DeCorScol : conception d'un outil d'assistance à l'annotation des chaînes de coréférence dans les écrits scolaires

Martina Barletta

DeCorScol : conception d'un outil d'assistance à l'annotation des chaînes de coréférence dans les écrits scolaires

Martina
BARLETTA

Sous la direction de Claude Ponton

Laboratoire : LIDILEM

UFR LLASIC
Département d'informatique intégrée en Langues, Lettres et Langage (I3L)

Mémoire de master 2 recherche en Sciences du Langage - 30 crédits

Parcours Industries de la langue

Année universitaire 2021-2022

DeCorScol : conception d'un outil d'assistance à l'annotation des chaînes de coréférence dans les écrits scolaires

Martina
BARLETTA

Sous la direction de Claude Ponton

Laboratoire : LIDILEM

UFR LLASIC
Département d'informatique intégrée en Langues, Lettres et Langage (I3L)

Mémoire de master 2 recherche en Sciences du Langage - 30 crédits

Parcours Industries de la langue

Année universitaire 2021-2022

Remerciements

Je tiens à remercier Claude Ponton, directeur de ce mémoire, pour ses enseignements tout au long des deux années de master, pour sa patience et le suivi bienveillant de ce travail de recherche.

Je souhaite aussi remercier Catherine Brissaud pour son aide. Je remercie également Luca Pallanti pour sa contribution indiscutable aux réflexions théoriques autour de ce projet ainsi que Rafaela Gutierrez-Caceres pour les conseils bibliographiques et les observations ponctuelles. Merci à Marie-Paule Jacques pour sa contribution à la dernière version de ce travail.

Merci à mes enseignants du Master Industries de la langue, en particulier à Nicolas David, Sylvain Coulange, Véronique Aubergé et Thomas Lebarbé.

Merci aussi à mes camarades de Master pour leur soutien et les discussions dans le patio de I3L (ou sur Zoom).

Grazie alla mia famiglia, sparsa per il mondo, per il loro amore e a Jesús, master of PowerPoint e tortillas.

Déclaration anti-plagiat

DECLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : Martina PRENOM : Barletta

DATE : 29/05/2022

Mise à jour mars 2021

Sommaire

MOTS-CLÉS : écrits scolaires, coréférence, traitement automatique des langues, cohérence, cohésion

Résumé

Quelles sont les qualités qui distinguent un texte « bien écrit » d'un texte « mal écrit », « incompréhensible », « peu ou pas cohérent » ? Divers facteurs entrent en jeu lors de la compréhension (et écriture) d'un texte sur plusieurs plans langagiers. Entre les différents éléments qui rendent un texte cohérent et cohésif, les chaînes de coréférence jouent sans doute un rôle de premier plan. Elles permettent aux lecteurs de suivre le cheminement des personnages impliqués dans les actions évoquées par le récit et, aux scripteurs plus avancés, d'attribuer des qualités à leurs personnages, ainsi que de les faire agir dans l'univers fictionnel qu'ils décrivent.

Cependant, comment se réalise la coréférence dans les écrits scolaires et comment est-il possible de la décrire de manière à fournir des critères d'évaluation objectifs aux enseignants ? L'un des buts de ce travail est de fournir un outil TAL d'assistance à l'annotation de ce phénomène, de manière à contribuer à cette « cartographie de l'emploi des formes linguistiques qui déterminent la cohérence et la cohésion textuelles » déjà initiée en France (Garcia-Debanc *et al.*, 2021).

Nous allons introduire dans le chapitre 1 la notion de coréférence en linguistique et en Traitement Automatique des Langues. Nous décrirons ensuite le domaine des corpus scolaires dans le chapitre 2 puis nous aborderons les approches utilisées pour l'annotation de la coréférence sur des corpus variés dans le chapitre 3. La question de la coréférence et de l'annotation de la continuité référentielle dans les corpus scolaires sera traitée dans le chapitre 4.

Depuis les premières recherches en TAL sur l'écriture des apprenants dans les années 80, ce sont des questions didactiques qui pilotent les travaux

linguistiques sur les textes d'élèves (Doquet et Ponton, 2021). En nous positionnant dans cette tradition, nous souhaitons ici pouvoir répondre à certains questionnements partagés avec la didactique, en recourant au développement et à l'application d'un outil de Traitement Automatique des Langues. Celui-ci servira d'aide à l'annotation des chaînes de coréférence sur les données du corpus scolaire Scoledit (CE2). Cet outil, nommé *DeCorScol* (*Détection des Coréférences sur les écrits Scolaires*), a comme but d'offrir une aide automatisée pour l'annotation des chaînes de coréférences dans le cadre du corpus *Scoledit*. La méthodologie de travail que nous avons adoptée sur le corpus Scoledit sera présentée dans le chapitre 5. Nous décrirons ensuite l'architecture de l'outil dans le chapitre 6. Cet outil nous a déjà permis d'effectuer des tests d'annotation. Sur la base de ces données, nous avons pu tirer nos premières réflexions que nous présenterons plus en détail dans le chapitre 7. Elles portent à la fois sur les configurations des chaînes de coréférence dans les écrits scolaires de niveau CE2 et sur la manière dont la modélisation linguistique et le TAL nous permettent de cerner cette configuration à savoir : 1. la nature facultative des ressources sémantiques pour la détection des mentions nominales dans des productions de CE2, donc l'observation d'une variété lexicale limitée dans ces types de mentions. 2. le postulat d'une prépondérance de continuité thématique dans la construction des récits par la part des élèves ; 3. la productivité de certaines règles déjà exploitées pour la détection de coréférences dans des textes rédigés par des scripteurs experts (comme les verbes pronominaux) et l'absence d'autre structures syntaxiques plus complexes, également exploitées afin de résoudre la coréférence dans des textes de scripteurs experts, dans les écrits scolaires du niveau objet de notre analyse.

L'outil que nous proposons est fonctionnel pour l'étude du développement des notions de textualité et de cohérence textuelle tout au long du parcours d'apprentissage de l'écriture chez l'enfant. Nous prévoyons

effectivement que cet outil puisse être davantage utilisé sur les niveaux restants du corpus objet de nos études, de manière à fournir par la suite une cartographie longitudinale des éléments qui composent la notion de coréférence dans les écrits scolaires de ces niveaux.

Légende

Dans les exemples extraits du corpus, nous avons utilisé un code couleur pour distinguer les mentions qui réfèrent aux quatre personnages du corpus *Scoledit* et les mentions qui portent sur d'autres référents

robot

sorcière

chat

loup

autre référent

Les productions du corpus *Scoledit* ont été indiquées avec la nomenclature utilisée au sein du projet. Elle est constituée de ces éléments :

NORM-EC-CE2-2016-104-D1-S835

NORM : normalisation

EC : école

CE2 : niveau

2016 : année de recueil

104 : identifiant de la classe

D1 : Devoir 1

S – corpus *Scoledit*

853 – identifiant attribué à l'élève

Dans ce mémoire, nous avons utilisé des guillemets dans le cas où nous avons reporté dans le discours des citations tirées des extraits du corpus, dans le cas de citations courtes, et nous avons indiqué en cursives les mots en tant que tels (lemmes, etc.) et les étiquettes d'analyse en dépendance (telles que *nsubj* ou *obl:arg*). Nous avons indiqué toujours en majuscules les étiquettes des catégories grammaticales telles que NOUN, ADJ ou PRON.

Partie 1

-

La coréférence et les écrits scolaires

Chapitre 1. Définitions de la coréférence, entre linguistique et Traitement automatique des Langues

Plusieurs facteurs entrent en jeu lors de la compréhension (et écriture) d'un texte. Parmi les différents niveaux stratifiés qui distinguent la compétence d'un scripteur expert, la cohérence textuelle est définie comme le rapport de sens entre énoncés qui constituent un discours (Charolles, 2011), où « [c]es rapports de sens rendent l'enchaînement des énoncés cohérent et par là même, donnent au texte entier sa cohérence. » (Bonnemaison, 2018 : 21)

Dans cette première partie nous allons donner quelques définitions de la coréférence et de la notion de chaine de coréférence. En particulier, nous allons définir ce phénomène dans les domaines de la linguistique et du Traitement Automatique des Langues pour poser les bases nécessaires à la présentation du travail de modélisation et de détection de ces chaînes.

Nous allons ensuite aborder la description des corpus d'écrits scolaires en France : quels projets se sont emparés de cette thématique, avec quelle démarche méthodologique, pour quelles possibles applications de TAL ? Nous allons nous intéresser en particulier au corpus *RésolCo*, qui a pour objectif l'étude de la construction des chaînes de coréférence, et au corpus *Scoledit*, qui suit une même cohorte d'élèves du primaire de manière longitudinale et qui constitue le contexte de cette recherche.

Nous allons également décrire les différents projets existants dans le cadre de l'annotation et de la résolution des chaines de coréférences pour la langue française. Deux approches principales coexistent dans le milieu, l'approche symbolique et l'approche neuronale. Les deux types de démarche seront abordés à travers des exemples concrets d'outils et de projets. Nous

allons discuter les points forts et les points faibles des différentes approches pour construire un outil applicable aux données à notre disposition.

En dernier lieu, nous allons présenter les approches outillées existantes pour l'annotation comme *CorefUD* (*Coreference in Universal Dependencies*,(Nedoluzhko *et al.*, 2021) et certains outils d'annotation automatique en coréférence.

Cet état de l'art nous permettra de poser les bases des choix d'annotations effectués dans le cadre de ce projet.

1. La coréférence. Bref cadre théorique

Pour définir plus en détail la notion de coréférence, nous allons en proposer quelques définitions tirées de la littérature en linguistique et en TAL, avant de décrire les expressions linguistiques qui peuvent constituer une chaîne de coréférence et selon quels critères ces chaînes peuvent être décrites et modélisées. Nous allons aussi mentionner les métriques utilisées dans l'évaluation de la qualité des outils dédiés à la résolution de la coréférence, avant de mentionner les problématiques qui existent en annotation de ce phénomène. Nous allons aussi citer quelques outils et projets qui ont déjà abordé cette question en français et à l'international.

1.1. Définition de la coréférence

La notion de chaîne de référence se configure comme notion plus récente par rapport à celle d'anaphore, comme observé par Schnedecker (2019 : 1) ; ce terme apparait en linguistique française avec Chastain (1975), Corblin (1985, 1995) et Charolles (1988b : 8), et elle est définie de la manière suivante par ce dernier :

> Les chaînes sont constituées par des suites d'expressions coréférentielles [...]. Seules peuvent appartenir (donner lieu à) une chaîne des expressions employées référentiellement, c'est-à-dire toutes et rien que les expressions nominales (ou

pronominales) permettant d'identifier un individu (un objet de discours) quelle que soit sa forme d'existence (personne humaine, événement, entité abstraite)

La coréférence peut être autrement définie comme la « détermination d'un set de mentions qui se réfèrent à la même entité » (Stuckardt, 2016). L'interprétation de ces expressions linguistiques, liées entre elles, peut construire une relation d'identité référentielle (Corblin, 1985). Nous allons aussi voir par la suite que l'exclusivité des expressions nominales ou pronominales comme appartenant aux chaînes de coréférence est parfois élargie à d'autres expressions linguistiques qui font également référence à un « individu » représenté dans la phrase ou, en d'autres mots, à un *référent* donné.

Pour construire une chaîne de coréférence, il faut au moins trois expressions faisant référence à la même entité, appartenant à l'univers du texte (Schnedecker, 1997). Si on a moins de trois composantes, les notions d'*anaphore* et de *coréférence* sont suffisantes pour une description pertinente des phénomènes de paires d'enchaînements référentiels.

En TAL, la différence entre coréférence, chaîne de coréférence et anaphore est moins tranchée. Depuis l'introduction du terme de *coréférence* en TAL lors du MUC-6 (MUC Consortium, 1995b : 6), dans le contexte d'une tâche très similaire à la résolution d'anaphore, ces deux termes, *coréférence* et *anaphore,* sont quasiment synonymes et il existe un certain degré de confusion en littérature par rapport aux différences et similarités entre ces deux tâches (Poesio, 2016) qui se chevauchent partiellement mais qui ne représentent pas toujours la même tâche en TAL. La même quasi-synonymie existe entre *coréférence* et *chaîne de coréférence* : « (…) les spécialistes de TAL assimilent coréférence et chaîne de référence en un seul objet d'étude, le plus souvent sous le nom chaîne *de coréférences*, où le mot *coréférence* apparaît indifféremment au singulier ou au pluriel. » (Landragin, 2021 : 202). C'est pourtant utile ici de définir les caractéristiques fondamentales du phénomène de coréférence et

d'expliciter également les différences et similarités qui existent entre coréférence et anaphore.

Le terme « anaphorique » est souvent utilisé pour indiquer « les expressions dont l'interprétation dépend des objets introduits dans l'univers du discours, étant mentionnés ou seulement inférés »[1] (Poesio, 2016 : 38). La relation anaphorique entre deux expressions linguistiques est aussi définie comme asymétrique car elle « suppose la mise en relation d'une expression non autonome du point de vue de la référence et d'une expression référentielle susceptible de la « saturer » (…) », comme reporté par Schnedeker (2019 : 11) en utilisant les termes de Corblin (1995). Pour ce qui concerne la coréférence, cette notion suppose « une forme d'identité référentielle entre les référents évoqués » (Schnedeker, 2019 : 13) sans pour autant reposer sur une interdépendance entre mentions afin de pouvoir les interpréter et relier au référent évoqué, ce qui configure la relation de coréférence comme symétrique.

Toujours selon Poesio (2016 : 39), deux expressions peuvent être coréférentielles lorsqu'elles désignent le même objet de l'univers du texte pris en considération. Toutefois, certaines expressions peuvent ne pas être coréférentielles mais seulement anaphoriques. C'est le cas d'expressions dans des contextes hypothétiques ou de négation (Partee, 1970) comme dans les exemples (1). Si l'objet n'existe pas, on ne peut pas parler de coréférence, mais on peut toujours désigner les deux référents comme anaphoriques.

(1) a. Si je possédais une voiture, je l'utiliserais très souvent.

(1) b. Je ne peux pas acheter une voiture. Je ne saurais pas où la garer.[2]

[1] Notre traduction depuis l'anglais. « As we said above, we use the term anaphoric to indicate expressions whose interpretation depends on objects introduced in universe of discourse U either by virtue of being explicitly mentioned (like engine E3 in (22)) or by being inferred (as in the cases of plurals and propositional anaphora). » (Poesio, 2016 : 38)

[2] Les exemples 1a et 1b sont librement inspirés et traduits depuis les exemples de Poesio (2019).

Il peut y avoir aussi des expressions qui ne sont pas anaphoriques tout en étant coréférentielles. C'est le cas où on fait référence à la même personne dans plusieurs documents et discours (par exemple, toutes les références à Benjamin Malaussène dans les romans de D. Pennac), ce qu'on définit comme *cross-document coreference*. Ces références ne sont pas anaphoriques car chaque document constitue un univers discursif en soi (Poesio, 2016).

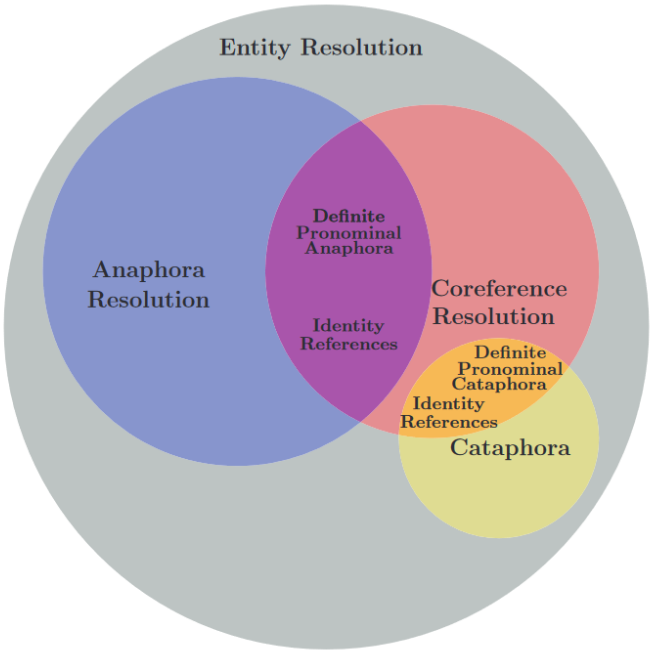


Figure 1. Illustration de la résolution d'entités. Tiré de Sukthanker *et al*., 2020

Ces deux phénomènes, tout en se chevauchant partiellement (comme illustré dans la figure 1), ne renvoient pas tout à fait à la même tâche du point de vue informatique et linguistique et présentent des difficultés dans leur définition surtout en TAL.

Par la suite, nous utiliserons de manière indistincte les termes de *coréférence* ou de *chaîne de coréférence* pour deux raisons principales. La première est que l'on considère que chaque texte de notre corpus constitue un univers discursif fictionnel en soi où les personnages mentionnés existent en tant qu'acteurs de l'histoire racontée. La deuxième est notre intérêt pour l'étude de la constitution des réseaux de mentions dans des textes produits par des

enfants au niveau textuel, plutôt que de réduire notre recherche à l'étude des paires mentions-antécédents. Nous allons aussi désigner de temps en temps les chaînes en tant que *clusters*, en prenant inspiration des nombreux travaux et outils de TAL anglophones qui décrivent les ensembles de mentions référées à un même objet en utilisant ce terme.

Des types différents de mentions peuvent constituer les chaînes de coréférence, et nous allons principalement faire référence à la taxonomie de Charolles (2002) pour les définir.

1.2. Les mentions ou maillons de la chaîne de coréférence

Pour Charolles (1988b), les éléments qui constituent une chaîne de coréférences sont ces expressions nominales ou pronominales qui permettent « d'identifier un individu (un objet de discours) quelle que soit sa forme d'existence (personne humaine, événement, entité abstraite) », et qui peuvent être appelées, par métaphore, *maillons* de la chaîne de référence (que nous allons appeler *mentions*, en suivant la terminologie anglophone).

Dans une chaîne, nous retrouvons des types de mentions différentes, soit par partie du discours soit par représentation de surface de l'individu mentionné. Ce premier critère de catégorie grammaticale a été récemment utilisé par Federzoni *et al.* (2021) dans leur travail sur la catégorisation des chaînes de coréférence pour décrire les différentes catégories de mentions que l'on peut détecter dans un corpus donné. Dans leur cas, 8 catégories de mentions ont été reconnues, catégories qui se recoupent avec celles que nous allons présenter par la suite, en faisant référence en particulier aux données issues du corpus analysé.

Dans notre corpus nous avons pu identifier les types de mentions suivantes, ce qui revient à une taxonomie complète des catégories de mentions possibles :

- ***Les noms propres*** : ils indiquent un individu précis et habituellement précisé dans l'univers du texte analysé. Cette précision peut se réaliser par simple juxtaposition du type de personnage au nom propre cité, comme dans l'exemple (2a), ou à travers une formule de présentation, comme dans l'exemple (2b). Cette formule est habituellement composée du syntagme nominal qui fait référence à un des quatre personnages et du verbe réfléchi « s'appeler ».

> (2) a. Il était une fois une sorcière. elle avait un chat magique, le chat avait le pouvoir de changer de couleur. La maison de la sorcière Gigi était noire donc vu que le chat Victor pouvait changer de couleur, Gigi tomba mais un jour elle a dit « j'en ai marre oust. » (…) [3]
>
> (2) b. Maitre chat a un robot, il nettoie tout, même le grenier. Le robot s'appela Bobis. (…) [4]

- ***Les expressions nominales définies***, qui peuvent être divisées en deux catégories, ***complètes*** ou ***incomplètes*** : ce sont des syntagmes nominaux introduits par des articles définis. Dans le cas de ce type de mention, nous pouvons relier l'expression à son référent sur la base du contexte d'énonciation. La distinction fondamentale entre expressions complètes et incomplètes est que les premières sont référentiellement autonomes, c'est-à-dire qu'elles « ne sont valides que pour un seul référent ou un ensemble de référents » (Charolles, 2002 : 75), et donc elles peuvent identifier un référent unique, alors que les deuxièmes peuvent s'appliquer « à un nombre indéfini d'êtres ou de groupes d'êtres particuliers » (Charolles, 2002 : 75-76). Dans l'exemple (3), la mention (a) peut être considérée comme une expression nominale incomplète, car elle fait référence à un groupe d'êtres, alors que la mention (b) est une expression

[3] Production NORM-EC-CE2-2016-102-D1-S212
[4] Production NORM-EC-CE2-2016-102-D1-S214

nominale complète car elle est liée à un référent spécifique, cité plus tard dans le texte.

(3) Un jour une navette spatiale s'écrase à Chatville, le pays des chats(a) quand descendit un drôle de bonhomme fait en un métal et en clous(b).[5]

- ***Les expressions nominales démonstratives*** : l'utilisation du syntagme démonstratif fait toujours référence à une entité déjà introduite dans le discours et il est perçu comme expression focalisante, donc qui détermine une entité en particulier. Dans l'exemple (4), l'entité introduite dans le discours à travers une expression nominale, introduite par une préposition ; (a) est ensuite reprise sous forme d'expression nominale démonstrative (b) :

(4) Un jour la sorcière le transforma en petite fille(a), et la sorcière trouva cette petite fille(b) très gentille et très mignonne.[6]

- ***Les expressions nominales indéfinies***, toujours introduites par un article indéfini. Ces expressions peuvent être spécifiques ou génériques : spécifiques quand elles servent à introduire un nouveau référent dans le discours, comme dans l'exemple (5a); génériques, comme dans l'exemple (5b).

(5) a. C'est l'histoire d'un chat très malin qui vit dans un château (…)[7]

(5) b. Le chat est un animal de compagnie, très diffusé dans le monde.

Il ne faut pas oublier qu'une mention peut être aussi constituée par un nom seul, pas forcément précédé d'un article.

- ***Les pronoms personnels, sujet ou objet*** : ils jouent un rôle syntaxique bien défini car ils sont liés au contexte d'énonciation et nécessitent la présence d'un référent explicite pour qu'on puisse leur attribuer un rôle sémantique (voir exemple (6)). Cependant, en français l'existence des pronoms impersonnels,

[5] Production NORM-EC-CE2-2016-6-D1-S2000
[6] Production NORM-EC-CE2-2016-96-D1-S1931
[7] Production NORM-EC-CE2-2016-17-D1-S584

vides de sens, peut être problématique pour un système de détection des chaînes de coréférence. Ce pronom est très présent dans le corpus analysé dans la formule d'incipit des productions écrites comme dans l'exemple (7)

> (6) La sorcière et sa chatte font une potion magique avec des plantes et des serpents et des myrtilles tellement elles étaient les plus gentilles et des fois étaient des fois méchantes.[8]
>
> (7) Il était une fois un chat qui se promenait dans la forêt et qui ramassait des fleurs.[9]

D'autres éléments peuvent manifester le référent, ce qui nous pousse à les inclure dans cette taxonomie des possibles mentions du référent : ce sont ***le pronom réfléchi*** et ***le déterminant possessif***. Dans certaines constructions, nous pouvons considérer que le pronom réfléchi est coréférent du sujet de la phrase où il est contenu, comme dans l'exemple (8).

> (8) Il était une fois une sorcière qui préparait une potion magique. il survint un homme$_{(a)}$ qui se transformait en chat(b).[10]

Ce dernier exemple contient aussi un exemple de *référent évolutif* (Charolles, 2001) : ces référents peuvent être fréquents dans des histoires fictives, dont les personnages font souvent face à des transformations. Tout en essayant de prendre en considération la coréférentialité présente dans l'exemple (8) de l'homme (a) et du chat (b) en lequel il est transformé par la sorcière, dans ce travail nous ne prenons pas volontairement en considération les référents évolutifs, pour des problématiques liées à l'automatisation du processus de détection de ces mentions.

[8] Production NORM-EC-CE2-2016-17-D1-S576
[9] Production NORM-EC-CE2-2016-130-D1-S2531
[10] Production NORM-EC-CE2-2016-19-D1-S3051

Le déterminant possessif peut aussi être pris en considération comme faisant partie de la chaîne de coréférence du sujet possesseur/agent, comme dans l'exemple (9).

> (9) Il y avait une fois une sorcière appelée Camille qui vivait heureuse avec son chat mignon, qui s'ennuyait.[11]

Dans cet exemple, la mention « son chat mignon » introduit un nouveau référent dans le discours, le chat, mais contient aussi une référence au référent précédemment mentionné, « la sorcière Camille ».

1.3. Les chaînes de coréférence : critères de description

Une chaîne de coréférence peut être décrite par plusieurs caractéristiques, comme celles proposées par Schnedecker (1997, 2005) et repris par Oberle (2017 : 17) :

a. La longueur de la chaîne, c'est-à-dire le nombre de maillons qui la composent ;

b. Sa portée : la chaîne analysée peut être soit locale (dans un paragraphe ou deux), soit globale (elle couvre tout le texte ou une bonne partie) ;

c. La distance moyenne entre maillons de la chaîne (combien de mots il y a entre un maillon et l'autre) ;

d. La catégorie et la fonction morphosyntaxique des maillons, notamment du premier maillon de la chaîne ;

e. Le patron de la chaîne : les séquences les plus présentes et catégories grammaticales des maillons ;

f. Le mode de cohabitation : « succession, entrecroisement, dérivation, partition, fusion, déroulement parallèle ».

Ces critères sont hautement variables selon le niveau de maîtrise du scripteur, le genre du texte analysé, sa longueur, et peuvent aussi servir de

[11] Production NORM-EC-CE2-2016-108-D1-S2978

lignes directrices pour établir les limites qu'on veut assigner aux chaînes analysées dans le présent travail.

D'autres études proposent deux critères supplémentaires qui sont la densité référentielle d'un texte et le coefficient de stabilité d'une chaîne. « La densité référentielle est calculée en divisant le nombre d'E[xpressions]R[éférentielles] par le nombre de mots du texte. » (Obry *et al.*, 2017 : 3) Ce critère peut être utilisé pour comparer la densité de mentions dans différents types de textes. Le deuxième critère, le coefficient de stabilité, était initialement créé pour mesurer la pauvreté lexicale dans les œuvres médiévales, mais s'est ensuite avéré très utile pour rendre compte des différents caractéristiques des textes modernes (Obry *et al.*, 2017 : 5).

> Le coefficient de stabilité renseigne sur la diversité des redénominations dans les C[haînes de]R[éférence] . Cette mesure s'inspire d'une étude de M. Perret (Perret, 2000) et s'obtient en divisant, pour un référent donné, le nombre total d'E[xpressions]R[éférentielles] nominales (G[roupes]N[ominaux] et noms propres) par le nombre de ses désignations nominales différentes. Pour être pleinement opératoire, la mesure du coefficient de stabilité doit neutraliser les variations formelles non pertinentes (…) et s'appuyer sur les lemmes, et non sur les formes des expressions référentielles nominales. Plus le coefficient de stabilité est fort, moins les désignations nominales sont variées et plus les CR peuvent sembler monotones.

Ces deux critères, appliqués aux chaînes plus longues rencontrées dans des récits anciens, ont permis d'étudier l'évolution diachronique de la composition des chaînes de coréférence en français (Obry *et al.*, 2017). Cette étude avait aussi comme but de proposer une méthodologie d'analyse de ces phénomènes qui puisse rendre compte de l'évolution de la composition de ces chaînes, en intégrant d'autres critères si nécessaire. Cette méthodologie provenant de la littérature nous semble intéressante, pour ce qui concerne les

possibilités de comparaison en vue d'une analyse longitudinale et donc, d'une certaine manière, diachronique.

Plusieurs études (Schnedecker, 2005, 2014; Longo & Todirascu, 2009; Schnedecker & Longo, 2012) ont pu montrer que le genre du texte conditionne « le matériau linguistique contenu dans les chaînes de référence » (Schnedecker & Longo, 2012), ainsi que faire varier sensiblement les critères précédemment mentionné. L'étude de Schnedecker et Longo (2012) a pour objet d'analyse les faits divers, que les auteurs ont choisi, entre autres, pour leurs brièvetés : « (…) il s'agit de textes courts, ce qui permet de rendre compte de l'intégralité des CR. » et, d'autre part, le choix de ce genre est dicté par le fait que ses « caractéristiques thématiques et structurelles sont déjà bien établies » (Schnedecker & Longo, 2012).

Cependant, nous pouvons faire l'hypothèse que la compétence des auteurs des textes analysés puisse aussi conditionner les modalités de construction de ces chaînes et le type de mentions qu'on peut y retrouver plus souvent, dans le sens où l'on va retrouver dans les productions des élèves des structures plus simples que dans les écrits des scripteurs experts et une variété lexicale assez réduite dans les mentions utilisées. Comme dans le cas du projet *RésolCo,* que nous allons présenter par la suite (cf. Chapitre 1, 3.1), l'un des enjeux de notre travail est d'offrir une aide à l'annotation qui puisse aider à recenser tous les éléments qui constituent les chaînes de référence dans les écrits scolaires, et en perspective, de pouvoir utiliser ces instruments pour élargir ce travail à la description de l'évolution dans l'appropriation des marques de cohésion tout au long du processus d'apprentissage.

C'est à partir de cette hypothèse que nous nous fixons l'objectif de concevoir des outils de Traitement automatique des langues d'appui à l'étude de la coréférence tout au long du parcours d'apprentissage de l'écriture à l'école primaire. Ces outils vont principalement cibler, dans un premier temps, le

corpus *Scoledit*, que nous allons présenter plus en détail dans le chapitre 2 de ce travail.

2. La coréférence en TAL

Les chaînes de référence jouent un rôle majeur dans la construction de la cohérence des textes et leur identification est un enjeu majeur pour d'autres tâches du traitement automatique des langues, comme la traduction automatique, l'extraction d'informations ou encore la génération automatique de résumés.

