficultés. *SHS Web of Conferences, 4e Congrès Mondial de Linguistique Française*, **8**, 961–976. DOI : 10.1051/shsconf/20140801349.

GARCIA-DEBANC C. & BRAS M. (2016). Vers une cartographie des compétences de cohérence et de cohésion textuelle dans une tâche-problème de production écrite réalisée par des élèves de 9 -12 ans : indicateurs de maitrise et progressivité. **Recherches textuelles**(13). HAL : hal-01987031.

GARCIA-DEBANC C., HO-DAC L.-M., BRAS M. & REBEYROLLE J. (2017). Vers l'annotation discursive de textes d'élèves. *Corpus*, **16**, 157–184. DOI : 10.4000/corpus.2783, HAL : hal-01558836.

GARCIA-DEBANC C., HO-DAC L.-M., FEDERZONI S., BRAS M. & REBEYROLLE J. (2019). RésolCo un corpus de manuscrits d'élèves et d'étudiants pour l'étude de la cohérence. In *10èmes Journées Internationale de la Linguistique de Corpus*, Grenoble, France. HAL : hal-02877122.

GARCIA-DEBANC C., REBEYROLLE J. & HO-DAC L.-M. (2021). La continuité référentielle dans le corpus RésolCo : méthode d'annotation et premières analyses. *Langue francaise*, **211**(3), 99–114. DOI : 10.3917/lf.211.0099, HAL : hal-03559961.

GARDENT C. & MANUÉLIAN H. (2005). Création d'un corpus annoté pour le traitement des descriptions définies. *Revue TAL*, **46**(1), 115. HAL : halshs-00168567.

GROBOL L. (2020). *Coreference resolution for spoken French*. Thèse de doctorat, Université Sorbonne Nouvelle - Paris 3. HAL : tel-02928209.

JACQUES M.-P. (2005). Pourquoi une linguistique de corpus ? In G. WILLIAMS, Éd., *La linguistique de corpus*, Rivages Linguistiques, p. 21–30. Rennes, presses universitaires de rennes édition.

KLIE J.-C., BUGERT M., BOULLOSA B., ECKART DE CASTILHO R. & GUREVYCH I. (2018). The INCEpTION Platform : Machine-Assisted and Knowledge-Oriented Interactive Annotation. In D. ZHAO, Éd., *Proceedings of the 27th International Conference on Computational Linguistics : System Demonstrations*, p. 5–9, Santa Fe, New Mexico : Association for Computational Linguistics.

LANDRAGIN F. (2016). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, (92), 11. HAL : hal-01347949.

LANDRAGIN F. (2021). Le corpus DEMOCRAT et son exploitation. Présentation. *Langages*, **224**(4), 11–24. DOI : 10.3917/lang.224.0011, HAL : hal-03474748.

LANDRAGIN F. (2022). Expressions référentielles et chaînes de référence en français : le projet Democrat et son exploration des rapports entre linguistique textuelle et linguistique de corpus. *Echo des études romanes*, **18**(1), 49–65. DOI : 10.32725/eer.2022.004, HAL : halshs-03876206.

LANDRAGIN F., GLIKMAN J., SCHNEDECKER C. & TODIRASCU A. (2024). Chaînes de référence dans le corpus Democrat : une analyse en diachronie longue. *Corpus*, (25). DOI : 10.4000/corpus.8581.

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 : European Language Resources Association.

MUZERELLE J., LEFEUVRE A., ANTOINE J.-Y., SCHANG E., MAUREL D., VILLANEAU J. & ESHKOL I. (2013). ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In ATALA, Éd., *TALN'2013, 20e conférence sur le Traitement Automatique des Langues Naturelles*, p. 555–563, Les Sable d'Olonne, France. HAL : hal-01016562.

PONTON C., GUTIÉRREZ-CACERES R., TERUGGI L., FARINA E., BRISSAUD C. & WOLFARTH C. (2021). Scolinter : un corpus trilingue. L'exemple de la segmentation en mots. *Langue française*, **211**(3), 37–50. DOI : 10.3917/lf.211.0037, HAL : halshs-03384027.

PÉRY-WOODLEY M.-P., AFANTENOS S., HO-DAC L.-M. & ASHER N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *Revue TAL : traitement automatique des langues*, **52**(3), 71. HAL : halshs-00935201.

SCHNEDECKER C. (1997). *Nom propre et chaînes de référence*, volume 21 de *Recherches linguistiques*. Université de Metz : Librairie Klincksieck. HAL : hal-00808797.

SCHNEDECKER C. (2021). *Les chaînes de référence en français*. Collection l'Essentiel français. Paris : Éditions Ophrys.

TUTIN A., TROUILLEUX F., CLOUZOT C., GAUSSIER É., ZAENEN A., RAYOT S. & ANTONIADIS G. (2000). Annotating a large corpus with anaphoric links. In *Third International Conference*

on Discourse Anaphora and Anaphor Resolution (DAARC2000), p.2̃, United Kingdom. HAL : hal-00373327.

WOLFARTH C. (2019). *Apport du TAL à l'exploitation linguistique d'un corpus scolaire longitudinal*. Thèse de doctorat, Université Grenoble Alpes. HAL : tel-02517320.

WOLFARTH C., BRISSAUD C. & PONTON C. (2018). Transcrire et normer un corpus scolaire : pour quelles analyses ? In C. BRISSAUD, M. DREYFUS & B. KERVYN, Éds., *Repenser l'écriture et son évaluation au primaire et au secondaire*, volume 36 de collection Diptyque, p. 121–145. Presses universitaires de Namur. HAL : hal-01883221.