La résolution des coréférences a une longue histoire dans le domaine du TAL, comme tâche autonome ou comme partie de chaînes de traitement plus larges (à partir des premiers travaux de Winograd (1972) sur la compréhension du langage). Depuis les années 1970, différents systèmes ont été développés en s'appuyant d'abord sur des approches symboliques à base de règles. Par la suite, cette approche a été largement remplacée par une approche neuronale au fur et à mesure de la diffusion et de la plus grande disponibilité des ressources langagières et techniques nécessaires à la mise en place de ces systèmes.

Le terme *coréférence* a été introduit dans le domaine du TAL comme partie de la campagne d'évaluation « SemEval » du MUC-6 (MUC Consortium, 1995b). Elle était définie de manière assez large (MUC Consortium, 1995a) :

> The basic criterion for linking two markables is whether they are coreferential. Whether they refer to the same object, set, activity, etc. It is not a requirement that one of the markables is 'semantically dependent'on the other, or is an anaphoric phrase.[12]

Où le terme *markables* était défini comme

[12] « Le critère de base pour lier deux marqueurs est de savoir s'ils sont coréférentiels: s'ils se réfèrent au même objet, ensemble, activité, etc. il n'est pas nécessaire que l'un des marqueurs soit "sémantiquement dépendant" de l'autre ou qu'il s'agisse d'une phrase anaphorique. » (notre traduction)

> The coreference relation will be marked between elements of the following categories: nouns, noun phrases, and pronouns. Elements of these categories are markables. [...] The relation is marked only between pairs of elements both of which are markables.[13] (MUC Consortium, 1995a)

Depuis l'introduction de ce premier modèle de définition de la tâche, assez libre et large, les chercheurs de ce domaine se sont interrogé sur la délimitation du phénomène, de façon qu'au moins trois modèles différents de coréférence se sont différenciés comme synthétisé par Grobol (2020, pp. 13-15) :

1. La résolution de coréférence, définie par la tâche du MUC Consortium, comme le processus de détection de chaque mention *m* dans un document, où cette mention répond à deux critères :

a. Elle appartient à une catégorie syntaxique déterminée (nom, syntagme nominale, pronom)

b. Elle est coréférente avec au moins une mention m' qu'on trouve précédemment dans le texte, son antécédent.

   La relation entre un antécédent et sa mention peut être formalisé comme lien direct $m' \leftarrow m$. Cette formulation aborde la coréférence comme relation d'équivalence afin de simplifier l'opération de détection des mentions de la chaîne.

2. Le modèle *mentions pair* ou *link-centric* : il réduit ultérieurement l'espace de recherche, en identifiant pour chaque mention dans le discours, son antécédent immédiatement précèdent, donc en modélisant des paires antécédent-mention en amont. (Recasens, 2010)

[13] « La relation de coréférence sera marquée entre les éléments des catégories suivantes : noms, syntagmes nominaux et pronoms. Les éléments de ces catégories sont marquables. [...] La relation est marquée uniquement entre des paires d'éléments qui sont tous deux marquables. » (notre traduction)

3. Le modèle *entity mention* : il modélise le lien entre les mentions et les entités correspondantes, comme l'approche utilisée pour le corpus OntoNotes (Hovy *et al.*, 2006) : dans cette approche, la mention est l'unité centrale et le lien pris en considération est le lien mention → entité.

Pour conclure, nous pouvons définir la résolution de coréférence en TAL comme la double tâche de détection des mentions tout au long du document et l'identification des chaînes de coréférence, dans une des trois manières alternatives que nous venons de présenter.

3. L'évaluation de la résolution de coréférence en TAL

Le choix d'une métrique d'évaluation d'une tâche donnée représente un des facteurs qui garantit la comparabilité d'un système ainsi que la reproductibilité des résultats obtenus. Cela constitue donc une étape importante pour la définition des benchmarks d'une tâche donnée. Cependant, dans le cas de la résolution de coréférence, ils existent plusieurs métriques : au moins cinq métriques ont été proposées en contrepartie de la première proposition, faite pour la tâche du MUC-6. En l'état actuel, il n'y a pas encore de consensus défini sur quelle est la métrique la plus adaptée à la tâche (Poesio *et al.*, 2016). Pour les citer brièvement, les métriques recensées sont :

- Le score MUC (Vilain *et al.*, 1995), qui a été la première métrique à être utilisée pour cette tâche. Elle a été conçue sur la base du modèle de la coréférence *link-based*. Le score MUC compte le nombre minimal de liens qu'il faut insérer ou éliminer lors qu'on compare le résultat d'un système à la référence établie (au *gold standard*) (Cai & Strube, 2010). Cette métrique ne prend pas en compte les *singletons*, c'est-à-dire les entités qui ne sont mentionnées qu'une fois dans le texte.

- Le score B^3 (Bagga & Baldwin, 1998), qui calcule la précision et le rappel pour chaque mention dans le document, avant d'utiliser ces deux

indices pour calculer la précision et le rappel global obtenus sur le corpus. Cette métrique, tout en prenant en considération les singletons, doit être modifiée pour inclure les mentions extraites par le système qui ne sont pas comprises dans le corpus de référence et les mentions du corpus que le système n'a pas détectées.[14]

- La famille de métriques CEAF (*Constrained Entity Alignement F-Measure*). Proposée par Luo (2005), elle représente la réponse à un défaut de B^3, qui prend en considération la même mention plusieurs fois dans le calcul de précision et rappel des mentions. Cette famille de métriques est fondée sur le meilleur alignement possible entre *reference entities* et *system entities*.

- Le score MELA (*Mention, Entity and Link Average Score*) (Denis & Baldridge, 2009) appelé aussi « CoNLL score » car choisi comme score d'évaluation lors des campagnes d'évaluation 2011 et 2012 de CoNLL. Cette métrique est une moyenne de MUC, B^3 et $CEAF_{e,}$ (Lion-Bouton *et al.*, 2020).

- Le score BLANC (*Bilateral Assessment of Noun-Phrase Coreference*, introduit par Recasens et Hovy (2011). Ce score est une variation de l'index de Rand, qui prend en considération les mentions isolées (les singletons), tout en gardant la validité du score en leur présence (ce qui fait « gonfler » les scores B^3 et CEAF). Il est doté d'une granularité plus fine qui permet de mieux discriminer entre systèmes (Recasens & Hovy, 2011).

Même si les chaînes peuvent indiquer n'importe quel type de référent, comme une personne, un événement, une organisation ou une période dans le temps, la plupart des auteurs en TAL se sont occupés de l'annotation de chaînes

[14] « Since B^3 's calculations are based on mentions, singletons are taken into account. However, a problematic issue arises when system mentions have to be dealt with: B^3 assumes the mentions in the key and in the response to be identical. Hence, B^3 has to be extended to deal with system mentions which are not in the key and key mentions not extracted by the system, so called twinless mentions (Stoyanov *et al.*, 2009). » (Cai & Strube, 2010)

à référent humain ou, au plus, des chaines à référents concrets (Oberle, 2017 : 15) car la détection d'objets abstraits peut s'avérer plus complexe. Dans ce travail, nous allons nous occuper principalement de l'annotation de chaînes à référent animés (car à cause des caractéristiques de notre corpus, nous ne pouvons pas nous limiter aux référents humains, car ils ne représentent pas les seuls acteurs/patientes dans les productions écrites de notre corpus), et nous allons restreindre ultérieurement le champ de recherche, dans le sens où nous allons concentrer nos efforts sur la détection des chaînes dont le référent principal est donné par les consignes qui dictent le contenu des productions écrites des élèves (cf. Chapitre 1, 3.2 pour la consigne citée ici). Pour cette raison, nous n'allons pas aborder dans ce travail la notion d'*entité nommée* qui est habituellement un des éléments fondamentaux de chaque système de résolution de coréférences : le choix de restreindre notre champ de recherche sera explicité plus en détail dans la deuxième partie de ce travail lors de la description du fonctionnement de l'outil développé dans le cadre de notre recherche (cf. Chapitre 6).

Ce choix de restreindre le champ de recherche des entités auxquelles les mentions font référence est aussi dû au corpus sur lequel nous allons effectuer nos analyses. Dans le chapitre suivant nous allons introduire le type de textes sur lesquels nous avons prévu d'effectuer nos recherches, c'est-à-dire les textes qui composent un corpus d'écrits scolaires. Ces textes ont la spécificité d'être éloignés de la norme, et donc ils présentent des difficultés pour leurs traitements automatiques.

Chapitre 2. Les corpus scolaires

Si l'on suit la définition donnée par C. Wolfarth (2019 : 45), les corpus dits scolaires sont des corpus qui rassemblent des textes produits par des élèves de l'enseignement primaire et secondaire. La caractéristique principale de ce type de corpus est que les auteurs de ces productions sont encore dans une phase d'apprentissage de l'écrit, ce qui implique que leurs productions sont pour la plupart éloignées de la norme établie dans la langue de référence.

Dans cette partie, nous allons passer en revue les différents corpus annotés qui existent dans le domaine de l'étude des écrits scolaires en français langue première, et nous allons aborder plus dans le détail deux corpus spécifiques : le corpus *RésolCo*, qui est le premier corpus scolaire annoté en continuité référentielle pour le français langue maternelle, et le corpus *Scoledit*, qui représente le cadre de notre travail de recherche. Nous allons aussi citer un corpus scolaire italien à cause de ses liens méthodologiques avec le projet contexte de notre travail, avant de proposer une conclusion sur le rôle des écrits scolaires dans la recherche en didactique.

1. L'accessibilité des corpus

Comme déjà observé dans un passé récent par M.-F. Elalouf et C. Boré (2007) même si des projets existent dans le domaine des corpus scolaires, les données restent encore difficiles d'accès et peu numérisées (Wolfarth, 2019 : 73; Ponton *et al.*, 2021).

Selon l'état des lieux, très exhaustif, effectué par C. Wolfarth dans sa thèse (2019 : 47-52), confirmé plus récemment par C. Doquet et C. Ponton (2021), les corpus scolaires en français se font toujours rares, et le seul changement est le partage de quatre des corpus faisant partie du projet E-Calm

(cf. Chapitre 2, 3.). Cependant, récupérer ces données ou les visualiser en ligne reste toutefois une opération non triviale, sauf dans le cas des corpus que nous allons présenter dans la suite. Sur les 22 corpus recensés pour le français L1 (Wolfarth, 2019 : 47-52), seulement quatre projets ont rendu leurs données publiques. Nous allons voir par la suite de quelle manière ces données ont été recueillies, dans quel but et comment elles sont mises à disposition en ligne. Toutefois, nous avons aussi pu remarquer que les étapes de transcription et d'annotation des corpus sont quasiment toujours effectuées ou du moins vérifiées manuellement, ce qui ralentit parfois le processus de traitement nécessaire à rendre un corpus public.

Depuis une dizaine d'années différents projets ont tenté de concevoir des corpus scolaires (Elalouf, 2005; Elalouf & Boré, 2007; Garcia-Debanc & Bonnemaison, 2014 ; Boré & Elalouf, 2017) mais la plupart du temps l'accès à ces corpus reste assez restreint sinon impossible. Le premier corpus à être diffusé dans une forme numériquement exploitable, celui de M.-F. Elalouf (2005), a permis d'ouvrir la voie aux corpus qui sont arrivés bien plus tard, associés à des outils numériques de traitement des données bien plus avancés qu'en 2005. Ce corpus était disponible sous forme de CD-Rom. La continuation de ce travail est représenté par le corpus *EMA*, désormais accessible en ligne[15] (Boré *et al.*, 2018), que nous allons maintenant brièvement présenter.

Depuis 2014, différents corpus d'écrits scolaires ont été collectés et mis à disposition, dont au moins quatre conçus dans le contexte du projet *E-Calm*[16] (financé par l'Agence Nationale de la Recherche et coordonné par Claire Doquet). Ces projets ont permis de récolter et rendre disponible en ligne en libre accès au moins quatre corpus d'écrits scolaires : le corpus *Scoledit*, qui

[15] Corpus *EMA*, écrits scolaires : https://www.ortolang.fr/market/corpora/ema-ecrits-scolaires-1 (consulté le 10/01/2022)
[16] Site du projet E-Calm : http://e-calm.huma-num.fr/le-projet/ (consulté le 10/01/2022)

représente le contexte de notre recherche, le corpus *RésolCo* (Garcia-Debanc *et al.*, 2017), que nous allons ensuite présenter plus en détail, le corpus *Ecriscol* (Doquet *et al.*, 2017; Doquet, 2020;) et le corpus *Littératie avancée* (Jacques & Rinck, 2017). Alors que les deux premiers corpus mentionnés ici ont été sollicités par les chercheurs, les corpus *Ecriscol* et *Littératie avancée* ont été collectés de manière écologique. Nous allons décrire ici plus en détail le corpus EMA ainsi que les corpus *RésolCo* et *Scoledit* qui sont intégrés au corpus *E-Calm.*

2. Le corpus EMA

Le corpus *EMA* (Boré *et al.*, 2018) a été développé par les chercheurs du laboratoire éponyme EMA de Cergy. Dans sa deuxième version, il est constitué de deux dossiers, l'un composé de textes argumentatifs, l'autre de textes narratifs liés à une tâche de lecture faite au préalable dans la classe concernée. Ce corpus est composé de textes collectés de manière écologique et les choix d'analyse de son sous-corpus de transcriptions suivent les mêmes choix que dans le cas du projet *Ecriscol*, dirigé par C. Doquet à Paris 3 au laboratoire Clesthia. Pour chaque production, en effet, sont disponibles son scan, sa transcription, son annotation et ses métadonnées.

Comme décrit par M.-L. Elalouf et S. Perrin (2019), l'utilité de ce type de corpus est à rechercher dans le besoin de « développer chez les enseignants une culture de l'écrit scolaire », à laquelle la recherche sur les corpus scolaires peut contribuer. En effet, la possibilité de faire une comparaison entre plusieurs écrits d'élèves peut permettre aux enseignants de se rendre compte des compétences mobilisées par les scripteurs à un certain niveau et pour une tâche prédéfinie, ce qui pourrait aussi mener à une amélioration des critères d'évaluation adoptés par les enseignants (Boré & Elalouf, 2017; Wolfarth *et al.*, 2018).

Cette approche du corpus au service de l'enseignement en plus de la recherche est la même que l'on retrouve dans les corpus du projet *E-Calm*, que nous allons décrire maintenant.

3. Le projet E-Calm et ses corpus

Initié en 2016, le projet *E-Calm* ambitionne le développement d'un large corpus scolaire et universitaire en langue française pour mener différentes études (orthographe, cohérence/cohésion des textes, interventions des enseignants). Parmi ses 4 sous-corpus composant le corpus *E-Calm*, deux contiennent des productions suscitées par la recherche :

- Le corpus *Scoledit* (Ponton, 2018), qui rassemble des textes narratifs récoltés de manière longitudinale du CP au CM2. Il a été conçu au sein du laboratoire LIDILEM de Grenoble. La méthodologie de ce corpus a été récemment appliquée à la collecte de textes en Italie et Espagne, dans le cadre de la constitution du corpus *Scolinter* (Ponton *et al.*, 2021).
- Le corpus *RésolCo* (Garcia-Debanc & Bonnemaison, 2014; Garcia-Debanc *et al.*, 2017, 2021), rassemble des textes collectés de l'école à l'université, avec une tâche qui présente une consigne de résolution de problèmes de cohérence et cohésion textuelles.

Les deux autres corpus contiennent des productions écologiques, c'est-à-dire des textes sollicités par les enseignants, à partir de tâches « conçues et mises en œuvre par l'enseignant de la classe au moment où elles doivent prendre place dans la programmation prévue » (Boré & Elalouf, 2017 : 8) :

- Le corpus *Ecriscol* (Doquet, 2020) rassemble des textes scolaires du CE1 jusqu'à l'entrée à l'université, ainsi que les notes et brouillons relatifs à leur genèse.

- Le corpus *Littératie avancée* (Jacques & Rinck, 2017) rassemble des textes recueillis en licence et en master. Il a été conçu au sein du laboratoire LIDILEM de Grenoble.

Nous allons dans un premier temps décrire plus en détail le corpus *RésolCo* car il a été conçu pour aborder les phénomènes de coréférence, objets de cette étude. Le corpus *Scoledit* sera présenté dans un deuxième temps car il constitue le contexte de notre recherche.

3.1. Le corpus RésolCo

Le but principal du corpus *RésolCo* est celui de « décrire de façon exhaustive et systématique les formes linguistiques utilisées pour construire la référence dans les textes d'élèves (…). » (Garcia-Debanc *et al.*, 2021 : 100).

Les textes récoltés sont des productions écrites d'étudiants de différents niveaux de scolarité (du CE2 à l'université). Les productions sont issues d'une tâche de résolution de problème de cohésion textuelle, la même pour tous les niveaux scolaires. 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 ») sont proposées aux étudiants. La tâche impose d'insérer ces trois phrases dans un texte narratif fictionnel. Les trois phrases et la consigne proposée forcent le scripteur à insérer dans le récit un personnage « à partir d'une expression anaphorique imposée » (Garcia-Debanc *et al.*, 2021 : 101), à la différence des consignes habituellement proposées à l'école, dans lesquelles « le choix de l'identité référentielle du personnage est réalisé antérieurement à la mise en texte ». Voici la consigne proposée (Garcia-Debanc et al., 2021 : 101) :

> Raconte une histoire dans laquelle tu insèreras séparément et dans l'ordre donné les trois phrases suivantes :
>
> P1 - **Elle** habitait dans cette maison depuis longtemps.
> P2 - **Il** se retourna en entendant ce grand bruit.
> P3 - Depuis cette aventure, **les enfants** ne sortent plus la nuit.

Les trois phrases sont données aux étudiants déjà inscrites sur des bandelettes de papier qu'ils peuvent coller à l'endroit de leur choix dans leur texte manuscrit. Ce dispositif a été mis en place pour éviter la rédaction des trois phrases et pour laisser aux scripteurs la liberté d'écrire autant qu'ils souhaitent entre une phrase et une autre.

Figure 2. Corpus RésolCo : la page de navigation entre les textes et leur version normalisée

Un site en ligne[17] contient le corpus qui est disponible au téléchargement sous différents formats d'exploitation ainsi qu'en exploration libre à travers deux pages différentes: une page contenant la version transcrite du texte et l'autre une version normalisée d'un point de vue orthographique et annotée selon les référents principaux de la tâche (à savoir les référents reliés par le scripteur aux pronoms personnels et au syntagme nominal pluriel explicités dans les trois phrases qui constituent la consigne : « elle », « il » et « les enfants », indiqués en gras dans les phrases). Les deux pages contiennent

[17] Le corpus est disponible en ligne au lien suivant : http://redac.univ-tlse2.fr/corpus/resolco/

également le scan de la copie d'origine présenté à côté de la transcription ou de la version normalisée.

La méthodologie de travail des transcriptions prend en compte la génétique textuelle pour interpréter certaines erreurs : les ajouts, suppressions ou substitutions sont considérés dans la phase de transcription et annotés.

8/01/19 Premier jet de l'expression écrite III (5) Tout est inveter de A à Z	Île-de-France	2019-01-08	collège public en zone urbaine hors REP. Plus de 500 élèves	Rédaction
Bonjour, moi c'est Zack Ston et je vais intégrer le lycée Mavis Draguillon, c'est le meilleur lycée du pays. Bon je vous laisse je dois aller en ~~eou~~ classe. Quand je suis arrivé en classe le proffesseur me demanda de me présenter. Donc j'ai dût dire à la classe " Bonjour, je me nomme Zack Ston, j'ai 15 ans et je maitrîse la magie anti-dragon de feu. (et oui la magie existe dans notre monde ai~~s~~ [n]ssi que les dragons et les démons.)				

Figure 3. Un texte accompagné de sa transcription. Texte CO-3e-2018-FSBJC6-D1-R9

Dans la page qui reporte la version normalisée et annotée à côté du scan de la copie, des filtres permettent de mettre en évidence les mentions des chaînes de référence prévues par les phrases de la consigne (ainsi que les phrases elles-mêmes).

Comme spécifié dans la page de présentation du site du corpus[18], chaque texte est accompagné de plusieurs couches d'annotation : outre les traces du processus d'écriture déjà mentionnées, nous avons la normalisation orthographique, l'étiquetage morphosyntaxique, l'analyse syntaxique en dépendances et l'annotation de structures discursives (Garcia-Debanc *et al.*, 2017) comme la continuité référentielle, les unités de discours élémentaires, les

[18] http://redac.univ-tlse2.fr/corpus/resolco/presentation.html (consulté le 10/01/2022).

relations de discours et problèmes de cohérence. L'étape d'analyse syntaxique a été réalisée à l'aide de l'outil *Stanza* (Qi *et al.*, 2020) et l'annotation de la continuité référentielle a été faite manuellement grâce à l'interface d'annotation *Glozz* (Widlöcher & Mathet, 2012). Dans sa forme actuelle, le corpus contient au total 385 textes[19].

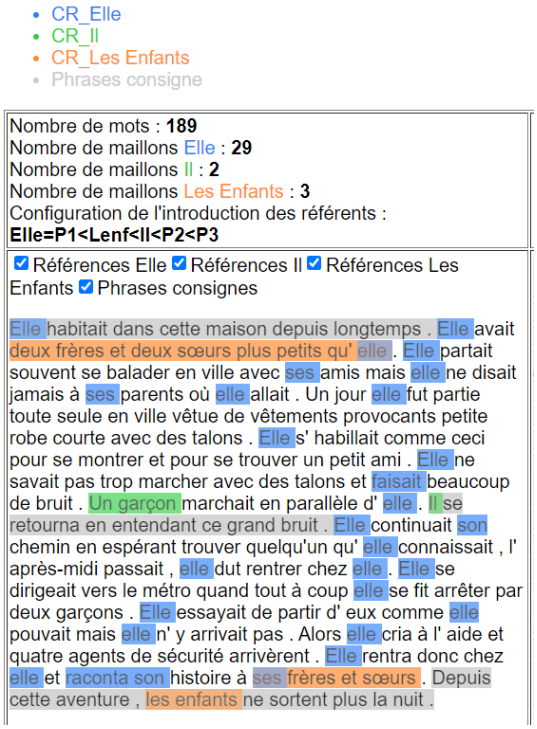


Figure 4. Version normalisée et annotée d'un texte aves les filtres possibles sélectionnés. Texte CO-3e-2016-VTAC305-D1-R5

Ces données sont aussi disponibles sous forme de deux différents fichiers téléchargeables : le *Corpus RésolCo*, qui contient la transcription des textes au format XML-TEI, les scans au format .png et les deux versions de texte, la copie d'origine et sa transcription au format .txt ; la ressource *RésolCo Continuité Référentielle*, qui contient les fichiers annotés en continuité référentielle au format .glozz. La ressource *RésolCo Cohérence*, contenant les

[19] Données tirées de la page de consultation du corpus http://redac.univ-tlse2.fr/corpus/resolco/telechargement.html (consulté le 10/01/2022).

textes segmentés en Unités de Discours élémentaires et annotées en relations de discours est encore à venir[20].

3.2. Le corpus Scoledit

Ce corpus représente le contexte de notre recherche et a été conçu par les chercheurs du laboratoire LIDILEM de l'Université Grenoble Alpes. Son but est de suivre de manière longitudinale le développement des compétences d'écriture des mêmes élèves, suivis du CP au CM2. Il représente actuellement un des seuls corpus longitudinal d'écrits scolaires entièrement transcrit et librement accessible en ligne[21] (Ponton *et al.*, 2021).

Les textes récoltés ont été sollicités par les chercheurs, sur la base de deux consignes de production : une consigne proposée en classe de CP et une deuxième utilisée pour les classes de CE1 à CM2.

Le recueil de CP a été effectué dans le cadre de la recherche *Lire-Ecrire au CP* (Goigoux *et al.*, 2015). La consigne consiste dans la présentation de quatre images à partir desquelles il est demandé aux élèves de raconter l'histoire du petit chat en 15mn ; ils pouvaient consulter les images à tout moment.

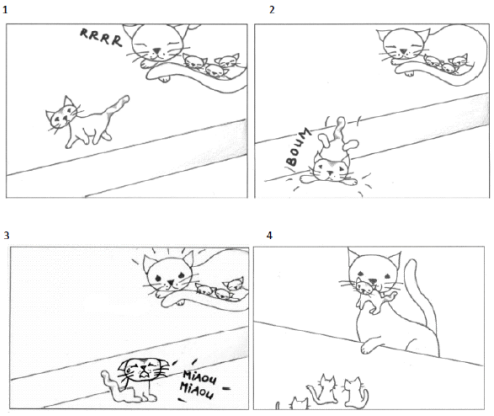


Figure 5. Images présentées aux élèves lors de la production écrite en CP.

[20] Informations contenues dans la page « Téléchargements » du site web du projet http://redac.univ-tlse2.fr/corpus/resolco/telechargement.html (consulté le 10/01/2022).
[21] Disponible sur son site dédié http://Scoledit.org/Scoledition/ (consulté le 10/01/2022).

La deuxième tâche proposait des images de quatre personnages : une sorcière, un chat, un robot et un loup. Chaque élève devait choisir un ou deux personnages, les indiquer sur sa feuille, et écrire une histoire comprenant ces personnages en 30 minutes. Les chercheurs conseillaient aux élèves d'écrire pendant 25 minutes de manière à avoir 5 minutes à disposition pour la relecture (Wolfarth, 2019).

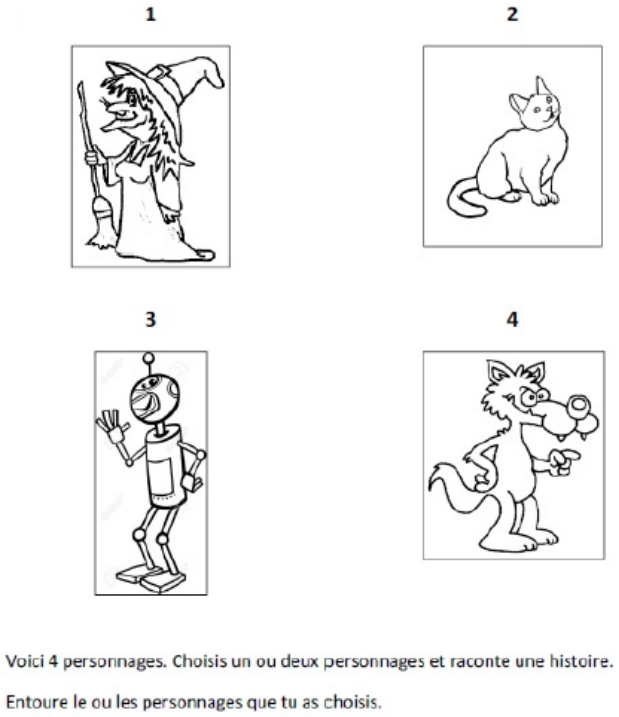


Figure 6. Images présentées aux élèves lors de la production écrite en CE1, CE2, CM1 et CM2.

Comme dans les expérimentations en psycholinguistique, les élèves écrivent leurs productions à partir d'images, pour éviter tout amorçage linguistique (Garcia-Debanc *et al.*, 2021). Un des intérêts de cette consigne pour notre recherche et que les images proposées limitent les identités référentielles des personnages insérés dans les textes.

En parallèle de cette constitution du corpus, l'équipe du LIDILEM a mené des recherches sur les possibilités d'exploitation de ces données à travers le TAL. Un des choix méthodologiques de construction de ce corpus a été, outre la phase de transcription, celle de mettre en place la normalisation manuelle des productions, afin de pouvoir exploiter de manière automatisée des textes

éloignés de la norme langagière. Cette phase de normalisation est ici définie comme la production d'un texte de comparaison, dont la réécriture le rapproche de ce qui était attendu, avec l'objectif de rendre « l'écrit plus standard et de se rapprocher de normes d'écriture » (Wolfarth, 2019 : 82)

Les transcriptions des productions écrites sont ensuite alignées de manière automatique avec les textes normés, à l'aide de l'outil *Aliscol* (Wolfarth, 2019). Cette phase d'alignement est effectuée en mettant en correspondance les segments transcrits avec les formes normalisées, sur la base d'un alignement graphique et phonologique. En parallèle, une étape d'étiquetage morphosyntaxique est réalisée sur les textes normalisés à l'aide l'outil TreeTagger[22]. Sur la base des comparaisons entre les textes produits par les élèves et les textes normalisés, il est possible de dégager des observations sur l'orthographe et sur d'autres niveaux de la langue, et le fait d'avoir à

Figure 7. Visualisation du scan et de la transcription d'une production sur le site du projet Scoledition

disposition un texte normalisé permet d'effectuer des analyses automatiques sur ce sous-corpus à l'aide du TAL (Wolfarth, 2019). Même si des tests de

[22] L'outil est disponible au lien suivant : https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (consulté le 01/02/2022).

normalisation semi-automatique ont été opérés, vu le « fort taux d'ambiguïtés générées » (Wolfarth, 2019 : 79), cette direction n'a pas été poursuivie.

Sur le site du projet, il est possible de consulter les productions des élèves, sous forme de scan, ainsi que leur transcription. Les versions normalisées des textes ne sont pas disponibles en ligne, mais seulement sur demande. Un tableau de consultation permet de visualiser les productions de chaque élève présentes sur la plateforme tout au long des années de scolarité.

Dans l'état actuel, le site *Scoledition*[23] héberge le corpus complet, qui contient 2.878 transcriptions récoltés dans 68 écoles[24]. Les contraintes dues à l'organisation de l'école, à l'implication des chercheurs et aux absences ou changements d'école des élèves ont eu pour conséquence la diminution des élèves participants : de 975 enfants initialement impliqués, seuls 373 ont pu effectivement produire un texte chaque année (Ponton *et al.*, 2021), ce qui a fait diminuer de manière considérable le nombre initialement prévu, au moment du recueil, de 7000 productions (Wolfarth *et al.*, 2018) à 2878 productions recueillies. Le corpus longitudinal, c'est-à-dire les productions qui ont été transcrites, normalisées et vérifiées manuellement, comporte actuellement 1820 textes ainsi distribués :

Niveau scolaire	Nb textes
CP	373
CE1	373
CE2	369
CM1	369
CM2	336

Tableau 1. Distribution des textes par niveau scolaire

[23] Disponible à ce lien http://scoledit.org/scoledition/ (consulté le 10/01/2022).
[24] Donnés tirés de la page du corpus http://scoledit.org/scoledition/corpus.php (consulté le 10/01/2022).

Sur le corpus total, restent donc 1058 textes à transcrire, normaliser et vérifier.

L'approche méthodologique de l'équipe du projet *Scoledit* a été reprise pour la conception de deux autres corpus récoltés respectivement en Italie et en Espagne dans le cadre du projet *Scolinter*. Ce projet, décrit par Ponton *et al.* (2021), est une collaboration entre l'Université Grenoble Alpes (France), l'Università di Milano Bicocca (Italie) et l'Universidad de Almeria (Espagne). Il prévoit d'effectuer une récolte similaire à celle du projet *Scoledit*, sur les mêmes tâches d'écriture, pour pouvoir mettre en comparaison le développement de l'écriture dans ces trois langues romanes.

La phase de transcription et de normalisation des deux corpus étant encore en cours, nous avons actuellement à disposition en libre accès seulement une petite partie des données récoltées[25].

Une première recherche a été effectuée sur les différences de segmentation en mots entre les trois corpus (Ponton *et al.*, 2021), ce qui a permis aussi de réaliser un premier essai de transfert, méthodologique et d'outils, entre le français et les deux autres langues.

Remarquons ici que, au moins en Italie, ce projet représente le premier cas de corpus longitudinal récolté sur les mêmes sujets dont les données et les transcriptions sont librement accessibles en ligne. Il existe toutefois d'autres exemples de corpus scolaires en Italie dont un en particulier que nous allons décrire par la suite, car il est lié méthodologiquement parlant au sous-projet Ecriscol appartenant au projet *E-Calm*.

[25] Le corpus est disponible ici http://*Scoledit*.org/scolinter/ (consulté le 10/01/2022).

4. Les corpus d'écrits scolaires en Italie : le corpus CoDiSSc

Plusieurs corpus existent pour l'italien langue étrangère, surtout pour l'italien oral, (nous citons les corpus DILS de Savy *et al.* (2012), et le LIPS de Vedovelli *et al.*, entre autres). Néanmoins, en l'état actuel, très peu de données ont été récoltées et mises à disposition pour ce qui concerne l'apprentissage de l'écrit des enfants de langue maternelle italienne.

L'un des projets existant en Italie est le projet *CoDiSSc* (Revelli, 2011). Ce projet récolte plus de 1200 documents de types différents mais toujours liés au milieu scolaire, rédigés depuis l'Unité d'Italie (1863) à nos jours (Doquet *et al.*, 2021). Ce corpus récolte en majorité des cahiers d'école de sujets variés, mais présente aussi un petit « sous-corpus », beaucoup plus restreint, constitué de documents administratifs à côté des cahiers de textes des élèves, ainsi que des documents produits par les enseignants, comme des registres scolaires ou les relevés de notes des étudiants.

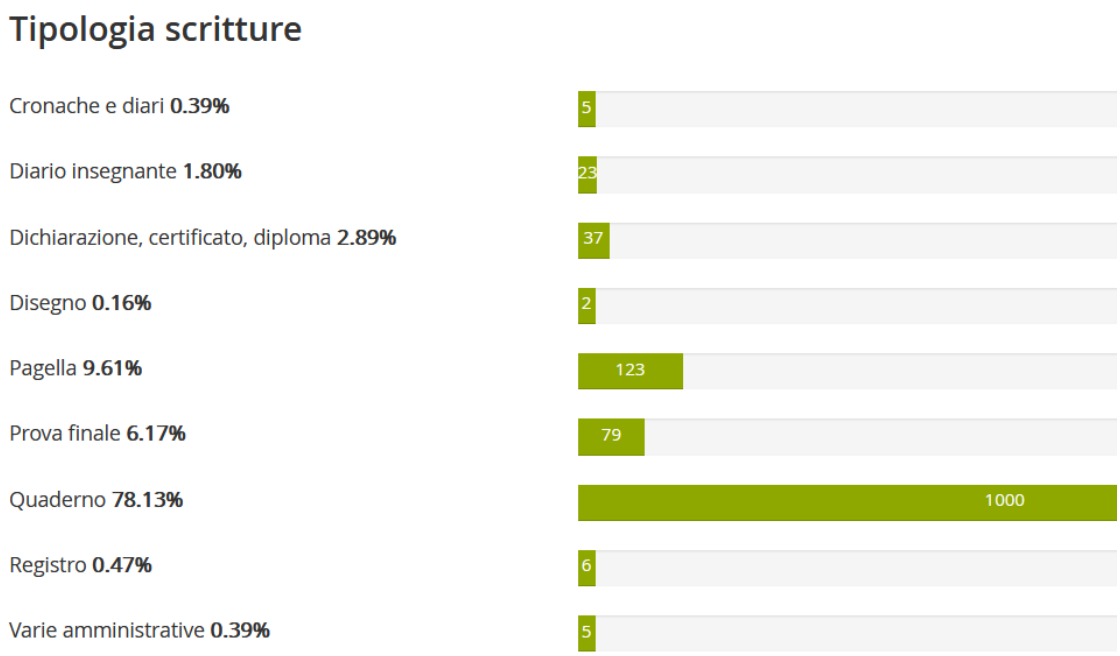


Figure 8. Typologies de textes du corpus CoDiSSc. Les cahiers (*Quaderni*) représentent la majorité du corpus. Graphique tiré de la page du projet http://www.codissc.it/consistenza-archivio (consulté le 19/01/2021).

Le but de ce corpus est d'exploiter les cahiers d'écoles récoltés en particulier en Vallée d'Aoste, mais aussi dans trois autres régions italiennes du Nord, afin d'étudier de manière diachronique les caractéristiques principales de « l'italien scolaire » et des modèles langagiers proposés par les enseignants tout au long de cette période (*Studi e ricerche del progetto CoDiSSc*). L'archive du projet dans son entièreté est visualisable en ligne, sous forme de scan. L'utilisateur peut ainsi naviguer entre différentes périodes et types de documents, qui peuvent être visualisés en plein écran et « feuilletés » numériquement. Des filtres permettent de sélectionner la période souhaitée, ainsi que, entre autres, la présence ou pas d'interventions des enseignants, le contenu du cahier, la région, ville et le niveau scolaire d'origine.

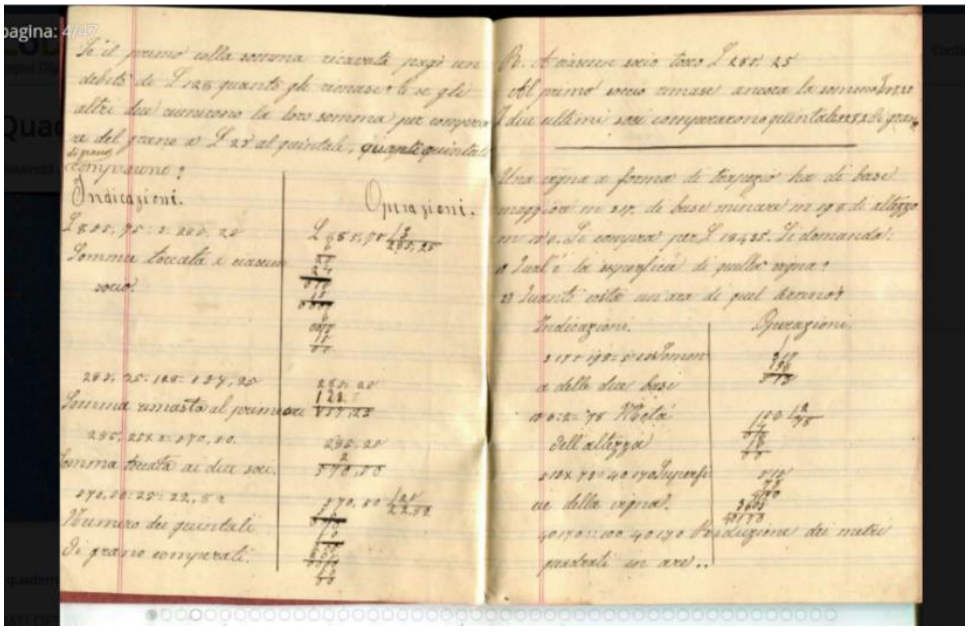


Figure 9. Visualisation d'un cahier d'arithmétique et géométrie de 1906.

Ce projet a adopté la méthodologie du projet *Ecriscol* pour ce qui concerne la transcription et l'annotation des textes (Doquet *et al.*, 2021). Cependant, les transcriptions et données enrichies ne semblent pas encore en libre accès au public. Cette initiative, néanmoins, permet de rendre compte du développement de l'écriture scolaire tout au long d'un siècle ; il en résulte un outil intéressant pour sonder l'évolution des pratiques d'enseignements de

l'italien et d'autres sujets (des cahiers de mathématique, géométrie, science etc. sont disponibles sur plusieurs années).

5. En guise de conclusion sur les corpus scolaires

Pour conclure ce chapitre, nous souhaitons rappeler la contribution fondamentale que les corpus scolaires ont dans la recherche sur le développement de l'écriture. Ces données « tracent des trajets développementaux de l'acquisition de la langue écrite et de ses usages » (Doquet & Ponton, 2021 : 11), en permettant de mettre au premier plan les obstacles persistants dans l'apprentissage et d'effectuer un véritable travail scientifique de modélisation des acquisitions. Cependant, ces corpus restent encore peu développés en nombre et surtout peu accessibles : comme recensé par C. Wolfarth (2019) et remarqué par C. Doquet et C. Ponton (2021), pour les corpus en français langue première, peu de ressources sont accessibles et exploitables pour des traitements automatiques à l'heure actuelle, et encore moins sous forme normalisée, même si trois des corpus faisant partie du projet E-Calm sont déjà à disposition du public et des chercheurs.

Cette pratique de partage des corpus n'est donc « pas encore courante, bien qu'elle soit, pour connaître la réalité des performances des élèves à l'écrit, tout à fait nécessaire. » (Doquet & Ponton, 2021 : 13). Pour combler cette insuffisance dans la quantité des données à disposition, une réponse possible est d'exploiter de manière transversale les ressources à disposition, en mutualisant les corpus existants, les méthodologies et les outils à disposition.

Si des études sur la continuité référentielle ont déjà été effectuées sur le corpus *RésolCo*, le corpus *Scoledit* n'a fait actuellement l'objet que de travaux sur l'orthographe. Dans ce travail, nous nous proposons d'apporter notre modeste contribution à la « cartographie de l'emploi des formes linguistiques qui déterminent la cohérence et la cohésion textuelles » (Garcia-Debanc *et al.*, 2021), en réalisant des études sur la construction de la coréférence sur le niveau

CE2 du corpus *Scoledit*, en appliquant certains choix méthodologiques déjà effectués au sein du corpus *RésolCo* dans notre analyse (cf. Chapitre 4, chapitre 5). Nous proposons d'apporter cette contribution à l'aide d'un corpus longitudinal, en partant d'un corpus de travail de niveau intermédiaire, car nous postulons de pouvoir tracer la variabilité de la composition de ces chaînes entre différents niveaux scolaires.

Chapitre 3. Approches pour l'annotation des coréférences

Une des problématiques à aborder concernant les chaînes de référence est le choix des critères à adopter lorsqu'on doit les annoter à l'intérieur d'un corpus donné. Comme des genres de textes différents contiennent des chaînes avec des caractéristiques assez différentes, il est important de pouvoir déterminer ces critères sur la base des propriétés intrinsèques des textes analysés et des objectifs du corpus analysé.

En TAL, ces choix d'annotation peuvent être faits en raison des possibilités techniques d'un système ou sur la base de critères linguistiques choisis en amont, ou des buts d'utilisation prévus pour un corpus donné, ou encore du type de son contenu textuel et des modèles langagiers qui saisissent mieux le phénomène abordé.

Nous avons déjà mentionné lors du chapitre 1.2 l'existence de deux types d'approches possibles :

1. l'approche symbolique ou par règles (comme celle adoptée par Oberle (2017) initialement au sein du projet *Democrat* ou l'outil *ArkRef* (O'Connor & Heilman, 2013) que nous allons décrire plus en détail dans la suite) ou

2. l'approche neuronale, qui exploite les réseaux de neurones, comme dans le cas des travaux de L. Grobol, que nous n'aurons par la possibilité de discuter en détail dans ce travail, bien qu'ils représentent probablement les travaux les plus récents qu'ont été menés en résolution de la coréférence sur le français en utilisant des architectures de réseaux de neurones profonds (Dinarelli & Grobol, 2019; Grobol, 2020, 2021).

Ces différentes approches peuvent être vues comme complémentaires, car là où les approches par règles réussissent mieux en termes de précision,

celles par apprentissage ont des meilleures performances en rappel. Nous allons maintenant présenter ces deux types d'approches existantes en intelligence artificielle et dans le contexte du TAL avant de mentionner quelques-uns des outils qui existent pour chacune d'elles.

1. Approche par règles ou approche symbolique

Comment pouvons-nous décider si un système est intelligent ?

Depuis plusieurs décennies les chercheurs des différentes disciplines (mathématiques, sciences cognitives, informatique, philosophie…) s'interrogent sur les critères de définition de l'intelligence dans le cas d'un système artificiel. Newell (1990) définit de cette manière l'intelligence d'un système :

> A system is intelligent to the degree that it approximates a knowledge-level system[26].

La notion d'intelligence est limitée pour Newell au niveau des connaissances ou aux niveaux basés sur les connaissances d'un système. Une manière de pouvoir représenter des connaissances complexes, pour permettre à un système de les manipuler, est sous forme de symboles.

L'approche symbolique a été introduite par Newell et Simons (1976) et décrit l'intelligence artificielle comme le développement de modèles qui utilisent la manipulation de symboles, lisibles par l'humain, pour accomplir une tâche donnée. Ces symboles peuvent être représentés à l'intérieur de structures de données définies (listes, arbres, graphes etc.) qui définissent les liens existants entre les symboles. Ce type de systèmes a été souvent utilisé dans le passé et fonctionne au mieux sur des tâches bien définies pour lesquelles on a à disposition des informations claires. Par exemple, la partie d'échecs de *Deep Blue* contre Kasparov en 1997 est un exemple très connu d'utilisation d'un

[26] « Un système est intelligent au niveau où il peut se rapprocher à un système de connaissance » (Notre traduction).

système symbolique. Un des avantages importants de ce type de système est l'explicabilité des décisions prises par un système et la transparence des étapes du raisonnement, ce qui n'est pas le cas dans les systèmes de *machine learning* ou dans les systèmes à base de réseaux de neurones.

L'approche symbolique se concentre sur une définition large d'intelligence et de raisonnements abstraits mais elle peut présenter quelques désavantages. En effet, elle obtient des meilleurs résultats avec des problèmes statiques et elle peut rencontrer des difficultés à s'adapter à la volée à un problème dont les informations lui sont inconnues. De plus, elle est plus difficile à maintenir dans le temps, car elle nécessite d'une maintenance continue dans le cas des données qui évoluent. Les systèmes symboliques présentent en fait souvent des bases de connaissances écrites à la main comme bases de leur fonctionnement, et donc une variation de tâche, des données, ou de système de connaissance implique la nécessité de renouveler ces bases de connaissance. Le renouvellement de bases existantes ou la création de nouvelles ressources est un processus parfois onéreux à mettre en place. Pour conclure cette partie introductive, nous pouvons définir les modèles symboliques comme des modèles où les connaissances à la base du système sont formalisées sous forme de règles, écrites à travers des symboles compréhensibles par l'humain. Nous allons ici présenter deux systèmes symboliques développés pour la résolution de la coréférence : le système *ArkRef* (O'Connor & Heilman, 2013) et, pour le français, l'outil *ODACR* (Oberle, 2017).

1.1. Systèmes à base de règles pour la résolution de la coréférence : le logiciel ArkRef

ArkRef (O'Connor & Heilman, 2013) est un outil développé principalement entre 2009 et 2010. C'est un outil déterministe à base de règles, qui exploite les informations syntaxiques d'un analyseur en constituants et les informations sémantiques d'une composante de reconnaissance d'entités

(O'Connor & Heilman, 2013; *The ARKref Noun Phrase Coreference System*, s. d.). L'architecture de ce système est largement fondée sur la description d'un outil similaire décrit par Haghighi et Klein (2009). Ce système de résolution de coréférence avait constitué un point de comparaison important à l'époque de sa sortie ; il surpassait alors en performance tous les systèmes supervisés et la majorité des systèmes non supervisés proposés jusqu'à cette date. Le système utilisé était composé d'un module syntaxique qui extrait des paires mentions-antécédents sur la base de règles établies en amont et grâce aux informations d'un parseur. Après avoir éliminé certaines paires sur la base de contraintes déterministes, il effectuait des décisions de compatibilité sur la base d'informations sémantiques. La dernière opération consistait dans l'élimination des antécédents incompatibles et la sélection des antécédents restants de manière à minimiser la distance syntaxique entre antécédent et mention (Haghighi & Klein, 2009; Sukthanker *et al.*, 2020). Ce dernier critère de sélection a été montré comme plus efficace que la sélection d'antécédents sur la base de la simple distance de surface (O'Connor & Heilman, 2013).

Le fonctionnement *d'ArkRef* est très similaire à celui du logiciel conçu par Haghighi et Klein : après avoir détecté toutes les mentions possibles dans un texte donné, *ArkRef* sélectionne un antécédent possible pour chaque mention, s'il en y a, et effectue les opérations suivantes :

1. Il prend des décisions immédiates pour certains patterns, comme l'apposition ou la construction prédicative-nominale, dans le cas où le sujet et l'objet de la phrase sont reliés entre eux par le verbe *être* ;
2. Dans le cas de mentions pronominales, il filtre les mentions précédentes par typologie syntaxique ;
3. Pour les mentions nominales, il filtre les mentions précédentes sur la base de la forme de surface et sur des critères de compatibilité sémantique.

4. Il choisit le candidat avec la distance syntaxique minimale entre les candidats filtrés s'il en y a sinon, il ne choisit pas de candidat (O'Connor & Heilman, 2013).

1.2. Systèmes à base de règles pour la résolution de la coréférence : le logiciel ArkRef

L'outil *ODACR* (Oberle, 2017) fonctionne sur la base de règles symboliques et répond à un manque de systèmes disponibles librement pour l'annotation des coréférences en français. Né comme une réponse à l'outil *RefGen* (Longo, 2013), initialement utilisé par le projet *Democrat*, *ODACR* est le prototype d'un outil de libre utilisation qui tente d'intégrer de manière efficace des bases de connaissances. Il est un système de bout en bout à base de règles linguistiques dont l'objectif est d'obtenir de meilleures performances par rapport à RefGen et de résoudre le problème de l'accessibilité en le rendant libre de droits.

En effet, RefGen, le prototype développé par Longo (2013) dans le cadre d'une thèse CIFRE est un système à base de règles linguistiques qui se base sur la théorie de l'accessibilité d'Ariel (2000). Il n'est plus malheureusement disponible car son *codebase*[27] avait été développé par un des ingénieurs de l'entreprise qui finançait le projet de Longo, entreprise ayant déposé le bilan. En outre, ses performances calculées par Longo (2013), semblaient peu prometteuses avec un score de 55%.

En plus de proposer deux ressources linguistiques (un dictionnaire d'entités nommées associé à des règles pour l'identification des dates, nombres et noms de personnes ; un dictionnaire d'hyperonymes), il a été nécessaire d'établir des règles de correction pour l'analyseur syntaxique (l'analyseur par apprentissage automatique *Talismane*[28] dans le cas de ce projet), jusqu'à arriver

[27] Ce terme désigne l'ensemble du code source utilisé dans le développement d'un logiciel donné (« Codebase », 2022).
[28] Disponible à la page web suivante : https://github.com/joliciel-informatique/talismane

à la réalisation de l'algorithme qui réalise trois étapes de détection sur trois éléments différents : les anaphores, les anaphores liées et la coréférence entre noms.

Bien que prometteur, ce système n'a pas été capable de tenir la comparaison avec les systèmes à base de réseaux de neurones développés pour le français plus tard et que nous allons présenter dans le paragraphe suivant.

2. Approche neuronale

Les systèmes à base de réseaux de neurones constituent actuellement les systèmes les plus développés et utilisés en TAL. Que ce soit dans les outils *off the shelf*[29] comme *SpaCy* (cf. Chapitre 5, 4.2) ou *GPT-3*[30], ou sous forme de logiciel *user-friendly*[31] qu'on utilise dans la vie de tous les jours, les réseaux de neurones sont omniprésents. Ces systèmes se basent sur une architecture complexe formée par plusieurs couches de neurones interconnectés. Ces « neurones », qui sont une formalisation du neurone au sens neurobiologique, sont inspirés du perceptron de Rosenblatt (1957), un automate utilisé comme classifieur linéaire.

Les avantages de cette architecture sont : la capacité presque infinie d'apprentissage de résolution d'une tâche définie et la bonne capacité de généralisation de l'application du modèle appris sur des données jamais « vues » par le système. Parmi les désavantages de ces systèmes, nous pouvons énumérer entre autres : la nécessité d'une grande quantité de données pour pouvoir obtenir un modèle de langue qui soit exploitable car le système a besoin d'accéder à une masse de données importantes pour pouvoir en dégager des

[29] Un outil *off the shelf* est un outil diffusé, sur le marché ou gratuitement, prêt à être utilisé dès le début ; ils ont développé avec le but d'être utilisé par un public varié.
[30] GPT-3 (acronyme de *Generative Pre-trained Transformer 3*) est un des plus gros modèles de langage existant, sorti en 2020. Il a été depuis appliqué dans plusieurs tâches de TAL différentes. (« GPT-3 », 2022, p. 3)
[31] On qualifie d'*user friendly*, un outil ou logiciel dont l'utilisation est intuitive et simple pour son utilisateur.

lois et des patterns qui soient assez généraux et applicables sur de nouvelles données ; l'exigence d'avoir à disposition une donnée annotée dans le cas de l'apprentissage supervisé, ce qui se traduit par le besoin de ressources et de corpus annotés à l'aide des méthodes symboliques pour pouvoir apprendre à la machine la tâche souhaitée. Enfin, le raisonnement de ce type de système est appelé « boîte noire » car il est très difficile de comprendre les étapes de raisonnement réalisées par la machine lors de la prise de décision (même si plusieurs recherches se penchent en ce moment sur l'interprétabilité des réseaux de neurones profonds, recherches qui comprennent parfois des TAListes) (Fan *et al.*, 2021; Sun *et al.*, 2021; Zhang *et al.*, 2021).

Les premiers systèmes de ce type pour la résolution des coréférences ont été développés pour la langue anglaise avant d'être appliqués à d'autres langues suite à l'apparition de plus grands corpus dans ces langues. C'est le cas du français où les premiers grands corpus annotés en coréférence sont parus entre 2011 et 2020 (nous citons ici seulement *Annodis*, *Ancor* et *Democrat*, et nous allons revenir sur ce dernier par la suite).

Plusieurs outils *off the shelf* sont actuellement disponibles en TAL : certains d'entre eux se servent d'un système à base de réseaux de neurones sous-jacent, comme l'outil *neuralcoref*[32].

Ce package fait partie de l'univers *huggingface*[33] et il est une extension de *SpaCy* 2.1+. Son but est d'annoter et résoudre les clusters de coréférence à l'aide d'un réseau de neurones. Il a été originellement développé pour être intégré dans un agent de conversation interactionnel ; il a été ensuite publié en *open source* par l'équipe qui l'a conçu. Toutefois, il est actuellement disponible uniquement dans une version compatible avec des versions de *SpaCy* plus anciennes que l'actuelle. En revanche, ce package permet de naviguer de

[32] Disponible au lien suivant https://github.com/huggingface/neuralcoref
[33] Disponible au lien suivant : https://huggingface.co/

manière assez intuitive entre les chaînes de coréférence contenues dans un texte. Enfin, il peut facilement être paramétré par ses utilisateurs. En l'état actuel, il contient un modèle en langue anglaise, entraîné sur le corpus OntoNotes 5 Dataset (Weischedel, Ralph *et al.*, 2013) qui contient des textes journalistiques, des conversations téléphoniques, et des sites web. Néanmoins, ce package permet de détecter des chaînes de coréférence en anglais avec une bonne précision sur la base de nos observations.

Nous avons effectué des tests de ce package sur des séquences de textes narratifs d'enfants en anglais et il semble présenter des résultats assez prometteurs. Cependant, le système interrompt la chaîne analysée et en recommence une nouvelle à chaque expression nominale présente dans le texte, ce qui pourrait ne pas être intéressant dans le cas de l'analyse de textes narratifs où on veut détecter la chaîne dans l'entièreté du texte, comme dans le cas du corpus *RésolCo*. En outre, ce package est disponible publiquement seulement pour l'anglais : bien que l'entrainement du modèle sur une autre langue soit possible, cette opération ne semble pas triviale et demande d'avoir à disposition un corpus annoté en coréférence et qui puisse correspondre à la typologie de textes qu'on souhaite analyser.

En France, deux corpus principaux annotés en coréférences existent : le corpus *RésolCo* que nous avons déjà mentionné car il fait partie des corpus scolaires et le corpus *Democrat* (Landragin, 2021). Ce dernier a été objet de recherches liées aux systèmes d'apprentissage machine dont nous allons discuter dans la suite.

3. Le projet Democrat

Le projet « Description et Modélisation des Chaînes de Référence : outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement automatique » est un projet financé par l'Agence Nationale de la Recherche et commencé en 2016. Porté par quatre laboratoires (LaTTICe de

Paris, LiLPa de Strasbourg, ICAR et IHRIM à Lyon), il a été coordonné par Frédéric Landragin. Son but était de « développer les recherches sur la langue et la structuration textuelle du français via l'analyse détaillée et contrastive des chaînes de référence […] dans un corpus diachronique de textes écrits entre le 9e et le 21e siècle, avec des genres textuels variés. » (Oberle, 2017 : 13).

Entre autres, ce projet a créé le premier corpus de grande taille (« même s'il n'atteint pas le seuil symbolique du million de mots » (Landragin, 2021 : 13)) annoté en chaînes de référence librement disponible pour le français écrit : il contient environ 560 000 mots dont 198 000 expressions référentielles annotés et 20 000 chaînes de référence. Il regroupe 58 textes de différentes époques dont 26 textes narratifs (des débuts de roman ou des nouvelles en intégralité) et d'autres genres de textes variés (Landragin, 2021).

En s'appuyant sur un cadre théorique largement partagé en France (Corblin, 1985; Schnedecker, 1997, 2005, Charolles, 2002), les chercheurs du projet *Democrat* ont pu aussi proposer deux outils pour la détection des chaînes de coréférences qui exploitent les réseaux de neurones : un pour le français oral transcrit, l'outil *Decofre* (Grobol, 2020) et un deuxième pour le français contemporain écrit (Wilkens *et al.*, 2020). Les apports du TAL à l'étape d'annotation consistent dans l'élaboration de plusieurs outils « (…) élaborés pour faciliter la tâche d'annotation (annotation ‹assistée› par opposition à une annotation strictement manuelle), pour visualiser le corpus annoté, pour faciliter le traitement des annotations ou l'export vers d'autres outils d'analyses statistiques. » (Quignard *et al.*, 2021 : 5) Néanmoins, l'étape d'annotation a été réalisée entièrement à la main, sans aucune pré-annotation automatique en mention proposée aux annotateurs participant à la campagne.

Le choix méthodologique fondamental de ce projet, qui le démarque de projets existants similaires, est le fait d'avoir annoté pour la première fois en France les chaînes de référence dans leur entièreté : « le projet *ANNODIS* s'est

intéressé aux chaînes topicales plutôt qu'aux chaînes de référence (Federzoni, Ho-Dac & Rebeyrolle, 2020) ; le projet *ANCOR* s'est intéressé aux relations anaphoriques (Muzerelle *et al.*, 2013) » (Landragin, 2021 : 12).

Ce corpus ayant été développé dans une démarche TAL, il n'est pas étonnant de constater qu'il a été exploité dans ce contexte comme dans la recherche présenté par L. Grobol (2021) qui, déjà dans sa thèse, préconisait l'utilisation du ce corpus comme test pour son outil, entrainé sur la base du corpus *ANCOR*.

Cependant, nous postulons ici que, bien qu'assez consistant pour offrir une ressource d'entraînement fiable, ce corpus ne pourra pas être adapté au but de notre recherche à cause de sa taille (trop restreinte pour un corpus d'apprentissage machine) et pour le type d'écrits qu'il contient (scripteurs experts en opposition aux scripteurs apprenants de notre corpus *Scoledit*). Le corpus *Democrat* a été intégré dans un projet de large échelle d'harmonisation des corpus annotés en coréférences de langues différentes, le projet *UD coreferences*.

4. Le projet UD coreferences

Ce projet, initialement présenté en 2021, est un projet pilote porté par la Charles University de Prague. Le but était d'harmoniser des corpus existants annotés en coréférences de langues européennes différentes avec un schéma commun d'annotation basé sur les *Universal Dependencies*. Cette action, inspirée de l'initiative *Universal Anaphora* portée par Massimo Poesio, a comme but celui de rendre possible des expérimentations multilingues à large échelle dans le domaine de la résolution de coréférence (Nedoluzhko *et al.*, 2021).

17 corpus existants ont été harmonisés, dont 13 sous licence libre. Seuls ces 13 corpus ont été rendus publics alors que les 4 corpus originairement sous licence non libre ont été harmonisés mais non publiés.

Figure 10. Liste des corpus tiré de la vidéo "Coreference meets Universal Dependencicies [LingMon #174]

Le choix de s'appuyer sur la tradition des *Universal Dependencies* est doublement justifié : d'une part, par leur notoriété et les outils déjà existants pour l'annotation en UD ; d'autre part, il est aussi justifié par des raisons théoriques comme la fréquente correspondance entre mentions et unités syntaxiques et le fait que les coréférences se manifestent souvent à travers des relations syntaxiques entre autres (Nedoluzhko *et al.*, s. d.). Cette première tentative d'uniformisation de la donnée ouvre une piste importante pour tous ces projets multilingues qui souhaitent exploiter un schéma d'annotation commun et viable pour les langues concernées, en plus de répondre à l'apparent manque de grands corpus annotés sur certains domaines spécifiques (comme les corpus scolaires).

Par exemple, l'application de ce schéma commun sur plusieurs corpus scolaires du français pourrait nous permettre d'obtenir une quantité de données assez importante laissant envisager la construction d'un corpus de grande taille dans ce domaine, ce qui n'est pas encore le cas. De la même façon, l'existence d'un schéma unique pour plusieurs langues pourrait nous permettre d'aborder des corpus scolaires de langues romaines différentes avec la même approche méthodologique, au niveau des choix d'annotation.

Chapitre 4. La coréférence en didactique : une question de cohérence

Nous avons précédemment posé quelques bases théoriques sur la notion de coréférence. Après avoir effectué un état des lieux sur les corpus scolaires en français langue première et avoir précisé les différents projets existants en annotation de la coréférence, il est utile ici de rappeler certaines des idées qui existent en didactique du français langue maternelle par rapport à l'enseignement de la textualisation, dont la coréférence peut être vue comme une des composantes principales. En abordant ce sujet, nous voulions rappeler une citation de Charolles qui écrit à propos de la cohérence : « Il n'est pas sûr que l'on puisse définir précisément en quoi consiste la cohérence, ni non plus qu'il faille chercher à le faire » (2006 : 26). Nous allons donc ici limiter nos réflexions à certaines recherches et tentatives de description de la cohésion et de la cohérence textuelles dans le champ de la didactique du français langue première.

1. La cohérence textuelle à l'école

Parmi les attendus de la fin du cycle 2 de l'école élémentaire (qui comprend CP, CE1 et CE2) par rapport à l'écriture, est cité le fait de savoir « Rédiger un texte d'environ une demi-page, cohérent, organisé, ponctué, pertinent par rapport à la visée et au destinataire. » (MEN, 2015). Mais à notre connaissance, aucun référentiel dédié à la cohérence textuelle n'existe actuellement.[34] La cohérence et la cohésion textuelles constituent un pan pour lequel la définition des critères d'évaluation reste floue.

[34] Nous avons trouvé une tentative assez récente de synthèse de différents publications, académiques ou issues du ministère de l'Enseignement Supérieur du Québec, afin de créer une sorte de référentiel de support aux enseignants pour l'évaluation de la cohérence textuelle (Lefebvre, 2020). Ce document, le RECO, est

Entre tous les critères mobilisés par les enseignants en évaluation de la production écrite, la cohérence semble être la moins stable : ignorée pendant longtemps comme objet d'enseignement (Carbonneau & Préfontaine, 2005), elle se configure comme notion extrêmement brumeuse, difficile à décomposer et décrire de manière constante d'un texte à l'autre, ou d'un scripteur à l'autre. Elle peut parfois apparaitre comme une question de style d'écriture plus que relevant d'un critère objectif, et être assujettie aux goûts et perceptions individuelles des enseignants ; par exemple, Rondelli (2010) décrit la cohérence en s'appuyant sur deux principes, le texte et le sujet interprétant, de manière à prendre en considération la tension existant, en description de la cohérence entre réception/perception subjective du lecteur et caractéristiques objectives du texte (Rondelli, 2010).

La perception de la cohérence textuelle peut ainsi dépendre fortement du genre de texte abordé, comme déjà observé par plusieurs chercheurs (Schnedecker, 2005; Longo & Todirascu, 2009; Schnedecker & Longo, 2012; Obry *et al.*, 2017), ce qui rend encore plus instable et potentiellement injuste l'utilisation de ce critère dans l'évaluation des productions écrites de typologies différentes.

« La cohérence interprétative consiste en une redondance assurant une homogénéité sémantique. Le texte construit donc un effet de monde. Mais si les mots ouvrent des mondes, ils le font parfois sans suivre la linéarité du texte. » (Rondelli, 2010 : 18) Dans ce manque de linéarité, la capacité des lecteurs à reconstruire les péripéties des personnages d'un récit est donné aussi par la tension entre répétition et variation des mots qui désignent le personnage. Mais comment déterminer de manière objective le juste équilibre à suivre entre

---

disponible pour le téléchargement au lien suivant : https://correspo.ccdmd.qc.ca/wp-content/uploads/2020/11/reco_referentiel_d-evaluation_de_la_coherence_du_texte_f.pdf

ces deux critères pour obtenir un texte bien écrit ? Comment cette problématique est-elle abordée dans la recherche en didactique ?

2. La coréférence et la continuité référentielle dans la recherche sur les écrits scolaires

À cause de ce manque de descripteurs largement partagés, la recherche a commencé à s'intéresser de plus en plus à l'étude de la cohérence à travers les productions des scripteurs experts et/ou en phase d'apprentissage.

C'est le cas des recherches menées sur le corpus *RésolCo*, que nous avons déjà mentionné (cf. Chapitre 2, 3.1), dont le but est d'effectuer une « cartographie de l'emploi des formes linguistiques utilisées (…) pour assurer la cohérence et la cohésion des textes. » (Garcia-Debanc *et al.*, 2021). La consigne qui préside aux productions de ce corpus impose l'intégration des pronoms personnels sujets dans un récit (*il* et *elle*), tout en laissant le scripteur libre de relier ou pas le dernier référent, *les enfants*, à ces deux pronoms[35]. Cette tâche s'appuie fortement sur les travaux portant sur le rôle du personnage dans les récits (Tauveron, 1995), ainsi que sur les travaux qui ont mis en évidence les difficultés des élèves dans la gestion du risque de confusion entre personnages dans une tâche rédactionnelle (Charolles, 1988a). À travers les productions suscitées par cette tâche, l'objectif est d'effectuer une analyse systématique à large échelle sur les formes linguistiques employées lors du processus d'apprentissage pour créer la continuité référentielle dans des récits, pour ensuite inventorier ces formes et construire ainsi des indicateurs de progression scientifiquement fondés (Garcia-Debanc *et al.*, 2021).

Dans ce travail nous partageons le choix méthodologique fait pour annoter le corpus *RésolCo* à savoir celui de se concentrer sur les référents induits par la consigne du corpus *Scoledit* : ce choix nous permet de nous

[35] La consigne de la tâche citée est reportée à la page 33 de ce mémoire et en annexe.

concentrer sur les personnages qui jouent un rôle central dans les productions écrite du corpus. Cette question sera ultérieurement approfondie dans la partie méthodologique de notre travail (cf. Chapitre 5).

3. Vers une approche outillée de l'annotation de la coréférence

Outre la pénurie de données relevée par Garcia-Debanc *et al.* (2021), l'autre problématique que l'on peut dégager lorsqu'on aborde le traitement outillé des corpus scolaires est la difficulté de mettre en place des annotations sur une large échelle. Notre travail vise à initier des recherches sur la conception d'un outillage assistant les chercheurs dans cette tâche d'annotation et d'analyse de la coréférence dans des corpus scolaires.

Comme décrit par C. Doquet et C. Ponton (2021 : 12), l'observation du langage à travers les grands corpus permet de faire « émerger des régularités, dans les acquisitions comme dans les erreurs commises ». Ces observations sont vitales pour une démarche didactique informée et efficace car elles permettent de pointer les difficultés persistantes dans le système langagier, parce que « peu visibles à l'œil nu, parce que rares malgré tout » ou aussi parce qu'elles restent inobservées (Doquet & Ponton, 2021 : 12).

En ce sens, le traitement automatique du langage peut constituer un véritable atout car il permet de dégager des phénomènes et des régularités qui risquent d'échapper aux annotateurs humains. Cependant, cette « automatisation » reste un simple appui à l'annotation humaine permettant de gagner du temps lors de la phase d'annotation sans remplacer les annotateurs experts. C'est le sens de l'outil que nous avons conçu et qui est décrit dans les parties suivantes.

Partie 2

-

Modélisation de la coréférence dans le corpus *Scoledit* : méthodologie et hypothèses de travail

Chapitre 5. Méthodologie

Dans cette deuxième partie, nous allons illustrer les choix méthodologiques effectuées au préalable pour la conception de l'outil de résolution des coréférences dans les écrits scolaires, introduit dans le chapitre suivant (cf. Chapitre 6). Pour traiter de manière efficace notre corpus de départ, composé de 373 textes de niveau CE2 du corpus *Scoledit*, nous avons dû mettre en place plusieurs étapes préliminaires : en premier, nous avons déterminé le périmètre d'un corpus de travail sur lequel nous allons conduire nos analyses manuelles, qu'on ne peut pas conduire sur un corpus dans son entièreté. Dans ce corpus, nous avons besoin de retrouver des textes qui soient assez représentatifs du corpus lui-même et qui contiennent les phénomènes qu'on souhaite étudier. Dans notre cas, le choix du niveau scolaire nous assure dans une certaine mesure la présence des chaînes de coréférences, le niveau CE2 étant considéré comme intermédiaire au sein de l'équipe responsable pour le corpus Scoledit : il s'agit du niveau pour lequel les élèves sont censés être déjà capables d'écrire un court texte cohérent (cf. Chapitre 4).Nous verrons par la suite quels sont les attendus en termes de réalisation de la cohérence textuelle et donc de la coréférence par rapport au niveau choisi. Nous avons donc à disposition un corpus de travail initial, composé de 50 textes de niveau CE2.

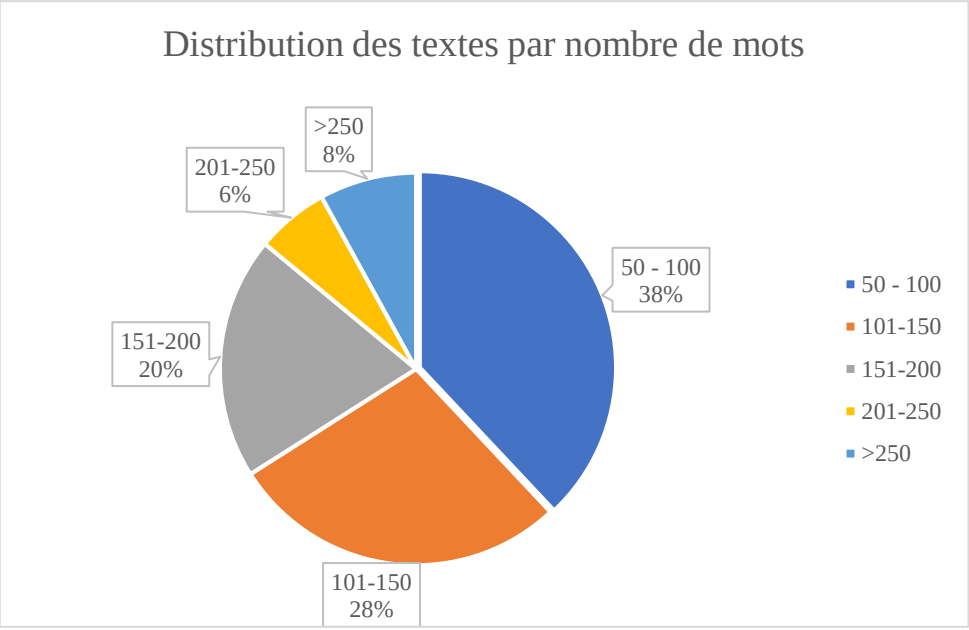
Après la constitution de notre corpus de travail, nous avons fait face à la nécessité d'effectuer des analyses morphosyntaxiques, afin de pouvoir effectuer une première analyse exploratoire manuelle, pour modéliser les différents éléments qui constituent des mentions référentielles au sein de notre corpus. Pour pouvoir réaliser ces étapes, nous avons dû dans un premier temps étudier la structure du corpus à notre disposition : en ayant choisi la version normalisée des textes du corpus *Scoledit*, nous avons dû effectuer une phase de prétraitement des textes, afin d'éliminer ou remplacer les balises d'annotation

qu'ils contiennent. Ces balises posent actuellement des problèmes dans le cadre d'une analyse morphosyntaxique automatique. Nous verrions par la suite la méthode de prétraitement que nous avons mise en place pour automatiser le processus de « renormalisation » des textes afin d'obtenir ensuite une analyse cohérente et le plus possible correcte.

Nous allons aussi présenter brièvement l'analyseur morphosyntaxique que nous avons choisi d'utiliser dans le cadre de ce projet, et les avantages qu'il présente. En dernier, nous allons illustrer certaines des analyses exploratoires conduites sur le corpus de travail et leurs résultats, avant de pouvoir passer à la modélisation de l'outil.

1. Choix du corpus

Comme déjà mentionné dans la partie 1, le corpus Scoledit est composé des scans des productions des élèves, des versions transcrites et des versions normalisées correspondantes. Notre travail a ciblé le traitement des versions normalisées des productions de CE2 du corpus longitudinal. Cela concerne 373 élèves, pour lesquels nous disposons des productions transcrites, normalisées et vérifiées. À l'intérieur de ces 373 textes, nous en avons choisis 50 au hasard, qui constituent notre corpus de travail principal. A l'intérieur de ce corpus de travail, nous avons ensuite sélectionnés plusieurs sous-corpus, composés d'environ 15 textes, que nous avons utilisé pour modéliser chacune des règles mises en place dans notre programme. La longueur des textes de notre corpus de travail varie entre les 50 et les 250 mots et plus, mais seulement le 34% des textes comportent plus de 151 mots, avec la grande majorité des textes qui comportent entre le 50 et les 150 mots.



Le choix du niveau de CE2 n'est pas anodin car il constitue le niveau scolaire intermédiaire du primaire. En CE2, la coréférence apparait déjà de manière plus substantielle tout en présentant encore de la variabilité dans la construction des chaînes et dans les mentions utilisées, ce qui présente des enjeux intéressants sur le plan didactique comme au niveau du traitement automatique. C'est enfin le deuxième niveau auquel les chercheurs ont soumis la consigne qui propose 4 personnages différents, ce qui nous fournit une plus grande variabilité dans les personnages décrits dans l'histoire par rapport à la consigne utilisé pour le CP, et donc une plus grande variabilité des mentions et des chaînes de référence. En outre, à ce niveau les élèves commencent à produire des textes plus longs par rapport aux niveaux précédents, ce qui leur permet de développer ultérieurement la coréférence au sein de leurs productions.

Le travail a été effectué sur les versions normalisées en posant aussi en perspective la possibilité de déporter les annotations en coréférences résultantes sur la version transcrite, « originelle » des textes, en s'appuyant sur l'alignement entre version transcrite et version normalisée, déjà mis en place au sein du projet *Scoledit* (Wolfarth, 2019; Wolfarth *et al.*, 2017).

2. Le rôle de la normalisation de l'écrit pour le traitement automatique et les corpus d'apprenants

Plusieurs corpus sont composés par des textes transcrits puis normalisés. C'est le cas, par exemple, des corpus d'apprenants cités dans la première partie de ce travail : (Elalouf & Boré, 2007; Garcia-Debanc & Bonnemaison, 2014b; Jacques & Rinck, 2017; Boré *et al.*, 2018; Doquet *et al.*, 2019, 2021; Doquet, 2020; Garcia-Debanc *et al.*, 2017, 2021). Les productions transcrites originelles présentent des formes non ou peu standardisées de langue et leur traitement automatique est fréquemment problématique (Wolfarth *et al.*, 2017). Les versions normalisées ont pour but de rendre possible l'exploitation de ces textes par des outils de traitement automatique, comme par exemple effectuer un étiquetage morphosyntaxique (opération impossible sur des textes trop fautifs). Dans le cas du corpus *Scoledit*, ces versions normalisées se proposent comme des « réécritures proches d'un attendu » des textes originaux (Wolfarth, 2019), et elles représentent le résultat d'une double opération de standardisation des formes non normées et d'annotation de la donnée manquante mais nécessaire à la compréhension et au traitement du texte. Dans le cadre de *Scoledit*, cette opération est régie par trois grands principes (Wolfarth, 2019 : 123) :

> a. Normaliser les productions au plus près de la production initiale de l'apprenant,
>
> b. Normaliser en considérant les phénomènes que l'on souhaite étudier,
>
> c. Normaliser en faisant appel le moins possible à l'interprétation (en cas de doute la primauté est donnée à l'oral).

Ces principes se traduisent notamment dans la mise en forme des textes normalisés dans le format standard XML-TEI par l'insertion de balises pour marquer certains phénomènes, comme l'omission d'une rupture forte entre deux propositions, l'omission d'une ponctuation faible marquant une rupture entre propositions ou encore la marque d'un discours direct dans la production. Une balise a aussi été introduite pour marquer l'omission de certains mots en

signalant la probable catégorie syntaxique d'appartenance du mot omis (ex. <omission type="nom">).

Ces choix de normalisation ne permettent pas l'utilisation directe des outils d'analyse syntaxique automatique existants. Nous avons dû pour cela mettre en place une phase préalable de prétraitements appelée « re normalisation ».

3. Structure du corpus et problématiques de prétraitement

Avant de poser les hypothèses qui ont guidé notre travail, il est nécessaire de décrire la phase d'exploration du corpus qui nous a permis de confirmer la faisabilité de construction de notre outil. Pour ce faire, nous avons exploré manuellement 50 textes issus de notre corpus de CE2.

Pour procéder, nous avons dû passer par une étape de prétraitement de notre corpus nécessaire à l'application d'une analyse. Nous allons présenter ici, à la suite, les différentes problématiques abordées lors de cette étape, comme le traitement des balises indiquant des ruptures discursives et les choix qu'ont été faits par rapport aux balises placées lorsqu'un mot était omis dans le texte par son auteur. Cette version « re-normalisée » des textes est surtout fonctionnelle pour l'analyse morphosyntaxique successivement appliquée par notre programme.

3.1. Phase de « re-normalisation » : choix opératifs

Dans la version normalisée des productions en format XML-TEI, les balises *omissions* (ponctuations ou mots) posent des problèmes pour l'analyse syntaxique automatique pour deux raisons principales. D'une part, les analyseurs actuels se fondent sur des suites de mots qu'ils étiquettent et ne peuvent pas prendre en compte ce type de balise. D'autre part, la suppression pure et simple de ces balises provoquerait une erreur de syntaxe du texte mettant en difficulté les analyseurs. Pour pallier ce problème, nous avons remplacé ces balises par des propositions « neutres » avant que les textes soit

soumis à la chaîne d'analyse[36]. Nous avons effectué ce processus sur 12 textes en total dans notre corpus de travail. Tous les choix effectués ont été faits sur la base du contexte du mot omis.

Par exemple, dans le cas des productions NORM-EC-CE2-2016-19-D1-S3051 (a) et NORM-EC-CE2-2016-104-D1-S831 (b), nous avons effectué les propositions suivantes dans le cas de l'omission d'un pronom :

> (1) Il était une fois un loup <**omission type**="pronom" **prop**="qui"/> vivait dans la forêt. Il mangeait les enfants qui n'étaient pas gentils.
>
> (2) Mais il se glaça d'horreur : ses pattes cédaient sous lui! Il était tombé à la cave! Et <**omission type**="pronom" **prop**="elle"/> était remplie de diamants!

3.2. Le traitement des balises de segmentation

Un des choix effectués par l'équipe du projet *Scoledit* était de garder dans la normalisation « que les éléments textuels produits par l'enfant : les mots, ou groupes de lettres, et la ponctuation » (Wolfarth, 2019 : 125) et donc de reproduire de manière le plus possible fidèle la mise en forme du texte originel, et de ne pas insérer dans la normalisation des marques de segmentation lorsqu'elles n'étaient pas présentes dans la production de l'élève. Toutefois, pour faciliter les analyses ultérieures, les chercheurs ont simplement marqué les phénomènes de segmentation du texte qu'ils soient implicites (mise en forme des titres par exemple) ou omis (signes de ponctuation par exemple).

L'omission des ponctuations a été annotée à l'aide de l'introduction de la balise <seg type="segm. forte"/> dans le cas de la présence évidente de rupture entre deux propositions, et de la balise <seg type="segm. faible"/> quand un signe de ponctuation est nécessaire mais qu'il n'y a pas de rupture forte, comme dans les exemples 1 et 2.

---

[36] Le pipeline d'analyse enchaine une série d'opérations comme la tokenisation ou l'étiquetage en parties du discours pour finalement proposer en sortie une analyse syntaxique du texte donné en entrée.

(3) je rentrai dans la maison **<seg type="segm. forte"/>** il y avait des potions de toutes les couleurs. Mais j'entendis du bruit et me cachai derrière la poubelle **<seg type="segm. forte"/>** elle était là **<seg type="segm. forte"/>**[37]

(4) Il était une fois une sorcière terrifiante à la voix lugubre, elle avait un chat noir comme une nuit sans étoile et gros **<seg type="segm. faible"/>** il était aussi très sale.[38]

Dans le cas de ces balises, nous avons remplacé automatiquement la balise de segmentation forte par le point et la balise de segmentation faible par la virgule. Ces choix, relativement neutres, permettent simplement d'éviter des erreurs d'interprétation aux analyseurs syntaxiques.

Un choix cohérent a été effectué pour ce qui concerne les balises qui marquent la présence d'un discours direct (<u> </u>), qui ont été substituées par les guillemets français « et », même en présence d'autres signes éventuels qui peuvent indiquer un dialogue, comme dans le cas du texte présenté dans la figure 11. Enfin dans la normalisation, la présence de titres ou de paragraphes est annotée respectivement par les balises <head> </head> et .


```
Filename:  "NORM-EC-CE2-2016-6-D1-S2000-V1.xml"
body:
  p_1:
    txt:  "Le robot martien"
  p_2:
    txt:  "Un jour une navette spatiale s'écrase à Chatville, le pays des chats  quand descendit un drôle de bonhomme fait en un métal et en clous. Bien sûr les humains savent qu'elle était un robot, mais les chats.... non. Alors le robot sortit de la navette. les chats restèrent muets mais quand il en descendit on pouvait entendre à des kilomètres :"
  p_3:
    txt:  "«  Va t'en! »"
  p_4:
    txt:  "«  Ce n'est pas ta planète! »"
  p_5:
    txt:  "«  Tu n'es pas comme nous, pars! »"
  p_6:
    txt:  "« File! »"
  p_7:
    txt:  "Alors le robot s'enfuit mais, dans le sens inverse de sa navette, mais vers la maison de l'inventeur. Il ouvrit la porte et là se dressait un autre robot. Alors tous deux devinrent amis mais quand l'inventeur chat se réveilla et il les chassait. Alors avec la navette il retournait sur la planète du robot où il était heureux."
  p_8:
    txt:  "Fin."
```


Figure 11. Exemple de sortie au format .json

[37] Production NORM-EC-CE2-2016-17-D1-S572
[38] Production NORM-EC-CE2-2016-37-D1-S1560

Comme nous tenions à conserver ces informations pour les analyses ultérieures, ces éléments de structuration du texte seront conservés sous forme de retour à la ligne dans les sorties de notre module, comme le montre la figure 11.

3.3. Le traitement des balises d'omission

Dans leurs productions, les élèves omettent parfois certaines formes, ce qui « rend la structure syntaxique non standard à l'écrit », et risque de faire échouer le traitement automatique subséquent (Wolfarth, 2019 : 126). Lorsque les éléments absents étaient identifiables de manière non ambiguë, ils ont été annotés dans la normalisation (comme par exemple la marque de négation *n'*) (Wolfarth, 2019 : 126). Si le choix du mot absent est ambigu et donc plusieurs alternatives sont possibles, une balise du type <omission type=" CATEGORIE " /> est inséré dans la normalisation, où CATEGORIE exprime « la catégorie du mot attendu (verbe, nom, pronom, adjectif, adverbe, préposition, déterminant). » (Wolfarth, 2019 : 126)

Ces balises n'étant pas facilement interprétables au niveau du traitement automatique, nous avons dû trouver une approche la plus neutre possible du point de vue du contenu de la production mais permettant l'analyse automatique. L'absence des mots omis faisait échouer l'analyse, nous avons pris l'option d'ajouter dans la balise omission un attribut proposition, "prop", avec comme valeur le mot le plus probable ; l'idée étant de garder la trace que le mot proposé ne fait pas partie à l'origine de la production. Le choix du mot a été fait selon le contexte dicté par chaque production écrite et en tenant compte de la catégorie du mot déjà indiquée dans l'annotation. Par exemple, dans le cas de l'exemple (10), le mot omis avait été identifié comme pronom, ce qui nous a amené à insérer le pronom le plus probable selon le contexte de la proposition.

(5) Il était une fois un chat qui se promenait dans la rue. Et il a vu un robot alors le chat va vers le robot et lui parla. Et <**omission type**="pronom" **prop**="il"/> lui dit <u> veux-tu jouer avec moi ? </u> dit le chat.[39]

Nous avons pu tester cette approche sur notre corpus de CE2. Ces insertions nous ont permis d'obtenir en sortie un étiquetage morphosyntaxique plus pertinent par rapport à celui obtenu en analysant le texte sans le mot de substitution. Ce choix a permis donc de repenser les choix méthodologiques d'annotation faits sur le corpus en entier, et de mettre en perspective la possibilité d'étendre ce type de modification non seulement au corpus ici utilisé mais au corpus *Scoledit* dans son entièreté. En outre, cette réflexion sera étendue aux autres langues actuellement dans le projet dès les premières phases de normalisation.

4. Module de pretraitement

4.1. Le module de prétraitement

Les textes normalisés du corpus *Scoledit* se présentent sous forme de textes au format XML contenant une production écrite par fichier accompagné des métadonnées « qui portent sur le recueil, sur les écoles et les classes participant au projet et les élèves », et « les consignes de recueil » soumises par les chercheurs (Wolfarth, 2019 : 103). Toutefois, nous n'avons pas exploité ces métadonnées lors de notre travail. La première étape du prétraitement consiste à extraire uniquement les textes des élèves depuis les fichiers XML d'entrée. Cette opération est largement simplifiée par le format XML-TEI des fichiers contenant les normalisations. Ensuite, nous nous sommes servi des annotations contenues dans les productions écrites pour constituer une nouvelle normalisation qui soit opérationnelle pour l'analyse automatique des textes. Le résultat de cette première étape de traitement nous a permis d'obtenir les productions écrites des élèves, identifiées par leur code unique, privées de leurs

[39] Production NORM-EC-CE2-2016-91-D1-S663

métadonnées et sous deux formats de fichiers distincts : format texte simple et format json.

Ce module de prétraitement a été développé en python en exploitant les fonctionnalités des librairies BeautifulSoup[40] et json[41].

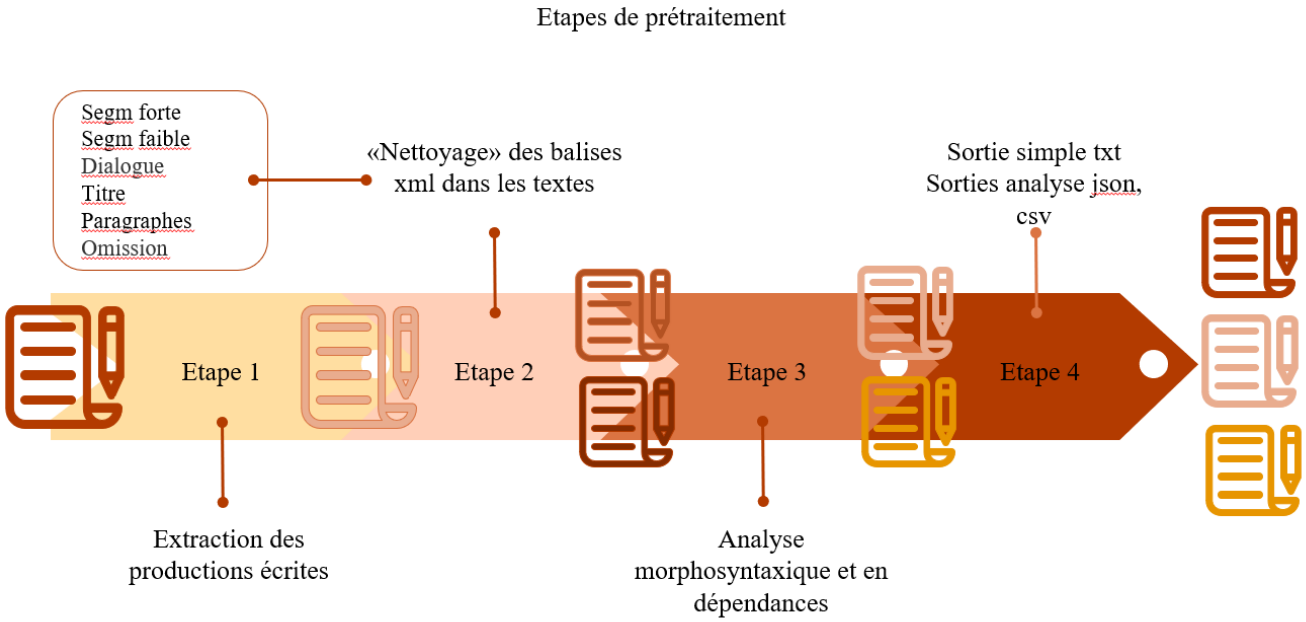


Figure 12. Schéma illustratif des modules de prétraitement et d'analyse morphosyntaxique initiale

4.2. Le choix de l'outil d'analyse morphosyntaxique : les points de force de SpaCy

Après avoir prétraité les textes issus du corpus, nous avons obtenus en sortie les deux versions txt et json du texte « propre » qui constitueront les entrées des étapes suivantes d'analyse de nos données. Pour réaliser ces étapes, et ensuite mettre en place notre outil de résolution des coréférences, nous avons

[40] Beautifulsoup est une librairie Python utilisée pour de l'extraction de la donnée depuis pages web (html) et fichiers xml. Elle est disponible au lien suivant https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[41] Le Javascript Object Notation (JSON) est un format standard utilisé pour représenter des données structurées, de manière similaire aux objets Javascript. Plus d'information sur ce format sont disponibles au lien suivant https://www.json.org/json-fr.html.

décidé d'exploiter les fonctionnalités de la bibliothèque Python *SpaCy*[42]. *SpaCy* est une bibliothèque disponible en libre utilisation, développée par explosion[43]. Elle peut traiter plus de 64 langues et elle est accompagnée d'une documentation extensive et approfondie. Autour de ce logiciel existe une grande communauté en ligne, très active sur Stack Overflow, ainsi que sur le repository officiel du logiciel hébergé sur GitHub. Sa communauté est composée d'utilisateurs et de développeurs qui contribuent en ajoutant et en maintenant des librairies, des plugins, des matériels d'étude et de documentation ainsi que d'autres outils autour des différentes tâches du traitement automatique des langues[44].

Actuellement arrivé à sa troisième version, c'est un logiciel qui permet de réaliser des pipelines de traitement dans des domaines variés et qui permet d'implémenter facilement plusieurs fonctionnalités comme la reconnaissance d'entités nommées ou la classification de textes (*Facts & Figures · SpaCy Usage Documentation*, s. d.). *SpaCy* réalise « sous le chapeau » les étapes de tokenisation, POS tagging, parsing etc.

Nous avons ainsi choisi d'exploiter cette libraire plutôt que d'autres disponibles (comme par exemple *stanza*[45]) pour une raison fondamentale : *SpaCy* est doté de *built-in* extrêmement utiles pour notre recherche, comme la fonctionnalité de *rule-based matching* et de *dependency matching*. La fonctionnalité de *rule-based matching*, et en particulier celle de *token matching*, nous permet de décrire une séquence de tokens sur la base de certaines de leurs caractéristiques (texte de surface, lemme, étiquette morphosyntaxique et d'analyse en dépendance, tête et dépendants syntaxiques), et de détecter

[42] La documentation du logiciel *SpaCy* est disponible au lien suivant https://*SpaCy*.io/.
[43] Explosion est une société informatique allemande spécialisée en développement de logiciels en intelligence artificielle et en traitement automatique des langues. La page officielle de la société est disponible au lien suivant https://explosion.ai/

[45] Disponible au lien suivant https://stanfordnlp.github.io/stanza/

automatiquement ces patterns à l'intérieur d'un texte donné. De manière similaire, la fonctionnalité de *dependency matcher* nous permet de détecter des séquences de tokens à l'intérieur de l'arbre en dépendance, sans la limite de la contiguïté des tokens recherchés. Nous avons largement exploité ces deux fonctionnalités pour construire la base de deux modules qui composent notre programme.

C'est aussi la présence de ces deux fonctionnalités dans *SpaCy* qui nous a fait envisager la possibilité de construire cet outil, et donc forcément le choix est tombé sur l'analyseur morphosyntaxique qui a fortement inspiré le début de ce travail.

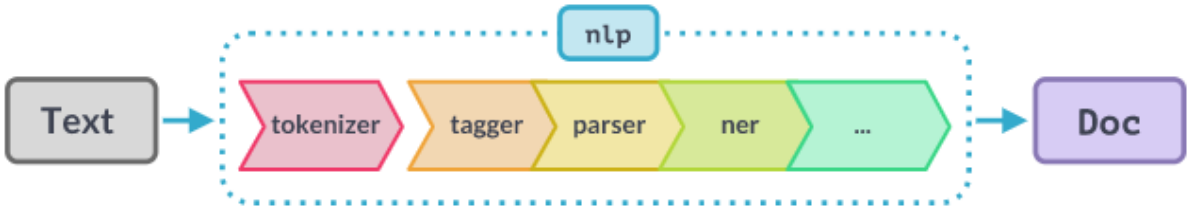


Figure 13. Illustration du pipeline de *SpaCy*, tiré de la page web https://*SpaCy*.io/usage/processing-pipelines

5. *Modélisation des mentions nominales et pronominales : bref état des lieux*

Comme discuté dans le chapitre 1, dans une chaîne de référence, nous pouvons retrouver des types de mentions différentes, soit par partie du discours, soit par représentation de surface de l'individu mentionné. Nous allons ici présenter les différents types de mention recensés dans les corpus de travail, sélectionnés pour construire des représentations le plus possible complètes des différentes structures internes des mentions contenues dans nos textes, en gardant en partie la taxonomie des mentions déjà citée dans ce travail (cf. Chapitre 1, 1.2). Précisons que notre démarche est empirique et incrémentale puisqu'elle s'appuie sur des exemples tirés de notre corpus sans hypothèse

particulière. Les observations faites sur ces exemples permettent de compléter le modèle.

5.1. Mentions nominales : noms communs

Une première étape exploratoire a ciblé l'étude des formes nominales contenues dans les productions écrites et extraites grâce à l'étape de POS-tagging des textes. Nous avons extrait une liste exhaustive des tokens étiquetés comme NOUN (nom commun) par *SpaCy* puis nous avons manuellement filtré les noms qui font référence aux quatre référents principaux de la production écrite. Nous avons pu observer peu de variations dans les formes nominales utilisées ; les plus fréquentes sont toujours les lemmes de base *sorcière*, *chat*, *loup* et *robot*. Voici leurs occurrences sur un total de 50 textes.

Mot	Documents où le mot est présent	Occurrences du mot
chat	40	192
sorcière	33	136
loup	16	67
robot	14	51

Tableau 2. Occurrences des mentions nominales dans le corpus de travail pour chacun des personnages.

Nous avons pu recenser au total 9 formes nominales faisant référence de manière non ambiguë aux 4 personnages, dont des synonymes, parfois présents à la fois dans les mêmes textes, ce qui témoigne d'un certain choix de variation lexicale par rapport aux entités principales déjà à ce niveau scolaire.

Voici la liste des mots que nous avons adopté en tant que lemmes de référence : *robot, loup, sorcière, magicienne, chat, chatte, chaton, minou, matou*.

Outre le fait qu'il est le personnage le plus fréquent dans les récits de notre corpus, le mot *chat* semble être celui qui présente le plus de variation : il est représenté à travers les formes *chat – chatte - chaton - minou – matou*. Le mot *sorcière* présente comme seule variation *magicienne*, et trois mots qui sont coréférents avec sorcière dans deux histoires différentes, *maitresse*, *vieille* et *dame*. Ces trois possibles coréférents apparaissent dans des histoires où la sorcière est présente avec le chat et joue le rôle de maîtresse de l'animal. Nous n'avons pas intégré dans notre liste de lemmes de références pour ces trois mots car ils sont peu fréquents et potentiellement ambigus, comme nous le verrons par la suite. Le mot *robot* présente essentiellement des variations de genre, comme dans le cas de l'occurrence *robote*, en contraste avec le masculin *robot*, au sein de la même production. On trouve également une périphrase assez élaborée : « *un drôle de bonhomme fait en un métal et en clous* »[46] L'hyperonyme *animal* est présent seulement une fois pour faire référence au chat. Enfin, le *loup* semble être le référent qui ne présente pas de variations lexicales dans le corpus, selon notre recherche manuelle et empirique.

Ces variations lexicales étant peu nombreuses dans le corpus par rapport aux occurrences des lemmes de base, nous avons donc décidé de sélectionner les groupes nominaux qui contiennent une mention explicite des personnages par lemme. Nous allons donc individuer les groupes qui contiennent des mentions des tokens dont le lemme est *chat*, *sorcière*, loup ou *robot*. Nous inclurons également quelques-uns des synonymes rencontrés dans notre liste et identifiables de manière non ambiguë comme faisant référence à une entité donnée (par exemple *chaton* ou *magicienne* ont été inclus mais pas *animal* car cet hyperonyme pourrait faire référence au chat comme au loup et donc

[46] Production NORM-EC-CE2-2016-6-D1-S2000

ambigu). Nous allons discuter plus en détail le fonctionnement du module chargé de cette opération dans le chapitre 6 (cf. 6.1).

Sur la base de la liste des formes nominales mentionnées précédemment, nous avons pu sélectionner un corpus de travail contenant des productions où l'on peut observer une certaine variation au niveau de la structure interne des mentions nominales. Sur ces productions nous avons ensuite conçu et testé le premier module de notre programme, celui chargé d'effectuer la détection des expressions nominales définies, indéfinies et démonstratives.

Par exemple, nous avons testé la fonctionnalité d'étiquetage des mentions sur le texte *NORM-EC-CE2-2016-108-D1-S2978*, car ce texte contient les deux personnages la sorcière et son chat ; il présente aussi une certaine variabilité dans les mentions utilisées.

> La sorcière et son chat
>
> Il y avait une fois une sorcière appelée Camille qui vivait heureuse avec son chat mignon, qui s'ennuyait. La petite magicienne eut un jour une visite imprévue qui parlait d'école pour sorcière, Camille très mécontente eut la mauvaise idée de répondre non. Le jeune homme très fâché, alors qu'il ne croyait pas avoir une réponse pareille préféra partir. Camille entièrement fâchée jetait un sort à son adorable minou qui doit maintenant supporter des poils bleus. Si la visite n'aurait pas eu lieu le pauvre mignon ne serait pas en colère contre sa maitresse. Trois jours plus tard la magicienne en bon état rendait sa couleur à son chaton (marron clair, les yeux bleus). Depuis ce jour le chat et la sorcière sont heureux. FIN.

Dans ce texte, en plus du passage de la mention indéfinie utilisée lors de la première parution du personnage, nous pouvons observer une certaine variation lexicale sur les deux personnages décrits, oscillant entre nom propre et expressions nominales définies, une fois le personnage introduit. En effet, sur la base de ce texte, nous avons adopté les différents lemmes *sorcière* et

magicienne pour les associer au cluster de *sorcière*, ainsi que décider d'inclure un token facultatif[47] dans le pattern à détecter, de manière telle à inclure la présence d'un adjectif éventuel entre l'article et le nom commun, comme dans la mention « la petite magicienne ». Sur la base de ce texte, nous avons aussi inclus les lemmes *minou* et *chaton* pour identifier le chat. Nous pouvons observer des mentions similaires dans le texte *NORM-EC-CE2-2016-96-D1-S1931*, où on peut aussi rencontrer des référents évolutifs (dans ce cas, « son chat qu'elle a transformé en fille »). Dans ce texte on peut aussi observer des occurrences d'expressions nominales démonstratives comme « ce chat », « ces enfants » et « cette petite fille » (que nous ne traitons pas car non référées de manière directe à un des quatre personnages dont nous délimitons la présence à travers notre analyse).

La sorcière et le chat

Il était une fois une sorcière, qui s'appelait Juliette. Elle n'aimait pas du tout ce nom, elle voulait s'appeler la sorcière de la rue Broca. Et elle avait un chat, elle ne l'aimait pas du tout, c'est pour ça qu'elle l'appelait Juliette. Ce chat n'aimait pas du tout ce nom et lui comparé à la sorcière, il aimait la sorcière. Un jour la sorcière le transforma en petite fille, et la sorcière trouva cette petite fille très gentille et très mignonne. La sorcière a appelé sa fille (son chat qu'elle a transformé en fille), la fille de la sorcière Broca et la sorcière acheta un nouveau chat et plein d'animaux pour sa fille. Mais sa fille voulait encore plus d'animaux mais tous ces animaux étaient méchants et la sorcière les transforma en petites filles, en petits garçons et en grandes filles et en grands garçons. Un chat apparait et la sorcière n'avait plus de magie et c'était un chat gentil donc la sorcière aimait ce chat et tous ces enfants aussi.

[47] La séquence de tokens cherchée dans le texte peut inclure un token dont on ne doit pas spécifier des caractéristiques, contrairement aux autres tokens dans ce type de module, où au moins une caractéristique entre texte de surface, lemme, POS tagging, étiquette d'analyse en dépendance, étiquette morphologique, doit être indiquée.

Dans cette étape nous avons aussi inclus la détection de l'article contracté comme dans le cas de l'occurrence « du chat » dans le texte *NORM-EC-CE2-2016-20-D1-S107.*

> Il était une fois une sorcière et son petit chat. Ils vivaient heureux et s'aimaient beaucoup. La sorcière nourrissait son chat et s'occupait de lui. Ils faisaient des jeux et s'amusaient à attraper des souris et des araignées. Mais un jour un méchant loup s'approcha du chat et... il sauta sur lui et le dévora en une bouchée. La sorcière était très triste mais un jour elle eut une idée. Avec sa magie elle allait retrouver le loup et lui jeta un sort! Grâce à ce sort le chat sortit du ventre et ils rentraient chez eux. Et ils étaient encore en train de jouer et la sorcière se sentit de nouveau heureuse. FIN.

La nécessité d'insérer la détection de l'adjectif antéposé au nom et postposé à l'article a été confirmée à plusieurs reprises dans notre corpus, où l'on trouve avec prépondérance la présence de l'adjectif « petit » dans cette position, référé aux lemmes de notre liste, comme dans le cas de la production *NORM-EC-CE2-2016-104-D1-S841*, où l'on trouve cette structure à la fois avec un article défini « le petit chat », « le petit chaton », et un démonstratif ; « ce petit chat ». Dans ce texte on trouve la raison pour laquelle nous n'avons pas inclus le mot *maîtresse* dans notre liste. En effet, même si on a plusieurs occurrences dans des textes différentes où *sorcière* et *maîtresse* sont coréférentiels, ici ces mots pointent clairement deux référents différents.

> C'est l'histoire d'un chat avec sa maitresse qu'il aimait beaucoup. Un jour ils se baladaient, le petit chat partit explorer les environs. il sautait, grimpait aux arbres et galopait si loin qu'il s'était perdu. à un moment donné il vit une grotte il rentra. tout d'un coup ce petit chat perdu vit une sorcière laide mais très gentille et elle lui dit " Que fais - tu là mon ami? ". Le chat répondit tremblant " Je me suis perdu. pouvez -vous m'aider s'il vous plait". "Oui bien sûr tu es si mignon". "Merci". Ils étaient devenus les meilleurs amis du monde. Ils cherchaient, cherchaient. le petit chaton perdu était désespéré. La sorcière lui dit " Ne t'inquiète pas

nous allons trouver mais tout de suite rentrons. il commence à faire nuit et à pleuvoir. Le lendemain ils cherchaient encore et encore, tout d'un coup le petit chat qui avait une bonne vue vit sa maison, par la fenêtre il voyait sa maitresse pleurer. il rentra et elle sauta de joie. la sorcière était triste alors on lui dit de venir vivre ici. Fin

Nous avons pris aussi en considération dans notre recherche les mentions nominales qu'on a appelées « simples ». Ce sont des mentions dont nous n'avons pas pris en considération la structure, comme dans le cas de « en petit minou méchant », ou des mentions précédées par un adjectif possessif, comme dans le cas de « son chat » dans le texte *NORM-EC-CE2-2016-108-D1-S2988*, « sa chatte » dans la production *NORM-EC-CE2-2016-17-D1-S576*, ou la mention simple « loup » dans le texte *NORM-EC-CE2-2016-98-D1-S815*.

Il était une fois une sorcière et son chat qui vivaient dans la ville de Pastise où il n'y avait que des sorciers et des sorcières. un jour le voisin de la vieille va transformer le chat en petit minou méchant et du coup le petit minou griffa la vieille. c'est devenu un gros désastre alors la petite dame fait une potion pour tuer son chat . elle attendait une semaine pour pouvoir tuer son chat . le minou fut mort et la dame fut triste mais elle est si vieille qu'elle meurt vieille et triste.[48]

La sorcière et sa chatte .

La sorcière et sa chatte font une potion magique avec des plantes et des serpents et des myrtilles tellement elles étaient les plus gentilles et des fois étaient des fois méchantes. La sorcière s'appelle Justine et sa chatte s'appelle Chillie.[49]

[48] Production NORM-EC-CE2-2016-108-D1-S2988
[49] Production NORM-EC-CE2-2016-17-D1-S576

> Un petit chat s'éloignait doucement de la ville. La nuit tombée, il décide de dormir dans la forêt. La pleine lune, les yeux du petit chat scintillèrent. Et d'un seul coup le petit chaton se transforma en un glouton de loup qui s'appela Nicolas François.[50]

Grâce à l'observation de ce corpus de travail, nous avons pu donc dégager trois « structures fondamentales » de mentions nominales présentes dans notre corpus de travail, composé par 6 textes :

- Une structure composée par : un article, défini, indéfini ou contracté ou préposition + un adjectif + un nom commun ;

- Une structure composée par : un article, défini, indéfini ou contracté ou préposition + un nom commun ;

- Un nom commun en isolement, auquel un déterminant possessif ou une préposition est probablement antéposé.

5.2. Mentions nominales : noms propres

Une autre étape de modélisation et d'exploration du corpus a concerné les noms propres présents dans les productions des élèves. Comme lors de l'exploration des noms communs, nous avons extrait une liste des tokens étiquetés comme noms propres et examiné sa composition à l'intérieur du corpus.

Dans le corpus de travail sélectionné pour cette première analyse sur les noms propres, nous avons pu observer deux structures fréquentes. La première est le nom commun qui suit un syntagme nominal, contenant un des lemmes de référence que nous avons mentionnés ci-dessus ; l'autre, le nom propre isolé.

Lors de l'étape précédente d'étude des formes nominales dans les textes, nous avions pu déjà observer que certains noms propres étaient erronément

[50] Production NORM-EC-CE2-2016-98-D1-S815

étiquetés comme noms communs (comme les noms propres *Sorcièla*, *Croquette*, *Jen-pouleloul* ou *Trololole*). Dans certains cas, très rares, un nom propre était reconnu en tant que tel dans une occurrence et étiqueté de manière erronée lors de sa deuxième occurrence au sein du même texte. Pour éviter des erreurs d'analyse due à ces problèmes d'étiquetage, que nous souhaitons plus tard étiqueter en tant que mentions, une ressource externe qui recueille ces tokens a été créée, en les associant à la catégorie correcte. Une fonction qui permet de forcer l'étiquette associée dans la ressource externe au token lors de son analyse a été intégrée dans notre programme, nous permettant ainsi de recatégoriser 44 noms propres. Nous avons utilisé cette même ressource pour forcer le POS-tagging de certains tokens fréquents et mal étiquetés, comme des onomatopées (*miaou*, *boum*, *plouf* entre autres).

5.3. Mentions pronominales : choix méthodologique

Nous avons décidé de restreindre le champ des mentions à la détection des quatre personnages qui peuvent représenter les protagonistes principaux des histoires décrites par les enfants : notre champ de recherche et de détection est d'un côté extrêmement réduit et, de l'autre, nécessite un effort supplémentaire lors de la détection des pronoms personnels présents dans les textes. Les pronoms étant très liés au contexte, ils représentent peut-être les mots les plus difficiles à désambiguïser. Dans ce programme, nous avions initialement détecté seulement les pronoms personnels, sujet et objet, de troisième personne du singulier avec l'objectif ensuite d'appliquer un filtre par genre pour les attribuer à leur chaîne de coréférence d'appartenance.

Nous avons consciemment décidé de sélectionner seulement la troisième personne car elle est la forme la plus fréquente dans le récit. Les formes de première et deuxième personnes sont plutôt présentes dans des contextes de dialogue direct ou parfois dans un récit à la première personne (cas extrêmement rare dans notre corpus de travail, composé de 50 textes). Les

pronoms de première et deuxième personnes, ayant souvent un rôle de déictiques dans les récits, s'avèrent être plus difficiles à traiter dans le cas d'un système pauvre en règles et en ressources. Nous avons donc fait le choix de ne pas traiter les extraits de textes représentant des dialogues directs et de ne pas inclure volontairement dans les mentions détectées la première et deuxième personne des pronoms personnels. Nous concentrerons donc nos efforts sur la troisième personne, plus fréquente et porteuse de signification.

En outre, nous avons pu observer sur plusieurs textes un phénomène de continuité thématique qui s'exprime à travers l'utilisation de chaînes denses de pronoms personnels sujet, liés au dernier référent explicitement cité. Dans ce but, nous avons délimité un corpus de travail où ce phénomène était évident, pour sonder la possibilité d'exploiter ensuite ce critère pour la sélection des pronoms. Par exemple, dans le cas du texte *NORM-EC-CE2-2016-104-D1-S855*, la sorcière est le référent principal d'une chaîne qui contient 8 pronoms personnels féminins de troisième personne d'affilée.

> Dans une forêt il y avait une maison où habitaient une sorcière et son chat. Son chat était noir avec des yeux perçants qui faisaient peur à un chien. Et la sorcière ce n'était pas mieux, elle attrapait un enfant et "hop" elle le mettait directement dans une cage avec des serpents, des asticots... et elle fermait la cage à double tour. comme ça l'enfant ne pouvait pas s'échapper. sinon elle le mettait directement dans la marmite. Mais un jour, elle attrapa un enfant et elle le mit dans la cage. elle voulait aller chercher un enfant à la ville qui était très très très loin. elle était bien pressée, du coup elle oublia de fermer la cage et du coup elle avait demandé à son chat de surveiller la cage mais le chat a vu une souris passer, du coup le chat la poursuit. le chat ouvre la porte et poursuit la souris. du coup le petit enfant sort de sa cage et comme la sorcière avait laissé les clés le petit enfant ferme la porte à double tour et la sorcière avait laissé sa baguette, elle jette un sort qui bloque la porte. Et ensuite la sorcière revient avec 3 enfants. Mais elle ne peut pas

rentrer. du coup elle pose le sac où il y a l'enfant et les 3 enfants s'échappent et le petit enfant qui est dans la maison appelle les 3 autres pour venir. Puis la sorcière voit son chat revenir avec une souris alors la sorcière dit à son chat « t'aurais dû la surveiller, j'avais laissé la cage ouverte ».

Ce cas en particulier se présente comme peu ambigu car un seul personnage féminin est présent dans l'histoire. Ce type de critère peut être moins productif si on traite un texte où deux personnages masculins sont présents. Toutefois, ce principe de continuité thématique semble tenir même dans des cas susceptibles de présenter des ambiguïtés comme dans le cas du texte *NORM-EC-CE2-2016-130-D1-S1125* où les pronoms personnels de troisième personne qui apparaissent sont toujours liés au dernier référent nommé explicitement dans le récit.

Il était une fois un chat qui alla dans la forêt. Il rencontra le loup noir. Le chat dit : « Qui es - tu? » Le loup noir lui dit : « Je suis le plus méchant loup de la planète et de la forêt ». Le chat s'en alla discrètement mais le loup l'avait vu. Le loup entendit un bruit mais avant que le loup aille voir d'où venait le bruit. Beee le loup enferma le chat dans une cage. Ensuite il alla voir le bruit, tout était calme mais tout à coup le loup entendit le bruit d'un mouton, il l'attaqua et le dévorait et tout à coup il entendit un bruit de chèvre un peu plus loin dans la colline mais il se disait : « non, je n'irai pas, je dois aller voir le chat ». Plus loin le chat ne réussit pas à s'en aller de la cage. Ensuite à plusieurs reprises de coup de griffes le chat réussit à sortir de la cage mais le loup venait d'arriver. le chat se cacha vite sous un grand arbre si grand que le loup ne l'avait pas vu. le loup cria « non » parce que le chat avait réussi à sortir de la cage et que le loup voulait manger le chat pour le diner. Le chat s'en alla discrètement et rentra chez lui. Mais il rencontra un chien abandonné et le ramena chez lui. Ensuite il sortit dehors pour jouer avec lui dans le jardin mais le loup arriva et ensuite le chat et le chien rentrent. le loup ne les voyait pas, il repartit dans la forêt. Mais le soir il revint mais il ne vit personne

dans la maison et il resta 1h puis il partit fatigué puis il revint le jour pour attraper le chat mais alors le chat avec le chien avaient déménagé à cause du loup . le loup désespéré attrapa un aigle pour le dîner. Il repartit dans la forêt fatigué. Il avait trouvé sa mère qui arrivait dans la forêt avec son petit frère et sa petite soeur.

6. Le modèle ArkRef – mise à l'épreuve des critères syntaxiques de sélection des antécédents

À la suite de l'étude précédente sur la détection des mentions nominales et pronominales dans les textes de notre corpus de travail, nous nous sommes occupée de l'étude des relations que les mentions coréférentielles peuvent entretenir. Nous avions en fait la nécessité, une fois les mentions détectées, de pouvoir relier entre elles les mentions qui font partie de la même chaîne de coréférence, sans nous appuyer exclusivement sur l'appartenance au même lemme. Par exemple, nous avons besoin de pouvoir mettre en relation les noms propres avec la mention nominale qui constitue son antécédent, ou de trouver l'antécédent nominal d'un pronom. Selon notre hypothèse, cette opération peut être effectuée en s'appuyant sur les résultats de l'analyse morphosyntaxique et en dépendance des textes.

Pour ce faire, nous nous sommes inspirée de l'algorithme de la première étape de sélection des antécédents effectuée par *ArkRef* (cf. Chapitre 5, 6.1). Nous avons donc sélectionné manuellement un corpus de travail contenant à la fois les mentions nominales citées précédemment (cf. Chapitre 5, 5.1) et des éléments de langage susceptibles de manifester les structures que l'on cherche à modéliser, à savoir, les verbes réfléchis, les groupes verbaux avec fonction d'attribut du sujet et l'apposition.

Nous nous sommes appuyée initialement sur les mêmes « chemins syntaxiques »[51] utilisés au sein du programme *ArkRef* en tant que modèles des

[51] Nous adoptons ici la terminologie utilisée par les auteurs du logiciel, en anglais « syntactic paths ».

structures à chercher dans le corpus, que nous avons ensuite « traduits » dans des symbolismes interprétables par les fonctionnalités de l'analyseur morphosyntaxique qu'on a cité précédemment (cf. Chapitre 5, 4.2), dons nous nous sommes servis pour sonder la présence de ces structures dans le corpus et en même temps affiner la modélisation utilisée pour la sélection d'antécédents dans notre programme. Toutefois, certains de ces critères ne semblent pas être applicables ou modélisables sur notre corpus, critères que nous allons présenter à fur et à mesure que nous décrivons notre processus de recherche ainsi que les occurrences de ces structures que nous avons pu retrouver ou pas dans nos données.

Grâce à cette analyse sur les données à disposition, nous avons dégagé des hypothèses descriptives des structures susceptibles d'être les plus productives lors de la mise en forme des règles pour notre programme, et nous avons pu constater la fréquence de la présence de certaines structures, ou l'absence tout court.

6.1. Le verbe pronominal ou réfléchi

Le verbe réfléchi est une structure que l'on retrouve assez fréquemment dans le corpus de travail notamment dans l'occurrence du verbe *s'appeler*, fréquemment utilisée pour présenter les noms des personnages protagonistes de l'histoire. Ce verbe est très présent dans le corpus avec des conjugaisons différentes, comme montrées par les exemples suivants.

Le robot s'appela Bobis. (…) [52]

Il était une fois une petite chatte qui s'appelait Boubou (…).[53]

Il était une fois une sorcière qui vivait avec un chat qui s'appelait Noisaitou et la sorcière s'appelait Rutabagae. (…)[54]

Il était une fois une sorcière qui s'appelait Carabistouie. (…) [55]

[52] Production NORM-EC-CE2-2016-102-D1-S214
[53] Production NORM-EC-CE2-2016-108-D1-S2972
[54] Production NORM-EC-CE2-2016-108-D1-S2986
[55] Production NORM-EC-CE2-2016-117-D1-S3003

Ces exemples montrent aussi la présence assez fréquente du pronom relatif *qui* pour introduire ces formules de présentation.

6.2. L'attribut du sujet et l'apposition

Nous n'avons pas relevé la présence de structures appositives dans le corpus, et même les groupes verbaux ayant pour noyau le verbe *être* en fonction attributive dont le sujet est un des personnages objets de notre recherche sont très rares dans le corpus de travail. Voici quelques exemples de cette structure extraits du corpus :

Je vous prouvais - je suis un chat perdu.[56]

« Je suis le plus méchant loup de la planète et de la forêt ».[57]

Maintenant le robot est le seul au monde.[58]

Bien sûr les humains savent qu'elle était un robot, mais les chats.... non.[59]

Mignon, lui était un chat mignon qui habitait dans la forêt, abandonné et qui cherchait souvent un refuge pour s'abriter.[60]

Nous avons décidé de garder également cette relation dans nos modélisations, car elle est probablement plus fréquente dans les niveaux de scolarité plus avancés.

6.3. Structures pas rencontrées dans le corpus de référence de CE2

D'autres structures présentes dans *ArkRef* n'ont pas été mises en place dans notre travail car nous n'avons pas trouvé d'occurrences de ces relations dans nos données :

1. la structure de « i-within-i constraint », à savoir le fait qu'un pronom ne peut pas faire référence au nœud qui le domine, comme dans l'exemple suivant, où le déterminant « its » ne fait pas référence à Walmart, qui est le nœud dominant :

Walmart says Gitano, its top-selling brand, is underselling.

---

[56] Production NORM-EC-CE2-2016-85-D1-S1981
[57] Production NORM-EC-CE2-2016-130-D1-S1125
[58] Production NORM-EC-CE2-2016-120-D1-S2439
[59] Production NORM-EC-CE2-2016-6-D1-S2000
[60] Production NORM-EC-CE2-2016-6-D1-S2035

(O'Connor & Heilman, 2013)

2. le rapport de non coréférentialité entre le sujet et la phrase nominale contenue dans un circonstant (comme dans la phrase « Pour appeler Yannick, il a pris le téléphone », où « Yannick » et « il » ne sont pas coréférentiels). À cause de la spécificité de nos données, nous avons dû réfléchir à quelles relations syntaxiques peuvent véritablement capter les phénomènes de coréférence au sein de notre corpus de travail, ce qui nous a mené, d'une part, à reprendre certains critères déjà décrits dans *ArkRef*, et d'autre part, à introduire des règles différentes pour saisir les patterns fréquents dans nos données.

7. Observations finales et hypothèses

Grâce à cette exploration, nous pouvons partir de l'observation, déjà faite par C. Wolfarth dans le contexte d'un travail sur le même corpus, que « le contexte de production du corpus constitue un élément susceptible de faciliter nos analyses » (Wolfarth, 2015 : 88). En effet, nous avons pu confirmer, par le biais d'explorations manuelles et ensuite dans l'application du programme, que le lexique et les structures utilisés dans les productions écrites des élèves sont relativement limités, ce qui permet la conception d'un système à base de règle avec une bonne capacité de généralisation sur les données à notre disposition.

Ici nous partons d'une hypothèse principale dotée de trois corollaires ou sous-hypothèses. Notre hypothèse principale est qu'il est possible de réaliser un système de résolution de coréférences à base de règles sur les écrits scolaires de niveau CE2, grâce à la simplicité des structures et du vocabulaire employé par les élèves. Les trois corollaires de cette hypothèse, qui ont guidé notre travail, sont les suivants :

1. À l'inverse d'autres systèmes de détection de mentions que nous avons pu observer, nous n'avons pas besoin d'exploiter des relations

sémantiques, comme l'hyperonymie ou la synonymie pour construire nos chaînes de coréférence. Ceci est dû notamment à la pauvreté lexicale observée dans la variation de nomination des référents.

2. Nous pouvons exploiter le principe de progression thématique à thème constant, observé dans la plupart des productions de notre corpus de travail de CE2 pour mettre en place un algorithme de construction de chaînes de références pauvre en ressources et en règles.

3. Certaines structures, sur lesquelles les outils de résolutions de coréférences à base de règles peuvent s'appuyer, ne sont pas fréquentes ou sont absentes des textes de notre corpus. Il est donc nécessaire de repenser quelles structures syntaxiques sont pertinentes pour les relations de coréférence, et peuvent être exploitées dans le cadre de notre programme. Dans notre cas, la structure du verbe réfléchi et celle du verbe réfléchi accompagné d'un pronom relatif semblent être très productives.

C'est surtout cette dernière hypothèse qui a inspiré le choix d'élaborer un nouvel outil plutôt que d'en appliquer un existant, sachant aussi que les outils *off the shelf* existants pour la résolution des coréférences en français ne sont pas simples d'utilisation et/ou ne sont pas maintenus régulièrement. Nous avons en fait rencontré des difficultés dans l'utilisation de *Decofre*, décrit par L. Grobol dans sa thèse (Grobol, 2020), ainsi que dans l'utilisation de *French-CRS*, publié par M. Mirzapour[61].

Nous allons décrire par la suite de quelle manière ces trois hypothèses ont été prises en compte dans la conception de notre outil, mais avant nous allons expliciter notre méthodologie de travail.

[61] Derniers tests effectués en janvier 2022.

8. Conclusion – poser les bases de notre méthodologie de travail

8.1. Restreindre le champ : étiquetage des quatre personnages principaux

Nous avons décidé dans ce travail d'étiqueter les chaînes de coréférence reliées aux quatre personnages principaux liés à la consigne proposée par les chercheurs lors de la récolte des données du corpus *Scoledit*. Le choix de se consacrer exclusivement à ces personnages vient, d'un côté, de la nécessité de restreindre le champ de notre recherche au niveau informatique et, de l'autre, est dicté par la comparabilité des résultats obtenus, garantie par la présence d'au moins un de ces quatre personnages tout au long du corpus, ce qui va nous permettre ensuite de vérifier l'évolution de ces chaînes par rapport aux différentes années scolaires analysées. Ces limites vont ainsi nous permettre de mettre en place un module de détection des phrases nominales relativement pauvre en ressources. Cette étude nous permet ainsi de sonder si la composition des mentions nominales modélisées au préalable est vraiment efficace et quels sont les pourcentages de structures internes de mentions nominales utilisées par les élèves.

8.2. Description du programme et des caractéristiques du corpus utilisé

En travaillant sur la forme normalisée des productions, nous allons dans un premier temps déterminer quelles sont les phrases ou les tokens nominaux qui peuvent reconduire à ces entités, sur la base de listes de lemmes extraits automatiquement à partir du corpus lui-même. Nous allons aussi sélectionner les tokens qui représentent des noms propres et des pronoms de troisième personne du singulier. Dans un deuxième temps, nous allons filtrer dans le texte des rapports comme la relation où un sujet et un objet sont reliés par un rapport de réflexivité. Dans un troisième temps, en suivant le postulat de la fréquence de continuité thématique, nous allons pouvoir relier les pronoms personnels aux entités coréférentes.

8.3. Modélisation et analyse morphosyntaxique – l'importance dans la transparence de l'analyse

Nous avons utilisé le même outil d'analyse morphosyntaxique en dépendance employé dans la construction de notre programme, en utilisant les mêmes paramètres, pour extraire des analyses complètes des textes objet des premières étapes d'analyse. Nous nous sommes largement basée sur les résultats d'analyse de *SpaCy* pour effectuer la modélisation des règles qui constituent le socle de notre outil, en excluant les analyses erronées que nous avons parfois rencontrées. Cet aller-retour continu entre analyse et performance de l'outil nous a permis d'un côté d'effectuer une modélisation le plus possible précise sur les structures présentes dans notre corpus et, de l'autre côté, nous a permis de relever toutes les erreurs d'analyse qui ont pu empêcher l'outil de fonctionner correctement. C'est le cas par exemple avec le dysfonctionnement détecté sur certains noms propres qui a été résolu avec l'implémentation d'une petite ressource externe créée à partir du corpus lui-même. Cette question sera décrite de manière plus approfondie dans le paragraphe 6.2.

Cette accessibilité de l'analyse tout au long du processus d'annotation nous permet de vérifier de manière transparente le fonctionnement de l'outil. En contraste avec les programmes à base de réseaux de neurones, dont la communicabilité et la transparence sont des problématiques pas encore résolue, un logiciel à base de règle est doté d'une communicabilité intrinsèque : cette communicabilité est pour nous un point de force incontestable, surtout dans le contexte des travaux sur les écrits scolaires, non-standardisés, où nous avons besoin de vérifier plus en détail sur quels points l'outil contient des faiblesses, pour quelles raisons il a échoué sur certains points, si l'échec provient d'une erreur commise par l'analyseur ou de l'insuffisance des règles mises en place.

Partie 3

-

Conception d'un outil d'aide à la détection des coréférences pour le corpus *Scoledit*

Chapitre 6. Présentation de *DeCorScol* et de son architecture

Après avoir testé quelques outils de résolution de la coréférence sur un corpus de travail assez restreint, constitué d'une dizaine de textes pris au hasard depuis notre corpus, nous avons pu constater de manière empirique, l'insuffisance des modèles langagiers utilisés dans l'application sur nos données. Nous avons testé ces outils sur peu de textes principalement pour des raisons techniques : l'outil était disponible par le biais d'un notebook Jupyter, et donc les affichages de résultats était de taille assez restreinte. Cette inadéquation est probablement dûe au fait que les outils actuellement disponibles publiquement sont principalement des outils à base de réseaux de neurones et que les modèles qu'ils exploitent sont entrainés sur des écrits rédigés par des scripteurs experts (comme des textes journalistiques ou des romans, dans le cas du corpus *Democrat*), ou entrainés sur des corpus de l'oral transcrit (comme dans le cas du corpus *ANCOR*). En plus, ces outils sont habituellement composés par un premier module de détection des entités nommées, qui comprennent des lieux, personnes, organisations etc., et les chaînes de référence sont ensuite construites en prenant ces entités nommées en tant qu'antécédents ; ce premier module en particulier est ce qui rend difficile la reconnaissance d'entités telles qu'un loup ou une sorcière comme référents principaux d'un texte. Sur la base de cette observation initiale, et ensuite des hypothèses qu'on a pu dégager à partir de l'observation des différents corpus de travail utilisés, nous avons donc décidé de concevoir un outil de résolution de la coréférence, fait sur mesure pour notre corpus.

Dans cette deuxième partie, nous allons présenter l'outil conçu lors de ce travail de master et son architecture : nous allons présenter le but principal de *DeCorScol* (***Dé****tection des* ***Cor****éférences dans les écrits* ***Scol****aires*) et les modules qui le composent. Ce logiciel a été conçu comme de bout en bout et à

base de règles : il prend en entrée les textes qui font partie du corpus, au format xml, et donne en sortie les mentions et les chaînes de coréférence du texte sous forme d'annotation (des balises xml entourent les mentions annotées), à partir de règles qui nous permettent de capturer les différentes mentions et les rapports intertextuels entre mentions. Le but principal de ce programme est de fournir de l'assistance à l'annotation manuelle des chaînes de coréférence sur les textes du corpus *Scoledit*.

Son architecture est librement inspirée du modèle de Haghighi et Klein (2009) et de l'outil conçu sur la base de leur publication, *ArkRef* (O'Connor & Heilman, 2013) que nous avons citée dans la partie précédente. Cet algorithme et le programme dérivé avaient été conçus et utilisés exclusivement pour la langue anglaise, ce qui a demandé un effort de réadaptation de l'algorithme et de ses règles au français, tout en tenant compte de la perspective d'application à des langues romanes autres comme l'italien et l'espagnol, dans le cadre du corpus *Scolinter*.

Les points en commun entre *ArkRef* et *DeCorScol* sont l'architecture à base de règles et un algorithme simple, qui s'appuie fortement sur le POS tagging et sur l'analyse en dépendance des textes en entrée ; les règles mises en place ont été développées en modélisant la structure interne des mentions et les relations possibles entre mentions et antécédents à l'intérieur d'un texte, sans faire recours à des ressources externes autres que les règles même et une ressource externe de correction du POS-tagging sur certains tokens fréquents.

Pour d'autres aspects, l'outil que nous avons obtenu s'éloigne considérablement de ses précurseurs mentionnés ci-dessus, dans le sens où nous avons mis en place un module de détection de mentions conçu sur mesure pour sur notre corpus de référence, et modélisé sur des référents spécifiques (à savoir, les chaînes qu'on détecte sont celles reliées aux personnages principaux de la consigne donnée aux élèves, ce qui restreint notre champ de recherche des

antécédents nominaux). En outre, les règles de détection syntaxique qu'on met en place visent exclusivement à saisir des phénomènes de coréférence dans le corpus objet de notre étude, sans se poser pour le moment des ambitions de généralisation sur des textes issus d'autres corpus. Une autre différence importante est l'absence de ressources externes autres que celles mentionnées ci-dessus, ou l'absence de modules sémantiques, ce qui nous permet de ranger notre logiciel dans la catégorie des outils « very low ressources ».

Le logiciel est constitué par trois grands modules :

1. Le premier module sélectionne des mentions sur la base du rôle syntaxique des mots qui les composent et par lemme, en créant une première version des clusters des quatre personnages des récits (le chat, la sorcière, le loup et le robot) ;

2. le deuxième module consiste dans la mise en place d'une fonction de sélection des mentions sur la base de parcours syntaxiques spécifiques ;

3. un troisième module crée des clusters sur la base des lemmes qui font référence aux quatre personnages principaux des productions écrites, avant de sélectionner les pronoms qui leurs sont connectés, pour enfin inclure dans les clusters crées lors de la première étape les mentions pronominales détectés. Les modules décrits ici de suite ont été développés en python à l'aide du package *SpaCy*. Nous avons notamment exploité deux fonctionnalités : la fonction de *Token matching* pour le premier module et de *Dependency matching* pour le deuxième.

1. Module d'identification des mentions : modélisation et représentation des quatre personnages

Un premier module s'intéresse à l'identification des éléments de langue qu'on peut définir comme mentions, c'est-à-dire comme maillons des chaînes de coréférence. Comme déjà mentionné, dans ce travail, le but est de détecter les chaînes de coréférence dont les référents principaux sont un ou plusieurs

des quatre personnages proposés dans la consigne de la production écrite (le chat, le loup, la sorcière ou le robot). Nous avons donc effectué une étape préalable d'exploration des mentions présentes dans le corpus, pour ensuite modéliser les éléments qui composent ces mentions dans le corpus de travail, en référence aux quatre personnages.

Deux réflexions principales sont ressorties de cette phase du travail :

1. la nécessité de prendre en compte toutes les formes lexicales possibles sous lesquelles ces quatre référents principaux peuvent se manifester, et les inclure dans la phase de détection ;

2. le besoin de modéliser les patterns de tokens qui peuvent constituer des mentions, selon leur rôle syntaxique.

À partir de ces deux idées, nous avons développé le premier module du programme, qui s'occupe de la détection d'un ensemble de mentions sur tout le texte. Les types de mentions qu'on a ciblé sont notamment les groupes nominaux, les noms propres, et les pronoms personnels. Nous allons décrire par la suite le double travail d'exploration et de modélisation effectué sur notre corpus de travail, en l'illustrant à travers des exemples de mentions tirés de notre corpus.

1.1. Détection des groupes nominales : noms communs

Suite à la première étape exploratoire, qu'a visée l'étude des formes nominales contenues dans les productions écrites et extraites grâce à l'étape de POS-tagging des textes, nous avons pu recenser en total 12 formes nominales connectés aux 4 personnages, dont quelque synonyme, parfois présents à la fois dans les mêmes textes.

Les variations lexicales sur les lemmes de base étant peu nombreuses dans le corpus par rapport aux occurrences des lemmes de base, nous avons donc décidé de détecter seulement les groupes nominaux qui contiennent une

mention explicite des personnages par lemme : nous allons donc idéntifier les groupes qui contiennent des mentions des tokens dont le lemme est *chat*, *sorcière*, loup ou *robot*, plus quelques-uns des synonymes rencontrés dans notre liste et identifiables de manière non ambiguë comme faisant référence à une entité donnée (par exemple *chaton* ou *magicienne* ont été inclus).

Nous avons enrichi ce pattern de détection par lemme, en incluant la possibilité que la mention de l'entité soit accompagnée par un déterminant, soit un article défini ou indéfini, soit un déterminant démonstratif, ou par une préposition ; ou qu'elle soit représentée à travers un groupe du type déterminant – adjectif – nom ou déterminant – adverbe – nom, ce que nous avons modélisé comme déterminant plus token inconnu plus lemme (personnage). Si une mention peut rentrer dans plusieurs patterns, nous prenons toujours en considération celle qui contient plus de tokens. Voici quelques exemples de mention sur l'entité *sorcière* et sa modélisation.

Mention	Modélisation
la sorcière une sorcière	DET, PronType=Art ; LEMMA (sorcière)
cette sorcière	DET,PronType=Dem; LEMMA (magicienne)
la magicienne	DET, PronType=Art ; LEMMA (magicienne)
la petite magicienne	DET, PronType=Art ; TOKEN (inconnu) ; LEMMA (magicienne)
chez la sorcière	ADP, token inconnu, lemme (sorcière)

Tableau 3. Modèles de mentions nominales adoptés dans notre programme

1.2. Détections des groupes nominaux : noms propres

Nous avons ensuite modélisé les possibles patterns de présence d'un nom propre dans les mentions du texte. Les patterns choisis ont été les suivants : celui, plutôt fréquent, composé par déterminant – nom commun – nom propre,

comme dans les exemples reportés en bas ; ensuite nous avons adopté le pattern de double nom propre, bien que moins fréquent, et en dernier le pattern d'un nom propre en isolement.

Mention	Modélisation
la sorcière Gigi le chat Victor	DET ; LEMMA(sorcière) ; PROPN
Nicolas Sarkozy François Hollande	PROPN ; PROPN
Gigi	PROPN

Tableau 4. Modèles de mentions formés par des noms propres, adoptés dans notre programme

Les séquences de tokens qui contiennent un nom commun qui fait référence à la liste des référents principaux sont ajoutés aux clusters de leurs référents à la fin de cette étape. Pour le moment, les noms propres en isolement ne sont pas pris en compte dans cette première clustérisation, même si le programme nous permet de visualiser s'ils ont été correctement détectés. Ils sont cependant annotés par genre, si l'information est disponible. Les noms propres seront insérés dans les chaînes de coréférence lors de l'étape de la troisième étape du programme, celle de détection des parcours syntaxiques et ensuite filtrés par similarité de surface : si un nom propre est associé à *sorcière* dans un texte, si des autres occurrences de ce nom propre sont dans le texte, elles seront automatiquement ajoutées au même cluster de *sorcière*.

1.3. Détection des pronoms ; choix méthodologique

Nous avons décidé de restreindre le champ des mentions à la détection des quatre personnages qui peuvent représenter les protagonistes principaux des histoires décrites par les enfants, notre champ de recherche et de détection est d'un côté extrêmement réduit, de l'autre nécessite un effort supplémentaire lors de la détection des formes pronominales présentes dans les textes. Les

formes pronominales étant très liées au contexte, elles représentent peut-être les formes les plus difficiles à désambiguïser. Dans ce programme, nous avions initialement détecté seulement les pronoms personnels, sujet et objet, de troisième personne singulière, avec le but ensuite d'appliquer un filtre par genre pour les attribuer à leur chaîne de coréférence d'appartenance.

Comme déjà mentionné (cf. Chapitre 5, 5.3), nous avons consciemment décidé de sélectionner seulement la troisième personne car elle est la forme la plus fréquente dans le récit, avec les formes de première et deuxième personne plutôt occurrentes dans des contextes de dialogue direct ou parfois dans le récit à la première personne, avec ce dernier point de vue très rarement présent dans notre corpus de travail. Les pronoms de première et deuxième personne, ayant souvent un rôle de déictique dans les récits, s'avèrent être plus difficiles à traiter dans le cas d'un système pauvre en règles et ressources. Nous avons donc fait le choix de ne pas traiter les extraits de textes représentant des dialogues directs et de ne pas inclure dans les mentions détectées la première et deuxième personne des pronoms personnels, et de concentrer nos efforts sur la troisième personne, plus fréquente et porteuse de signification.

1.4. Détection des incipit

En nous inspirant d'un mémoire de master qui s'est occupé des chaînes de références et des incipit dans les productions d'élèves du corpus *RésolCo* (Pons, 2019), nous avons inséré un pattern qui détecte l'incipit des productions, lorsqu'il est constitué de la formule « Il était une fois ». Le verbe être est ici détecté par lemme, pour éventuellement comprendre toutes formes erronées. Ce pattern a été inséré pour pouvoir ensuite exclure des chaînes de référence le pronom vide qui introduit la formule d'incipit.

1.5. Première clustérisation des noms communs

A la suite de cette étape de détection des mentions nominales, nous allons créer quatre clusters, un pour chacun des personnages possibles de l'histoire. Chaque cluster contient les mentions associées au lemme ou aux lemmes d'un personnage donné, y compris les mentions qui contiennent des noms propres.

À la fin de cette première étape, nous pouvons déjà visualiser en sortie du programme quelles mentions ont été détectées sur tout le texte, en ayant à disposition leur représentation en tant que *Spans*. Les *Spans* sont des structures de *SpaCy* : elles constituent des ensembles de tokens, avec un indice de début et un indice de fin ; en outre, les étiquettes morphosyntaxiques et en dépendance de chaque token restent accessibles tout au long du programme[62].

Ce module constitue déjà en soi un système d'assistance pour l'annotation, car il nous donne en sortie toutes les mentions d'un texte en exprimant la catégorie d'appartenance des mentions, sans pourtant construire des chaînes de référence. Ce premier module nous a aussi permis de conduire les analyses présentées dans la partie 6 de ce mémoire.

2. Module de sélection des mentions par relations syntaxiques

Cette section présente le fonctionnement du module de sélection des mentions par relations syntaxiques et en dépendance. Inspiré de programmes déjà existants et développés pour la langue anglaise, ce module a comme but de filtrer certaines mentions présentes dans le texte, notamment les pronoms et les noms propres, en exploitant l'analyse morphosyntaxique et en dépendance fournie par *SpaCy*.

[62] La documentation de *SpaCy* sur les Spans est disponible au lien suivant https://*SpaCy*.io/api/span.

2.1. Le modèle ArkRef

Le deuxième module qui compose *DeCorScol* est fortement inspiré du module de sélection des antécédents de l'algorithme de Haghighi et Klein (2009) et surtout aux « chemin syntaxiques » présents dans le logiciel *ArkRef* (O'Connor & Heilman, 2013). Comme précisé dans le chapitre 7, nous nous sommes fortement appuyés sur les modélisations de règles syntaxiques proposées par cet outil pour réaliser les règles employées par notre outil. Cependant, à cause de la nature de notre corpus, nous avons dû repenser ces « chemins syntaxiques » de manière à obtenir des règles le plus possible productives pour notre cas spécifique. Une différence avec les chemins *d'ArkRef* est dans la formalisation des éléments détectés : alors que leur parseur (celui de Stanford) se base sur une représentation par phrases et par syntagmes, nous exploitons les étiquettes morphosyntaxiques et d'analyse en dépendance fournie par *SpaCy* pour mettre en place notre module de détection des relations entre mentions.

2.2. Le verbe réfléchi

Ce critère de sélection de l'antécédent d'une mention donnée était déjà présent entre les critères de sélection des antécédents dans *ArkRef*, et il a été aussi intégré au sein de notre programme à travers deux différentes modélisations.

La phrase contenante un verbe réfléchi est une structure qu'on retrouve assez fréquemment dans ce corpus de travail, notamment dans la structure du verbe *s'appeler*, fréquemment utilisé pour présenter les noms des personnages protagonistes de l'histoire. Comme déjà remarqué, les noms propres peuvent présenter des difficultés d'insertion dans des chaînes, surtout dans le cas de noms inventés ou rares. Pour cette raison nous avons testé plusieurs règles de détection de ces structures, de plus en plus fines, pour s'adapter au mieux à nos données.

Nous avons rencontré dans le corpus plusieurs modélisations possibles des verbes réfléchis, dont deux ont été modélisées sur les formules de présentations qui impliquent la présence d'un nom propre. Grâce à la fonction de *Dependency matcher* de *SpaCy*, nous avons pu détecter ce type de pattern :

1. Un nom commun qui fait partie des référents principaux précède immédiatement un pronom relatif et les deux font partie du même arbre syntaxique ;
2. Le pronom relatif précède immédiatement le pronom réfléchi et les deux font partie du même arbre syntaxique ;
3. Le pronom réfléchi précède un nom propre et les deux font partie du même arbre syntaxique.

Le deuxième pattern exploite une modélisation similaire, qui inclut la conjonction coordonnante comme premier token du pattern (à laquelle *SpaCy* attribue l'étiquette d'analyse en dépendance *cc*) et n'inclut pas le pronom relatif.

Même si ces modèles provoquent encore du bruit dans les résultats, ils sont particulièrement efficaces car ils permettent de ne pas relier totalement ces types de formules de présentation à la présence d'un verbe spécifique, et permettent ainsi de contourner le possible obstacle de la distance entre la mention de l'antécédent et le nom propre successif.

Les autres modélisations possibles faites au préalable semblent avoir été surpassées par celle ici mentionnée, qui a été donc le seul pattern du verbe pronominal retenu dans notre logiciel.

2.3. Résultats du Dependency matcher

Tous les patterns décrits dans cette phase contiennent initialement tous les éléments décrits, donc une opération de filtrage est nécessaire avant de passer à l'étape suivante pour pouvoir éliminer des tokens qui ne font pas partie de nos chaînes de référence (comme les verbes pronominaux par exemple).

3. Sélection des pronoms et clustérisation

Le but de ce troisième module est de sélectionner des pronoms qui peuvent faire partie d'une chaîne de coréférence sur la base de leur relation syntaxique et en dépendance avec une mention qui en fait déjà partie.

Pour chaque phrase du texte, le programme sélectionne tous les éléments nominaux ou pronominaux qui représentent des sujets des phrases (dont l'étiquette en dépendance est soit *nsubj,* soit *expl :subj,* soit *ROOT,* étiquettes attribuées aux tokens ayant le rôle de sujet dans la phrase). Tous ces éléments sont insérés dans une structure de type liste ; ensuite l'algorithme compare chaque élément contenu dans la liste avec celui qui le suit. Si un nom commun est suivi d'un pronom, le pronom est ajouté au cluster correspondent à celui du nom commun qui le précède ; si un pronom est suivi d'un autre pronom, le deuxième pronom est ajouté au cluster où se trouve déjà le premier pronom.

Ce mécanisme apparemment simple permet d'obtenir des résultats plutôt corrects, si nous postulons la prépondérance des stratégies de progression thématique à thème constant dans les productions analysées.

La dernière étape consiste dans la sélection des mentions détectées par le *Dependency matcher* du deuxième module et de les insérer dans un des quatre clusters disponibles.

Dans la toute dernière étape, à la suite de la complétion des clusters et au filtrage des mentions qui se superposent, une fonction permet d'annoter le texte en entrée avec des balises xml, du type , où CHARACTER est le personnage représenté dans la mention.

En perspective, nous souhaitons aussi intégrer l'analyse effectuée dans le fichier json qu'on obtient en sortie du module de prétraitement des données, ce qui nous permettra ensuite d'obtenir un corpus complet en annotation morphosyntaxique et en coréférence, en obtenant ainsi le premier corpus scolaire annoté automatiquement en coréférences.

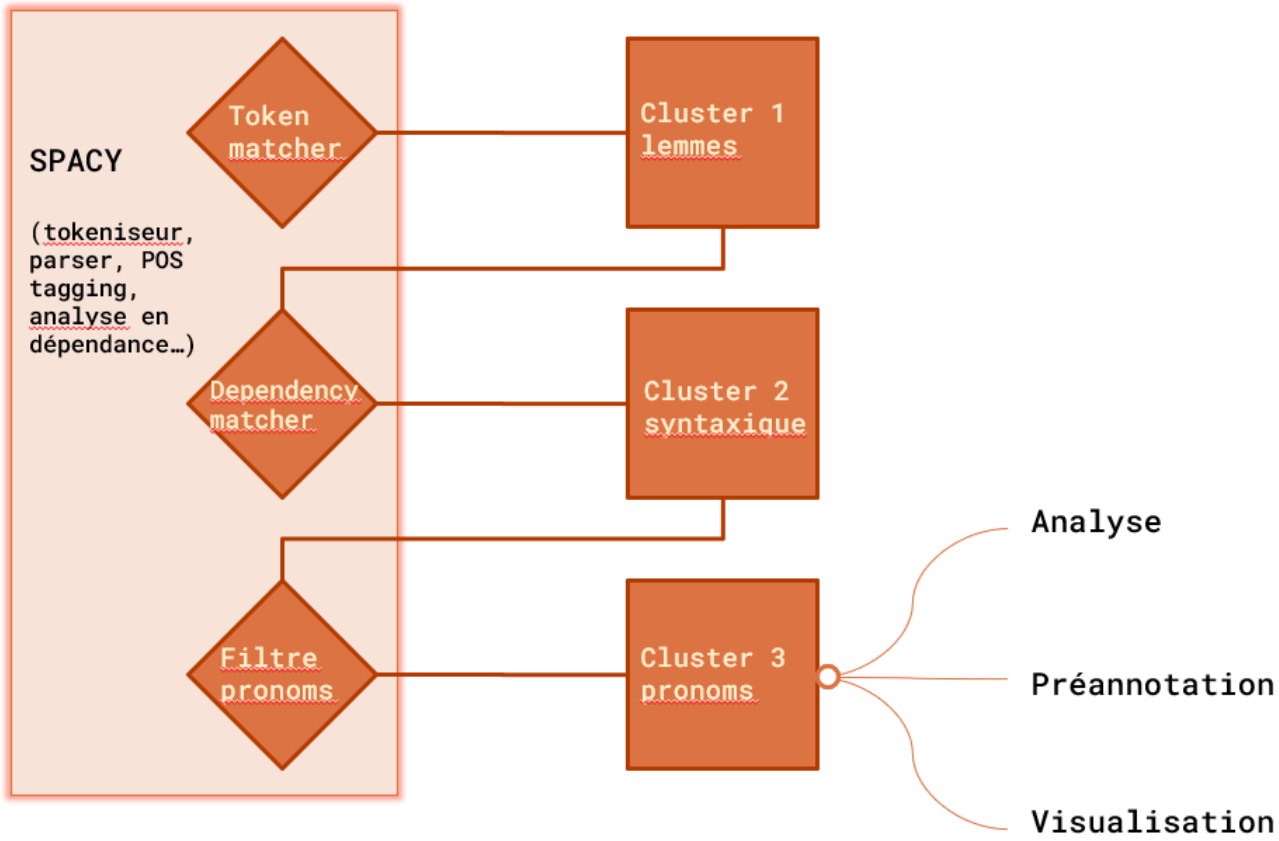


Figure 14. Schéma qui représente les différents modules et sorties du programme DeCorScol

Partie 4

-

Résultats observés grâce à l'outil

Chapitre 7. Quelques analyses sur les résultats obtenus grâce à *DeCorScol*

Dans cette quatrième et dernière partie nous allons présenter les observations issues des résultats obtenus en utilisant l'outil *DeCorScol* (***Dé****tection des* ***Cor****éférences dans les écrits* ***Scol****aires*) sur notre corpus de travail de CE2, composé de 50 textes. Comme déjà mentionné, l'objectif de ce programme est de fournir un outil d'assistance à l'annotation manuelle des chaînes de coréférence sur les textes du corpus *Scoledit*, et les résultats obtenus avec cette première version de logiciel semblent être assez prometteurs pour l'utilisation qu'on souhaite en faire.

Le travail présenté dans ce mémoire constitue la première expérience d'annotation en coréférence de ce corpus : à cause du manque d'une version déjà annotée en coréférence (en absence de ce qu'on appelle un *golden standard*[63]), nous ne pouvons pas employer les métriques habituellement utilisées pour la mesure de l'efficacité des outils de résolution des coréférences en termes par exemple de précision, de rappel et de F-mesure. Nous allons donc ici présenter principalement des observations manuelles, effectuées sur le corpus de référence de CE2.

Nous allons initialement vérifier si le premier module du programme, chargé de détecter les mentions nominales et pronominales dans le texte, a été capable d'accomplir son rôle ; ensuite nous allons évaluer le fonctionnement du deuxième module, et nous allons voir de quelle manière les structures que nous avons utilisées pour relier mention et antécédent ont produit des résultats

[63] Le *golden standard* est habituellement un corpus déjà annoté qui sert de comparaison pour mesurer la réussite de l'outil à évaluer.

différents des attendus et pourquoi. Enfin, nous allons estimer dans quelle mesure le critère de filtrage de pronoms basé sur la progression thématique à thème constant a effectivement fourni les résultats espérés.

1. *Première analyse : détection des mentions et mise à l'épreuve du modèle « sans sémantique »*

1.1. Les mentions nominales : noms communs

Dans la phase d'évaluation des performances de l'outil, nous avons initialement ciblé les textes contenants plusieurs formes de référence pour indiquer le même référent, afin de vérifier la capacité du logiciel à mettre en relation les différentes mentions qui appartiennent à une chaîne, sans pour autant avoir recours à une ressource externe autre que les lemmes codés dans le programme.

Sur la base de la liste des formes nominales mentionnée précédemment, qui nous a permis de corriger quelques erreurs importantes commises par l'analyseur, nous avons pu sélectionner les productions contenant le plus de variation au niveau de la structure interne de la mention : sur ces productions nous avons testé le premier module de notre programme.

Par exemple, nous avons initialement testé la fonctionnalité d'étiquetage des mentions sur le texte *NORM-EC-CE2-2016-108-D1-S2978*.

La sorcière et son chat

Il y avait une fois une sorcière appelée Camille qui vivait heureuse avec son chat mignon, qui s'ennuyait. La petite magicienne eut un jour une visite imprévue qui parlait d'école pour sorcière, Camille très mécontente eut la mauvaise idée de répondre non. Le jeune homme très fâché, alors qu'il ne croyait pas avoir une réponse pareille préféra partir. Camille entièrement fâchée jetait un sort à son adorable minou qui doit maintenant supporter des poils bleus. Si la visite n'aurait pas eu lieu le pauvre mignon ne serait pas en colère contre sa maitresse. Trois jours plus tard la magicienne en bon état rendait sa couleur à

son chaton (marron clair, les yeux bleus). Depuis ce jour le chat et la sorcière sont heureux. FIN.

Dans ce texte nous pouvons observer une certaine variation lexicale sur les deux personnages décrits. En effet, sur la base de ce texte nous avons adopté les différents lemmes *sorcière* et *magicienne* pour les associer au cluster de *sorcière*, ainsi qu'inclure un token dont les attributs ne sont pas spécifiés dans le pattern à détecter pour pouvoir inclure la présence de l'adjectif entre l'article et le nom commun. Nous avons aussi pu associer les lemmes *minou* et *chaton* aux cluster de *chat*. Pour nous assurer que ces différents éléments rentrent dans le même cluster, nous avons exploité une fonction d'étiquetage des mentions : chaque mention est étiquetée avec son lemme de référence. Tous les mentions qui portent des étiquettes reliées à la même entité sont insérées dans le même cluster, de manière telle à rendre la clustérisation des mentions un procès sans ambiguïté. Ce choix présente l'avantage de relever d'une implémentation simple au niveau informatique.

Cependant, cette fonction n'est pas suffisante si une ressource sémantiquement plus complexe est associée au programme : par exemple, dans un texte où le lemme *animal* est présent, et dans le cas où différents personnages peuvent lui être reliés, ce critère d'étiquetage simple n'est pas suffisant afin de résoudre cette ambiguïté. À cette première règle, il sera indispensable d'associer une fonction supplémentaire qui puisse sélectionner l'antécédent correct par rapport au lemme hyperonymique, par exemple.

Sur une plus large échelle, ce critère semble donner des bons résultats sur la majeure part du corpus : il est rare que les mentions qui font référence à la sorcière par exemple ne soient pas incluses dans la chaîne, comme dans le cas du texte *NORM-EC-CE2-2016-108-D1-S2988*.

Dans ce cas, la liste de lemmes n'a pas été suffisante : la sorcière est en fait nommée dans le texte comme « la vieille », et ensuite « la petite dame » ou « la dame ».

> Il était une fois une sorcière et son chat qui vivaient dans la ville de Pastise où il n'y avait que des sorciers et des sorcières . un jour le voisin de la vieille va transformer le chat en petit minou méchant et du coup le petit minou griffa la vieille. c'est devenu un gros désastre alors la petite dame fait une potion pour tuer son chat . elle attendait une semaine pour pouvoir tuer son chat . le minou fut mort et la dame fut triste mais elle est si vieille qu'elle meurt vieille et triste.

Aussi, les deux adjectifs, préposé et postposé qui sont censé faire partie de la mention « petit minou méchant », n'ont pas été insérés dans les limites de la mention « minou ». C'est-à-dire que notre programme ne prend pas en compte des patterns de ce type ou des enchaînements différents par rapport à ceux qu'on a décrit précédemment (cf. Chapitre 6, 1)

Une autre source d'erreur dans la clustérisation des mentions nominales, peut être reconduite aux lemmes qu'on peut potentiellement attribuer à plusieurs des entités représentées dans nos textes ; donc à ces lemmes qu'on peut considérer sémantiquement ambigus dans ce contexte. Cette problématique invalide l'utilisation exclusive des lemmes contigus (du type chat, chaton, minou) pour constituer une première phase de clustérisation dans le programme.

Dans le cas de ces mentions « ambigües », c'est le contexte qui est nécessaire pour pouvoir attribuer correctement la mention à son cluster d'appartenance (déjà observé dans le chapitre sur les mentions nominales, cf. Chapitre 5, 5.1).

Un exemple évident est celui du lemme « maitresse ». Ce lemme est parfois relié à la sorcière, d'autre fois il introduit un nouveau référent. Par exemple, dans le contexte du texte *NORM-EC-CE2-2016-104-D1-S841*, il est

clair que la maîtresse du chat (dont les mentions sont surlignées et en gras) est un personnage à part entière par rapport à la sorcière (dont les mentions sont en cursive et en gras), elle aussi présente dans cette production écrite.

> C'est l'histoire d'un chat avec **sa maitresse** qu'il aimait beaucoup. Un jour ils se baladaient, le petit chat partit explorer les environs. il sautait, grimpait aux arbres et galopait si loin qu'il s'était perdu. à un moment donné il vit une grotte il rentra. tout d'un coup ce petit chat perdu vit une sorcière laide mais très gentille et elle lui dit "Que fais-tu là mon ami?" Le chat répondit tremblant "Je me suis perdu. pouvez-vous m'aider s'il vous plait". "Oui bien sûr tu es si mignon". "Merci". Ils étaient devenus les meilleurs amis du monde. Ils cherchaient, cherchaient. le petit chaton perdu était désespéré. La sorcière lui dit "Ne t'inquiète pas nous allons trouver mais tout de suite rentrons. il commence à faire nuit et à pleuvoir." Le lendemain ils cherchaient encore et encore, tout d'un coup le petit chat qui avait une bonne vue vit sa maison, par la fenêtre il voyait **sa maitresse** pleurer. il rentra et **elle** sauta de joie. la sorcière était triste alors on lui dit de venir vivre ici. Fin

Dans le cas du texte *NORM-EC-CE2-2016-108-D1-S2978*, la sorcière est aussi individuée comme maîtresse du chat :

> La sorcière et son chat
>
> Il y avait une fois une sorcière appelée Camille qui vivait heureuse avec son chat mignon, qui s'ennuyait. La petite magicienne eut un jour une visite imprévue qui parlait d'école pour sorcière, Camille très mécontente eut la mauvaise idée de répondre non. Le jeune homme très fâché, alors qu'il ne croyait pas avoir une réponse pareille préféra partir. Camille entièrement fâchée jetait un sort à son adorable minou qui doit maintenant supporter des poils bleus. Si la visite n'aurait pas eu lieu le pauvre mignon ne serait pas en colère contre sa maitresse. Trois jours plus tard la magicienne en bon état rendait sa couleur à son chaton (marron clair, les yeux bleus). Depuis ce jour le chat et la sorcière sont heureux. FIN.

Dans ce deuxième texte, il est évident, à partir du contexte, que les mentions de la *sorcière Camille* et celle, unique, de la *maîtresse* du chat, sont

reliées à la même entité. Nous pouvons faire ce constat sur la base de l'absence d'autres personnages de genre féminin dans le texte, confirmé par les phrases qui clarifient le rapport entre chat et sorcière.

Pour pointer les différentes occurrences de structure interne des mentions nominales détectées, nous avons utilisé le premier module de notre programme, celui qui effectue la détection des groupe nominaux. Sur un total de 985 formes nominales dans tous les textes, nous avons appliqué le premier module seulement sur les groupes nominaux qui comprennent les lemmes des quatre personnages de la consigne. Nous avons ainsi détecté 445 mentions nominales, dont les trois structures internes plus fréquentes sont les suivantes :

Catégorie	Exemples	Nb
Article défini sing.+nom	Le chat Les chats	267
Article indéfini sing.+nom	Un chat	74
Nom	Chat	63
Article défini singulier + adjectif + nom	Le petit chat	9
Article indéfini singulier + adjectif + nom	Un petit chat	6
Article défini pluriel + nom	Les chats	4

Tableau 5. Occurrences de mentions nominales détectées par le programme

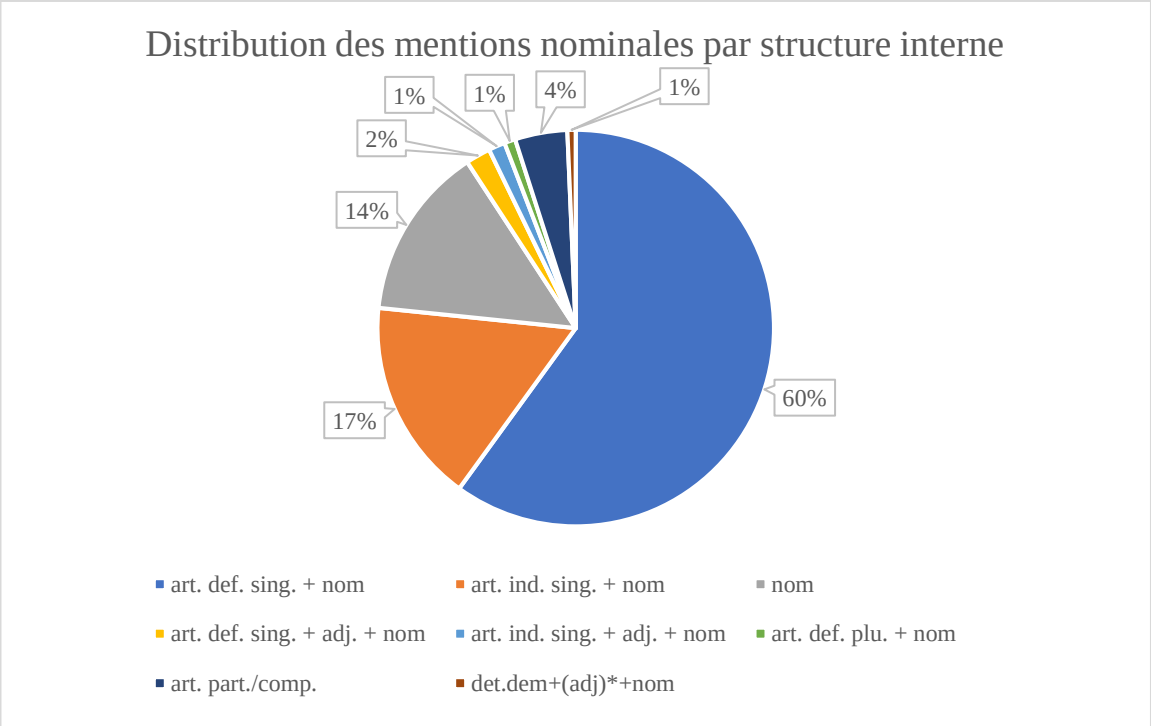
Cette liste nous permet d'observer que les mentions nominales composées par un article singulier (défini ou indéfini) et un nom commun, sont les plus fréquentes dans le corpus, 341 mentions sur un total de 445 mentions nominales détectées par ce module, soit 77% des mentions détectées. Les mentions introduites par un article défini représentent donc la majorité des mentions nominales détectées.

Ces mentions surpassent largement celles introduites par un article indéfini et les mentions qui contiennent le mot seul. Nous avons ainsi pu observer le fait

que dans toutes les occurrences de la structure article défini + adjectif + nom, l'adjectif présent est toujours « petit » éventuellement décliné. Les mentions introduites par un déterminant démonstratif, avec ou sans un adjectif, ont été comptabilisées au sein d'une seule catégorie car peu fréquentes. Une dernière catégorie est présente, où nous avons placé toutes les occurrences de articles partitifs et composés, à cause du manque de distinction et de l'ambiguïté de ses déterminants, catégorie qui comprend 19 mentions.

Nous pouvons ainsi constater, depuis la liste complète des éléments détectés, que cette méthode nous permet d'obtenir une liste de mentions nominales sans aucun bruit.

Cependant, dans les textes mêmes de notre corpus de travail, nous avons



pu observer des occurrences de mentions nominales qui contiennent des adjectifs postposés aux noms, parfois précédés d'un ou plusieurs adverbes. Dans cette version du programme, ce type de pattern n'est pas pris en compte, mais nous allons discuter plus tard des possibles futures implémentations de ce type de structure interne de mention au sein de notre logiciel (cf. Chapitre 8).

1.1. Les mentions nominales : noms propres

En ce qui concerne les noms propres, l'outil est bien capable de détecter ceux qui sont rattachés à un token dont le lemme est un de ceux défini dans notre lexique. C'est le cas, par exemple, dans le texte *NORM-EC-CE2-2016-102-D1-S212*, grâce à une fonction du premier module qui modélise la co-occurrence de noms propres et des lemmes définis en amont.

> Il était une fois une sorcière. elle avait un chat magique, le chat avait le pouvoir de changer de couleur. La maison de la sorcière Gigi était noire donc vu que le chat Victor pouvait changer de couleur, Gigi tomba mais un jour elle a dit « j'en ai marre oust. » Un jour un robot est entré dans la maison de Gigi et rentra dans l'ordre. Victor n'avait plus de pouvoir, Gigi arrêtait de tomber et le robot s'en alla.

Dans ce cas, et dans des cas similaires, l'outil a été capable de relier le nom propre déjà rattaché à un lemme donné à l'occurrence suivante du nom propre. Nous allons approfondir la manière dont les noms propres ont été pris en considération dans les chaînes de coréférence détectées par le programme dans le chapitre dédié aux résultats finaux.

Cependant, nous avons pu constater des erreurs de l'analyseur. Sur un total de 118 occurrences de 68 mots, 22 occurrences de 18 mots ne sont pas des noms propres. Nous allons aussi constater que notre outil semble échouer sur la reconnaissance de certains noms propres, probablement à cause d'un problème interne de l'analyseur.

Par ailleurs, une structure peu fréquente, observée dans notre corpus et que le système n'est pas en mesure de reconnaitre, est celle du nom propre antéposé au syntagme nominal qui désigne un référent, comme dans le texte *NORM-EC-CE2-2016-108-D1-S2981*.

> Il était une fois une sorcière qui habitait seule avec son balai qui faisait le ménage. Un jour elle eut l'idée de construire un robot pour l'accompagner. Mais ce jour -là Bot le robot n'arrêtait pas de tout casser. Plus tard elle jeta un sort pour

que l'engin mécanique parte d'ici mais ça n'avait pas marché. Alors elle l'a pris son ami et l'a balancé dehors. Le lendemain le robot cassa la porte pour tout réparer ses bêtises. Et il alla réveiller Marianne la sorcière pour se faire pardonner. Depuis ce jour ils faisaient des jeux de cartes et des blagues, même une fois le balai a pris une photo de Bot et la vieille en train de cuisiner.

Dans cet exemple, les deux noms propres présents dans le texte, surlignés en jaune, n'ont pas été correctement connectés aux référents présents dans le texte, même si le programme avait été capable de les reconnaitre en tant que noms propres (ce qu'on a pu observer sur la base de la sortie de notre programme, qui nous permet de visualiser une liste contenant toutes les mentions individuées avant qu'elles soient divisées en clusters).

1.2. Les mentions pronominales

Comme mentionné précédemment (cf. Chapitre 5, 5.3), nous avons aussi dans un premier temps inclus les pronoms personnels au sein des mentions individuées par le premier module de notre programme. Nous avons décidé de traiter exclusivement les pronoms personnels de troisième personne et de ne pas traiter a priori les pronoms insérés dans des contextes de dialogue direct. Cependant, la détection de pronoms du premier module du programme n'a pas été employé car non nécessaire à cause de la fonction de filtrage des pronoms, décrite dans le chapitre 8. Nous allons décrire plus en détail par la suite les résultats obtenus sur ces types de mentions.

2. Deuxième analyse : le verbe pronominal et les formules de dénomination

Dans le chapitre dédié au deuxième module de l'outil (cf. Chapitre 6, 2), nous avons déjà mentionné les tests que nous avons effectués selon différents patterns dans le but de détecter correctement les rapports de réflexivité, surtout dans les cas de formules d'introduction des personnages cibles de notre analyse. Ce critère de sélection, qui nous paraît le plus pertinent pour détecter tous les noms propres, ne se révèle pas toujours efficace. Nous allons ici décrire

quelques exemples de ces formules de présentation, et comment le programme a réussi ou pas à les détecter. Dans le cas du texte *NORM-EC-CE2-2016-108-D1-S2972*, à cause probablement d'un dysfonctionnement du programme, nous n'avons pas correctement relié le nom propre « Boubou » à sa chaîne d'appartenance, celle de « petite chatte ». Néanmoins, en observant la sortie du pattern de détection du verbe réfléchi accompagné d'un pronom relatif, nous pouvons remarquer que les deux occurrences du nom propre étaient correctement reliées à leur antécédent, bien que le programme n'ait pas été capable de les étiqueter ensuite.

> Il était une fois une petite chatte qui s'appelait Boubou, on l'avait appelée Boubou parce qu'elle adorait faire des blagues. Mais un jour, ***cette petite farceuse*** a eu l'idée de dire une blague méchante aux policiers. Mais en plus qu'elle était toute petite elle ne savait pas qu'elle pouvait aller en prison. Mais quand elle vit les habits du monsieur tomber elle vit que c'était un robot qui lui dit « C'est carnaval aujourd'hui, viens je vais te déguiser ». Et ils firent la fête toute la journée, et depuis ce jour le chat et le robot ne se quittèrent plus. « FIN »

Il est possible de trouver des autres exemples similaires dans notre corpus de travail, où les patterns détectés sont corrects mais l'étiquetage résultant ne l'est pas, ce qui est probablement dû à des problèmes de la fonction d'étiquetage, comme dans le cas des incipit de plusieurs textes.[64]

Le tableau 6 montre l'étiquetage obtenu en sortie du programme en regard de celui obtenu par le module de détection des verbes réfléchis sur la première phrase du texte.

[64] Les incipit ici cités proviennent respectivement des textes NORM-EC-CE2-2016-6-D1-S2035, NORM-EC-CE2-2016-108-D1-S2986, et NORM-EC-CE2-2016-117-D1-S3003

Dans certains cas, ces patterns peuvent aussi apporter du bruit dans l'étiquetage mais ils sont neutralisés grâce aux fonctions de filtrage mises en place successivement dans le programme.

Sortie obtenue par le programme	Sortie obtenue par le Dependency Matcher (deuxième module)
Il y avait une sorcière qui s'appelait Sorcièla et un chat qui s'appelait Mignon.	Il y avait une sorcière qui s'appelait Sorcièla et un chat qui s'appelait Mignon.
Il était une fois une sorcière qui vivait avec un chat qui s'appelait Noisaitou et la sorcière s'appelait Rutabagae .	Il était une fois une sorcière qui vivait avec un chat qui s'appelait Noisaitou et la sorcière s'appelait Rutabagae .
Il était une fois une sorcière qui s'appelait Mademoiselle Cascou et un chat qui s'appelait Minou.	Il était une fois une sorcière qui s'appelait Mademoiselle Cascou et un chat qui s'appelait Minou.

Tableau 6. Comparaison entre sorties étiquetées par le programme et sorties prévues par le module

3. Troisième analyse : vérification de progression thématique à thème constant

En dernier lieu, nous avons analysé les chaînes produites par notre programme, en portant attention à l'efficacité du critère de sélection de pronoms personnels selon le postulat d'une utilisation fréquente de progression thématique à thème constant de la part des élèves. Voici quelques exemples de sorties étiquetées pour lesquelles le programme a pu détecter plusieurs chaînes de coréférence pour des personnages différents.

Par exemple, dans le texte suivant, la sorcière est évidemment la protagoniste de l'histoire et l'application de ce critère fait qu'on peut effectivement associer

la majorité des pronoms à son cluster. Seulement trois pronoms restent en dehors de l'étiquetage (en jaune dans le texte) : les pronoms précédés par des prépositions (b) et (c) car ils ont étiqueté comme *obl:arg*, ce qui n'est pas pris en compte par le programme ; le premier prénom (a) qui est étiqueté comme *nsubj*, est probablement absent à cause d'une erreur dans la fonction de clustérisation du programme.

> Il était une fois une sorcière terrifiante à la voix lugubre, elle(a) avait un chat noir comme une nuit sans étoile et gros, il était aussi très sale. La sorcière partit souvent à la chasse sur son balai magique, un jour qu'elle partait à la chasse elle rencontra un loup, comme elle en avait grand besoin pour ses potions, elle fit boire à Cupidon une potion dont elle ne se séparait jamais, et qui faisait qu'elle pouvait mettre ses idées dans la tête de Cupidon, ainsi elle pouvait lui faire faire ce qu'elle voulait. elle put lui faire planter une flèche dans les fesses du loup pour qu'il soit amoureux d'elle(b). Elle put l'emporter chez elle(c) pour s'en servir pour ses potions magiques. Fin.[65]

Dans le cas du texte suivant, où deux personnages sont présents dans l'histoire, nous pouvons observer que la clustérisation est bien réussie du point de vue des regroupements par lemme. Cette fonction a toutefois échoué sur un bon pourcentage des pronoms. La moitié des pronoms féminins de troisième personne a été correctement reliée au cluster de *sorcière*. L'autre moitié (en jaune) contient des erreurs dues au manque de détection ou des clustérisations erronées sur le cluster de *chat*, l'autre personnage de l'histoire. On peut observer que les deux catégories des pronoms absentes de la détection appartiennent toujours aux deux catégories *nsubj* (les pronoms (a), (b) et (c)) et *obl:arg* (d).

> Il était une fois une sorcière qui voulait manger un chat très futé et qui faisait rater toutes ses potions. Un jour elle a essayé de lui jeter un sort

[65] Production NORM-EC-CE2-2016-37-D1-S1560

> qui rend instantanément aveugle mais il(a) se cacha derrière un miroir et la sorcière devint aveugle. Après tous ses échecs elle(b) a voulu jouer loyal. Alors elle proposa un combat. Le chat accepta. La sorcière utilisa son balai sans succès. Alors elle fonça pour l'attraper, il(c) évita habilement. La sorcière se retenait pour ne pas tomber. Le chat attaqua ses mains et elle tomba dans le puits. Depuis ce jour on n'entendit plus jamais parler d'elle.(d) [66]

Dans le cas suivant, nous pouvons remarquer que le pronom vide de l'expression « il était une fois » a été inclus dans la chaîne du sujet de la phrase suivante, c'est-à-dire il a été inclus à la chaîne de la sorcière. Même si un pattern de détection de cet incipit du récit a bien été mis en place, nous n'avons pas réussi à éliminer la reconnaissance de cette expression tout au long de l'analyse. Le pronom qui fait partie de l'expression rentre donc parfois dans l'analyse lors de la détection des pronoms mise en place dans le troisième module.

De plus, nous pouvons remarquer une erreur sur le premier pronom personnel sujet féminin de troisième personne. Il n'a pas été détecté, tout en étant bien étiqueté comme *nsubj*, comme dans le cas des textes précédents.

> La sorcière et le chat Il était une fois une sorcière, qui s'appelait Juliette. Elle(a) n'aimait pas du tout ce nom, elle voulait s'appeler la sorcière de la rue Broca. Et elle avait un chat, elle ne l'aimait pas du tout, c'est pour ça qu'elle l'appelait Juliette. Ce chat n'aimait pas du tout ce nom et lui comparé à la sorcière, il aimait la sorcière. Un jour la sorcière le transforma en petite fille, et la sorcière trouva cette petite fille très gentille et très mignonne. La sorcière a appelé sa fille (son chat qu'elle a transformé en fille), la fille de la sorcière Broca et la sorcière acheta un nouveau chat et plein d'animaux pour sa fille. Mais sa fille voulait encore plus d'animaux mais tous ces animaux étaient méchants et la sorcière les transforma en petites filles, en petits garçons

[66] Production NORM-EC-CE2-2016-20-D1-S98

et en grandes filles et en grands garçons. Un chat apparait et la sorcière n'avait plus de magie et c'était un chat gentil donc la sorcière aimait ce chat et tous ces enfants aussi.[67]

Un autre questionnement sur les possibilités de développement ultérieur de l'outil provient des observations sur la détection des pronoms de troisième personne plurielle comme dans le cas du texte suivant. Ces pronoms ont été détectés de manière erronée. Cependant, bien qu'ils soient peu fréquents dans notre corpus, grâce à cette erreur nous avons pu observer les résultats que le programme nous propose à l'état actuel en les traitant.

Dans ce texte, les deux personnages font des actions ensemble, ce qui est indiqué par la présence d'une chaîne de pronoms de troisième personne, qui ont été systématiquement rattachés au cluster du *loup*. Ici nous avons à traiter une mention qui indique deux référents séparés. Il faudra donc décider entre

- insérer ces types de mentions dans les deux (ou plus) chaînes des antécédents,
- constituer une chaîne à part,
- ne pas prendre en considération ces éléments dans l'étiquetage.

Il était une fois un loup et un robot qui étaient les meilleurs amis au monde. Un jour le loup avait très très faim. Le robot avait une faim de loup, mais le loup lui a dit : « Tu ne veux pas me manger quand même! » « Non, pas du tout!» « Ouf! » Ils partirent aller chercher de la viande dans une boucherie. Ils ont pris : du jambon et de la mortadelle. Ils ont pris aussi un dessert : une glace au caramel. Ils ont mangé et encore mangé. Le lendemain matin, le robot avait encore faim. Le loup était encore en train de dormir. Alors le robot s'est faufilé dans sa chambre, et lui a mangé ses deux oreilles, et puis ensuite le nez, les bras, les pieds, les jambes, le ventre... Et puis après... la tête! Le

[67] Production NORM-EC-CE2-2016-96-D1-S1931

robot était vraiment triste d'avoir mangé son meilleur ami. Maintenant le robot est le seul au monde.[68]

En général, nous avons pu remarquer que le critère adopté pour la sélection des mentions pronominales semble fonctionner sur la majorité des textes. Ceci nous permet de confirmer notre hypothèse de départ, à savoir, le fait que les scripteurs se tiennent à une progression thématique à thème constant dans la majorité des cas de productions narratives incluant plusieurs personnages. Dans le cas où ce critère a échoué, le problème pourrait être résolu avec un critère ultérieur de filtrage qui prendrait en considération le genre de l'antécédent et du pronom filtré, ou une règle qui exclurait le pronom vide qui fait partie de certaines formules (comme les formules d'incipit). De plus, nous avons observé qu'un bug du programme ne permet pas de reconnaitre certains pronoms coréférentiels bien qu'ils soient étiquetés comme *nsubj*. La correction de ce bug permettra d'améliorer les performances de l'outil, tout en gardant le même critère de sélection des pronoms déjà utilisé.

Un autre élément que nous avons pu remarquer est la présence des pronoms associés à des prépositions que n'ont pas été incluses dans les clusters, comme dans les exemples suivants :

Elle put l'emporter chez elle pour s'en servir pour ses potions magiques. Fin.[69]

Elle lui dit « D'accord, tiens » il tendit la main, elle le prend et elle l'emmène chez elle. [70]

Ces éléments sont assez fréquents dans le corpus et il faudra donc insérer ce pattern dans le système de reconnaissance développé pour le deuxième module de notre programme de manière à pouvoir prendre en compte aussi ces occurrences des pronoms de troisième personne.

---

[68] Production NORM-EC-CE2-2016-120-D1-S2439
[69] Production NORM-EC-CE2-2016-37-D1-S1560.
[70] Production NORM-EC-CE2-2016-91-D1-S949.

En dernier lieu, il faudra aussi prendre des décisions méthodologiques par rapport au traitement des pronoms personnels de troisième personne quand ils représentent un ou plusieurs des référents dont on souhaite détecter les chaînes de mentions coréférentielles.

4. Description des chaînes de coréférence : quelques statistiques sur le corpus de travail

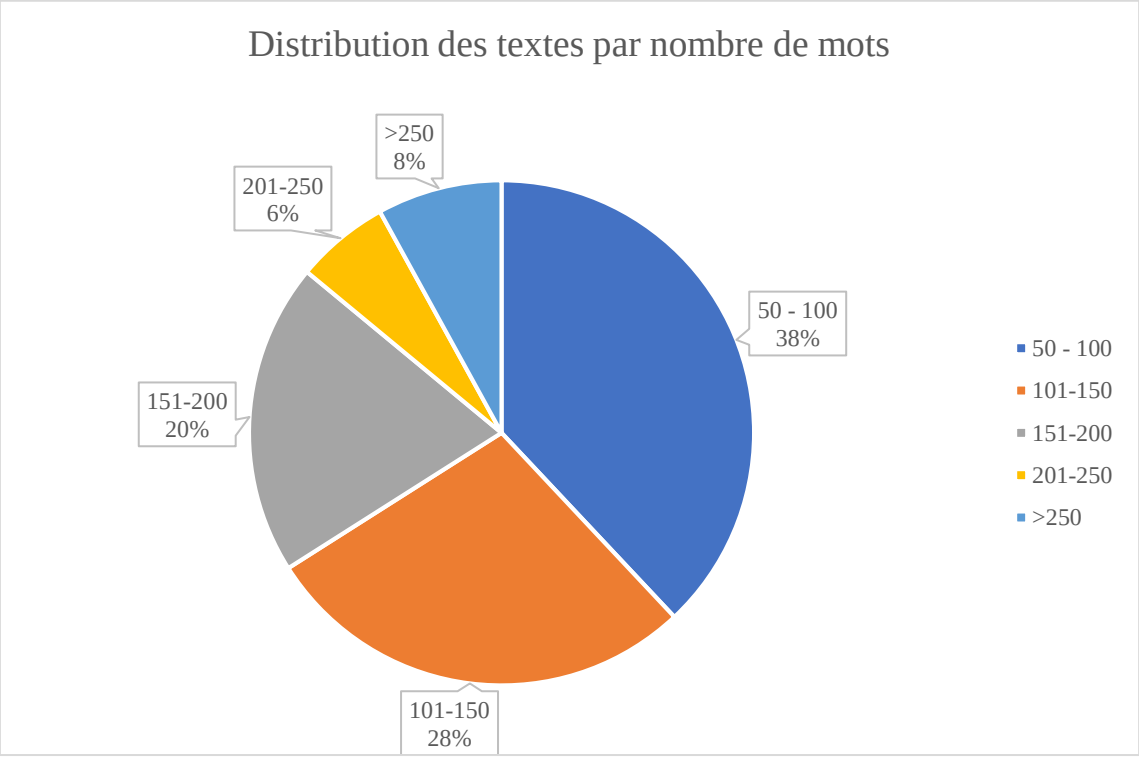
Pour donner un cadre d'orientation par rapport aux critères de définition des chaînes de coréférence (cf. Chapitre 1, 1.3) appliqués aux textes de notre corpus de travail, nous avons conduit quelques analyses quantitatives sur les caractéristiques principales des textes de notre corpus de travail et des annotations obtenues en sortie de notre programme. Sachant que ces descriptions peuvent être partiellement faussées par la persistance d'erreurs dans le fonctionnement du programme, elles donnent toutefois des indications sur la qualité de ces chaînes au sein des productions analysées.

Nous avons exploité les fonctionnalités déjà implémentées dans notre programme pour calculer la densité référentielle des chaînes de mentions des quatre personnages des histoires en nous inspirant d'une recherche sur les mentions des personnages dans des œuvres littéraires de F. Landragin (Landragin *et al.*, 2014).

Nous avions aussi commencé à annoter une partie restreinte du corpus de travail afin de pouvoir appliquer des calculs de précision et de rappel sur la base de ce corpus de référence annoté manuellement. Faute de temps, ce travail n'a pas pu être conclu et ne peut pas être utilisé pour fournir une évaluation objective de notre outil.

4.1. Longueurs des textes

Les textes qui composent notre corpus de travail ont des longueurs très variables entre le 50 et le 357 tokens[71]. Cependant, la portion des textes la plus grande par nombre de tokens, qui constitue le 38% du corpus de travail, se positionne dans le créneau des 50-100 tokens, et seulement 6% du corpus est constitué de textes qui comptent plus de 250 tokens. Nous pouvons donc affirmer que la majorité des textes comportent entre 50 et 200 tokens avec une moyenne de 133 tokens sur 50 textes.



[71] Ici nous avons pu calculer la longueur de nos textes en termes de tokens, donc des mots repérés par *SpaCy* au préalable. Cette opération de tokenisation est susceptible de présenter des erreurs, notamment sur les mots composés qui contiennent un tiret (selon nos observations empiriques). Cependant, nous avons retiré tout signe de ponctuation de ces analyses quantitatives.

4.1. Présence des personnages dans les textes et longueur moyenne des chaines par personnage

Le chat est le personnage le plus présent parmi les 50 textes. Il apparait dans 40 textes suivi par la sorcière avec 33 textes. Le loup et le robot ont des résultats similaires avec respectivement 16 et 14 occurrences.

Pour ce qui concerne les longueurs moyennes des chaînes pour chacun des personnages, nous pouvons observer qu'en moyenne, les chaînes comportant le chat semblent être les plus longues ce qui peut être est influencé par la fréquence de présence de ce personnage dans les textes.

En suivant la même méthodologie ensuite appliquée dans le calcul de la densité référentielle, donc en retirant les clusters qui contiennent zéro mention de l'analyse, nous avons calculé la moyenne des longueurs des chaînes de coréférence pour chaque personnage dans les textes.

Voici le comparatif résultant des calculs des moyennes des longueurs des chaînes sans clusters à zéro.

sorcière	chat	loup	robot
8,4242	7,175	7,5	5,5714

Tableau 7. Moyennes des longueurs des chaînes pour les quatre personnages dans tous textes du corpus.

Ces résultats nous montrent que, bien que le chat soit le personnage le plus fréquent en termes d'occurrences dans les textes, les chaînes qui indiquent la sorcière a une longueur majeure, et donc les élèves ont la tendance à utiliser plus de mentions pour décrire la sorcière dans un même texte que les autres personnages. Ce calcul est de toute manière présentée ici de manière indicative, sachant que le programme prend en considération un nombre limité des variations sur les mentions détectées, qui dépends directement de la structure du premier module. Cependant, une utilisation intéressante de cette métrique

serait de la calculer sur chaque année du corpus et ensuite de vérifier s'il y a une évolution dans la longueur moyennes des chaînes ainsi calculée tout au long des cinq ans de l'école primaire.

4.2. Densité référentielle des chaînes des personnages

Entre les différents critères de description des chaînes que nous avons mentionnés dans la description du cadre théorique (cf. Chapitre 1), nous avons mentionné le coefficient de stabilité et la densité référentielle. Nous avons décidé de ne pas calculer le coefficient de stabilité car nous aurions obtenu probablement un résultat faussé au niveau méthodologique (cf. Chapitre 1, 1.3). Ce critère descriptif se calcule sur la base du nombre de désignations différentes qu'on retrouve dans les chaînes détectées. « Plus le coefficient de stabilité est élevé, moins il y a de désignations différentes par rapport au nombre d'anaphores, et plus la stabilité référentielle est grande. » (Schnedecker, 2014 : 29). Comme c'est principalement notre programme qui établit les différentes désignations nominales détectées et incluses dans les clusters, sur la base de la liste de lemmes très restreinte que nous avons précédemment citée, nous avons décidé de ne pas appliquer ce critère d'analyse. En revanche, nous avons pu exploiter la structure de notre programme pour en tirer des données quantitatives sur la présence des personnages dans les textes, sur la longueur des chaînes qui leur sont associées, et nous avons ainsi pu facilement calculer la densité référentielle pour chacun des personnages dans les différents textes (le récapitulatif complet de ces données est disponible en Annexes).

Pour effectuer le calcul de la densité référentielle, nous nous sommes appuyés sur les résultats obtenus à l'aide de la fonction de clusterisation de notre programme (cf. Chapitre 6, 3), ce qui nous permet d'obtenir un cluster de mentions explicites des référents dans les textes pour chaque personnage. Nous avons ensuite divisé le nombre de mentions contenues dans le cluster de chaque

personnage pour la longueur totale du texte (signes de ponctuation retirés de l'analyse).

Nous pouvons ainsi observer une similarité entre les densités référentielles moyennes obtenues sur les référents sorcière et chat et la même tendance se répéter sur le couple loup-robot.

Toutefois, cette moyenne ne nous semble pas vraiment indicative des densités référentielles dans nos textes. En effet, il faut rappeler que certaines mentions nominales explicites n'ont pas été prises en compte par notre programme à cause du fait que le lemme de référence n'a pas été inclus dans la liste des lemmes détectés (cf. Chapitre 5, 5.5). Toutefois, ces mentions constituent la partie mineure des mentions nominales présentes dans les textes.

Pour corriger cette erreur, nous avons donc essayé de calculer ces moyennes en retirant de l'analyse tous les textes où le personnage n'apparait pas de manière explicite (et donc où son cluster contient 0 mention). Voici une comparaison entre les densités référentielles calculées des deux différentes manières.

sorcière		chat		loup		robot	
Moyenne avec 0	Moyenne sans 0	Moyenne avec 0	Moyenne sans 0	Moyenne avec 0	Moyenne sans 0	Moyenne avec 0	Moyenne sans 0
0,044	0,068	0,048	0,06	0,014	0,044	0,014	0,050

Tableau 8. Moyennes des densités référentielles des chaînes pour les quatre personnages dans tous les textes du corpus.

Ce calcul de la densité référentielle, déjà utilisé en littérature pour rendre compte des différences entre genres textuels différents, pourrait être ensuite appliqué sur le corpus dans sa totalité. De cette manière, nous pourrions

envisager de comparer les différences entre densités sur les différents niveaux scolaires, pour vérifier s'il y a une évolution au long du parcours d'apprentissage et une augmentation de la densité dans les textes ou si, au contraire, la densité diminue car d'autres stratégies sont déployées par les élèves pour construire la cohérence dans leurs productions écrites.

Chapitre 8. Résultats généraux et conclusions sur le fonctionnement de l'outil

Dans ce dernier chapitre, nous allons présenter les perspectives de travail ouvertes par l'utilisation de cet outil sur le corpus *Scoledit*, à la fois sur le versant de la description du phénomène de la coréférence au service de la didactique, et sur celui des possibles améliorations qu'on peut apporter à cet outil côté TAL.

Nous remarquons ici que nous envisageons d'élargir les analyses aux textes restants du niveau CE2 (nous avons en fait utilisé 50 textes sur les 373 textes du corpus de niveau CE2). Le reste du corpus pourrait être utilisé pour évaluer le modèle des règles mise en place dans notre programme et permettre ainsi d'y apporter des améliorations dont certaines déjà préconisées dans ce chapitre.

1. Perspectives de travail et améliorations à apporter à l'outil

Nous allons décrire par la suite les améliorations possibles de l'outil *DeCorScol* ressorties depuis les analyses menées sur ses résultats et leurs évaluations (cf. Chapitre 7). Ces problématiques ont été abordées dans une perspective pluridisciplinaire, avec les contributions de chercheurs et chercheuses en linguistique, didactique du français et TAL impliqués dans les recherches autours du corpus *Scoledit* et du projet *Scolinter* (cf. Chapitre 2).

1.1. Les mentions nominales

Nous avons décrit précédemment les typologies les plus fréquentes de mentions nominales présentes dans notre corpus de travail. Cependant, lors des analyses exploratoires conduites sur les résultats du module de détection de ces mentions, nous avons remarqué l'existence de mentions référentielles plus

étendues par rapport aux patterns que nous avions insérés dans notre programme (cf. Chapitre 7, 1.1).

Une autre implémentation prévue du programme, et fortement soutenue par les chercheurs en linguistique et en didactique du projet, concerne l'inclusion des déterminants possessifs dans l'annotation. Le déterminant possessif détecté sera associé à la chaîne de coréférence de l'entité possesseuse/sujet profond de l'action de possession comme déjà fait dans le cadre du projet *Democrat* (Delaborde, 2020). Toutefois, faute de temps, la détection de ce type de mentions n'a pas pu être intégrée dans notre logiciel.

Un autre aspect d'évolution est l'extension des entités détectées par le premier module en incluant des patterns plus « souples » de mentions nominales. Il s'agirait de s'intéresser toujours aux adjectifs préposés et postposés mais également aux adverbes qui les modifient et d'y inclure, comme nous venons de le proposer, les déterminants possessifs.

Dans l'exemple suivant nous pouvons observer une mention de ce type, « un loup très très mechant » : notre programme a effectivement détecté la mention de loup mais pas le syntagme nominale contenant deux adverbes et un adjectif (en jaune). Dans l'exemple suivant, le programme n'avait pas détecté un pattern du type préposition + article + nom commun + nom commun

> Et sa grand-mère arrive à la maison, elle voit un loup **très très méchant** (…)[72]
>
> le lendemain elle alla chez la maitresse sorcière qui lui avait appris les formules (…)[73]

Ces types d'erreurs du programme nous poussent à préconiser un module de détection qui puisse inclure des variations de patterns comme celles ici reportés.

[72] Production NORM-EC-CE2-2016-85-D1-S1981
[73] Production NORM-EC-CE2-2016-108-D1-S2986

1.2. Les dialogues

Les instances de dialogues directs de notre corpus de travail contiennent souvent des pronoms de première et deuxième personne que nous avons décidé, dans ce premier travail, de ne pas inclure dans les chaînes de coréférence (cf. Chapitre 5, 4.3). Faute de temps, nous n'avons pas réussi à implémenter un module d'exclusion des dialogues de nos analyses. Grâce à la présence des balises de dialogue dans la normalisation des productions écrites, nous pourrions dans un deuxième temps mettre en place une fonction qui permette d'effectuer un analyse morphosyntaxique complète de ces phrases pour fournir un cadre global de leur statut aux annotateurs humains, sans pourtant inclure les mentions qu'ils pourraient contenir dans l'étape de détection des mentions.

1.3. Le filtrage des pronoms

Bien qu'apparemment assez performant, le critère initialement choisi pour le filtrage des pronoms et leur inclusion dans des chaînes de coréférence a montré des limites lors de l'étape d'évaluation de l'outil. Dans la perspective d'apporter une amélioration à cette fonction de notre programme, nous prévoyons de corriger certaines erreurs techniques du programme, probablement issus du niveau algorithmique plutôt que du résultat de l'analyse morphosyntaxique, et d'implémenter un critère de filtrage « de niveau deux ». Le critère préconisé en ce sens est celui de filtrage des pronoms sur la base de la correspondance entre antécédent et pronom en genre et nombre lorsque le programme doit choisir parmi plusieurs propositions d'antécédents pour le même pronom.

Nous nous proposons aussi d'aborder la problématique des pronoms pluriels et de leur rattachement à une ou plusieurs chaînes coréférentielles de manière à inclure ces types de mentions dans nos études (cf. Chapitre 7, 3.).

1.4. Autres perspectives

Comme déjà mentionné lors de la description des résultats obtenus grâce au premier module de l'outil (cf. Chapitre 7, 1.1), l'absence de ressources lexicales et/ou sémantiques dans notre programme est suffisante pour une première ébauche mais présente néanmoins des limites. L'intégration d'une ressource lexicale qui puisse rendre compte de la variation lexicale des syntagmes nominaux au sein du corpus est envisageable si on souhaite améliorer ultérieurement les résultats obtenus à travers l'étape de reconnaissance des mentions nominales. Ceci est réalisable en exploitant des listes des mots créées à partir du corpus lui-même et mises en relation avec d'autres listes contenant des relations de synonymie ou hyperonymie, comme par exemple la ressource lexicale JeuxDeMots[74]. Nous souhaitons aussi, en nous appuyant sur des ressources lexicales adaptées, élargir l'analyse des chaînes de coréférences aux autres entités qui jouent les rôles de personnages des productions écrites. La mise en place d'une analyse étendue à toutes les entités dans les textes pourrait réduire le bruit causé par la présence de mentions erronément reliées aux quatre chaînes de coréférences actuelles.

En dernier lieu, nous souhaitons également exploiter la sortie en format json évoquée lors de l'explication sur la phase de prétraitement du corpus (cf. Chapitre 5, 3.3). Cette sortie couplée à une visualisation graphique devrait faciliter le travail des annotateurs experts. Cette visualisation permettra aussi de signaler quels sont les tokens insérés à partir des balises omissions (cf. Chapitre 5, 3 sur le prétraitement) qui ne font pas partie des productions originelles mais qui étaient fonctionnels à l'analyse morphosyntaxique.

[74] « JeuxDeMots est un jeu sérieux développé par le LIRMM et relevant du modèle de *game with a purpose* (ou GWAP : jeu avec un but) dont l'objet est la construction d'une base de connaissance sous la forme d'un réseau lexical. » (« *JeuxDeMots* », 2021). Les ressources sémantiques créées à partir de ce jeux ont une structuration très simple, ce qui nous fais envisager d'avantage leurs utilisation.

Conclusion

Grâce à ce travail préliminaire de détection des chaînes de coréférence sur un corpus de travail composé de textes issus du niveau CE2, faisant partie du corpus *Scoledit*, nous avons pu confirmer certaines des hypothèses faites lors de l'exploration du corpus et de la modélisation des éléments de langage faisant partie des chaînes de coréférence.

Notre hypothèse de départ sur la possibilité de créer un outil de résolution de coréférences à base de règles sur les écrits scolaires de niveau CE2, fondé sur des structures simples et sur le vocabulaire mobilisé par les élèves, a été partiellement confirmée par les résultats obtenus par le système *DeCorScol*. Tout en remarquant que l'outil présente encore des marges d'amélioration, les résultats obtenus jusqu'à présent sont assez prometteurs du point de vue de l'assistance à l'annotation manuelle.

Une opération que nous préconisons est d'élargir l'utilisation de l'outil à tous les textes de CE2 présents dans le corpus, de manière à pouvoir conduire une évaluation plus complète de l'outil, et éventuellement de pouvoir affiner le modèle de règles utilisé jusqu'à présent. Nous avons effectivement à disposition encore plus de 300 textes sur lesquels l'outil n'a pas été appliqué.

Nous pouvons aussi remarquer que les outils actuellement disponibles en TAL permettent vraiment de mettre en place un programme de pré-annotation de ce phénomène quasi totalement à base de règles. Néanmoins, il faut ici spécifier que le modèle exploité par l'analyseur morphosyntaxique est un

modèle entrainé à l'aide d'un réseau de neurones profond, ce qui pourrait qualifier notre outil d'hybride plutôt que totalement à base de règles[75].

En conclusion, cet outil pourra probablement répondre au besoin d'annotation du corpus *Scoledit* en relations de coréférence. Ces données vont ensuite permettre de fournir des descriptions de ce phénomène sur les différents niveaux de scolarité et sur leurs évolutions. Ces connaissances permettront de pointer les réussites et les difficultés persistantes des élèves dans le domaine de la cohérence textuelle permettant de faire évoluer par là même les démarches didactiques.

[75] Plus d'informations sur le modèle de transformeur utilisé par *SpaCy* sont disponibles au lien suivant https://*SpaCy*.io/usage/embeddings-transformers.

Bibliographie

Bagga, A., & Baldwin, B. (1998). Algorithms for scoring coreference chains. *The first international conference on language resources and evaluation workshop on linguistics coreference*, *1*, 563-566.

Bonnemaison, K. P. (2018). *Anaphore et référence en production écrite : Étude de textes narratifs d'élèves de 9 à 11 ans, du CE2 au CM2* [Phdthesis, Université Toulouse le Mirail - Toulouse II]. https://tel.archives-ouvertes.fr/tel-02627042

Boré, C., & Elalouf, M.-L. (2017). Deux étapes dans la construction de corpus scolaires : Problèmes récurrents et perspectives nouvelles. *Corpus*, *16*, Article 16. https://doi.org/10.4000/corpus.2731

Boré, C., Roubaud, M.-N., & Elalouf, M.-L. (2018). *Corpus ÉMA, écrits scolaires*. https://hdl.handle.net/11403/ema-ecrits-scolaires-1/v2

Cai, J., & Strube, M. (2010). Evaluation Metrics For End-to-End Coreference Resolution Systems. *Proceedings of the SIGDIAL 2010 Conference*, 28-36. https://aclanthology.org/W10-4305

Carbonneau, C., & Préfontaine, C. (2005). Enseigner et évaluer la cohérence textuelle. *Québec français*, *138*, 78-81.

Charolles, M. (1988a). La gestion des risques de confusion entre personnages dans une tâche rédactionnelle. *Pratiques*, *60*(1), 75-97. https://doi.org/10.3406/prati.1988.1498

Charolles, M. (1988b). Les plans d'organisation textuelle : Périodes, chaînes, portées et séquences. *Pratiques*, *57*(1), 3-13. https://doi.org/10.3406/prati.1988.1468

Charolles, M. (2001). Référents évolutifs et évolution de la référence. In *Les référents évolutifs entre linguistique et philosophie* (p. 39-97). Klincksieck. https://hal.archives-ouvertes.fr/hal-01404846

Charolles, M. (2002). *La référence et les expressions référentielles en français*.

Charolles, M. (2011). Cohérence et cohésion du discours. In K. H. ; C.Marello (Éd.), *Dimensionen der Analyse Texten und Diskursivent—Dimensioni dell'analisi di testi e discoursi* (p. 153-173). Lit Verlag. https://hal.archives-ouvertes.fr/hal-00665838

Chastain, C. (1975). *Reference and Context*. http://conservancy.umn.edu/handle/11299/185224

Codebase. (2022). In *Wikipédia*. https://fr.wikipedia.org/w/index.php?title=Codebase&oldid=189951303

Corblin, F. (1985). Les chaînes de référence : Analyse linguistique et traitement automatique. *Intellectica*, *1*(1), 123-143. https://doi.org/10.3406/intel.1985.851

Corblin, F. (1995). *Les formes de reprise dans le discours. Anaphores et chaînes de référence*. Presses Universitaires de Rennes. https://jeannicod.ccsd.cnrs.fr/ijn_00550962

Delaborde, M. (2020). *Analyse en corpus de chaînes de coréférence : La coréférence non-stricte à l'épreuve de la linguistique outillée*.

Denis, P., & Baldridge, J. (2009). *Global joint models for coreference resolution and named entity classification*. http://rua.ua.es/dspace/handle/10045/10549

Dinarelli, M., & Grobol, L. (2019, juillet). Modèles neuronaux hybrides pour la modélisation de séquences : Le meilleur de trois mondes. *TALN-RECITAL 2019 - 26ème Conférence sur le Traitement Automatique des Langues Naturelles*. https://hal.archives-ouvertes.fr/hal-02157160

Doquet, C. (2020). Analyser linguistiquement l'écriture à l'école : EcriScol, un corpus génétique. In *CLUB Working Papers in Linguistics Volume 4* (Vol. 4, p. 127-140). https://hal.archives-ouvertes.fr/hal-02883152

Doquet, C., Enoiu, V., Fleury, S., & Maziotti, S. (2017). Problèmes posés par la transcription et l'annotation d'écrits d'élèves. *Corpus, 16*, Article 16. https://doi.org/10.4000/corpus.2776

Doquet, C., & Ponton, C. (2021). Écrire de l'école à l'université : Corpus, traitements, analyses outillées. Présentation. *Langue francaise, 211*(3), 11-20.

Doquet, C., Revelli, L., & Moysan, A. (2021). Écriture et forme scolaire : Spécificités de transcription et de traitement. *Langue francaise, 211*(3), 21-36.

Doquet, C., Silvia, F., Fleury, S., Ho-Dac, L.-M., Mazziotti, S., Moysan, A., & Ponton, C. (2019). *The É:Calm Resource : Transcription, Encoding and Annotation of Handwritten Manuscripts produced by French Pupils and Students*.

Elalouf, M.-F. (2005). Enseigner à écrire entre 10 et 14 ans : Un corpus, des analyses, des repères pour l'enseignement. *Scérén & CRDP Académie de Versailles*. https://www.persee.fr/doc/airdf_1776-7784_2005_num_37_2_1676

Elalouf, M.-L., & Boré, C. (2007). Construction et exploitation de corpus d'écrits scolaires. *Revue Française de Linguistique Appliquée, XII*(1), 53. https://doi.org/10.3917/rfla.121.0053

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*. https://hal.archives-ouvertes.fr/hal-03161554

Facts & Figures · spaCy Usage Documentation. (s. d.). Facts & Figures. Consulté 27 avril 2022, à l'adresse https://spacy.io/usage/facts-figures

Fan, F.-L., Xiong, J., Li, M., & Wang, G. (2021). On Interpretability of Artificial Neural Networks : A Survey. *IEEE Transactions on Radiation and Plasma Medical Sciences, 5*(6), 741-760. https://doi.org/10.1109/TRPMS.2021.3066428

Federzoni, S., Ho-Dac, L.-M., & Fabre, C. (2021). Coreference Chains Categorization by Sequence Clustering. In A. for C. Linguistics (Éd.), *2nd Workshop on Computational Approaches to Discourse* (p. 52-57). https://hal.archives-ouvertes.fr/hal-03513356

Garcia-Debanc, C., & Bonnemaison, K. (2014a). La gestion de la cohésion textuelle par des élèves de 11-12 ans : Réussites et difficultés. *SHS Web of Conferences, 8*, 961-976. https://doi.org/10.1051/shsconf/20140801349

Garcia-Debanc, C., & Bonnemaison, K. (2014b). La gestion de la cohésion textuelle par des élèves de 11-12 ans : Réussites et difficultés. *SHS Web of Conferences, 8*. https://doi.org/10.1051/shsconf/20140801349

Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M., & Rebeyrolle, J. (2017). Vers l'annotation discursive de textes d'élèves. *Corpus, 16*, Article 16. https://doi.org/10.4000/corpus.2783

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.

Goigoux, R., Jarlégan, A., & Piquée, C. (2015). Évaluer l'influence des pratiques d'enseignement du lire-écrire sur les apprentissages des élèves : Enjeux et choix méthodologiques. *Recherches en didactiques, 19*(1), 9-37.

GPT-3. (2022). In *Wikipédia*. https://fr.wikipedia.org/w/index.php?title=GPT-3&oldid=193313041

Grobol, L. (2020). *Coreference resolution for spoken French* [Phdthesis, Université Sorbonne Nouvelle - Paris 3]. https://hal.archives-ouvertes.fr/tel-02928209

Grobol, L. (2021). Exploitation du corpus Democrat par apprentissage artificiel. *Langages*, *224*. https://hal.archives-ouvertes.fr/hal-03475070

Haghighi, A., & Klein, D. (2009). Simple Coreference Resolution with Rich Syntactic and Semantic Features. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 1152-1161. https://aclanthology.org/D09-1120

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006). OntoNotes : The 90% Solution. *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 57-60. https://aclanthology.org/N06-2015

Jacques, M.-P., & Rinck, F. (2017). Un « corpus de littéracie avancée : Résultat et point de départ. *Corpus, 16*, Article 16. https://doi.org/10.4000/corpus.2806

JeuxDeMots. (2021). In *Wikipédia*. https://fr.wikipedia.org/w/index.php?title=JeuxDeMots&oldid=183383812

Landragin, F. (2021). Le corpus Democrat et son exploitation. Présentation. *Langages*, *224*, 11-24.

Landragin, F., Tanguy, N., & Charolles, M. (2014). Références aux personnages dans L'Occupation des sols : Apport de la linguistique outillée. *Revue Sciences/Lettres*. https://doi.org/10.4000/rsl.816

Lion-Bouton, A., Grobol, L., Antoine, J.-Y., Billot, S., & Lefeuvre-Halftermeyer, A. A. (2020). Comment arpenter sans mètre : Les scores de résolution de chaînes de coréférences sont-ils des métriques ? In G. Adda, M. Amblard, & K. Fort (Éds.), *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)* (p. 10-18). ATALA. https://hal.archives-ouvertes.fr/hal-02750222

Longo, L. (2013). *Vers des moteurs de recherche « intelligents » : Un outil de détection automatique de thèmes : méthode basée sur l'identification automatique des chaînes de référence* [These de doctorat, Strasbourg]. http://www.theses.fr/2013STRAC041

Longo, L., & Todirascu, A. (2009). UNE ÉTUDE DE CORPUS POUR LA DÉTECTION AUTOMATIQUE DE THÈMES. *Texte et Corpus.*, *4*, 143.

MEN. (2015). *Programmes d'enseignement de l'école élémentaire et du collège* (Bulletin officiel spécial N° 386; p. 386). Ministère de l'Éducation Nationale.

http://www.education.gouv.fr/cid95812/au-bo-special-du-26-novembre-2015-programmes-d-enseignement-de-l-ecole-elementaire-et-du-college.html

MUC Consortium. (1995a). Appendix D : Coreference Task Definition (v2.3). *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*. MUC 1995, San Francisco, CA, USA. https://aclanthology.org/M95-1025

MUC Consortium. (1995b). Sixth Message Understanding Conference (MUC-6) : Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995. *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*. MUC 1995, San Francisco, CA, USA. https://aclanthology.org/M95-1000

Nedoluzhko, A., Novák, M., Popel, M., Žabokrtský, Z., & Zeman, D. (s. d.). *Coreference meets Universal Dependencies – a pilot experiment on harmonizing coreference datasets for 11 languages ÚFAL Technical Report*. 73.

Nedoluzhko, A., Novák, M., Popel, M., Žabokrtský, Z., & Zeman, D. (2021). Coreference in Universal Dependencies 0.2 (CorefUD 0.2). *https://ufal.mff.cuni.cz/corefud*. https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-4598

Newell, A. (1990). *Unified theories of cognition* (p. xvii, 549). Harvard University Press.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry : Symbols and search. *Communications of the ACM*, *19*(3), 113-126. https://doi.org/10.1145/360018.360022

Oberle, B. (2017). *ODACR (Outil de Détection Automatique des Chaînes de Référence)*. https://hal.inria.fr/hal-01837101

Obry, V., Glikman, J., Guillot-Barbance, C., & Pincemin, B. (2017). Les chaînes de référence dans les récits brefs en français : Étude diachronique (XIIIe – XVIe s.). *Langue française, 2017/3*(195), 91-110. https://doi.org/10.3917/lf.195.0091

O'Connor, B. (2021). *ARKref* [Java]. https://github.com/brendano/arkref (Original work published 2014)

O'Connor, B., & Heilman, M. (2013). ARKref : A rule-based coreference resolution system. *arXiv:1310.1975 [cs]*. http://arxiv.org/abs/1310.1975

Partee, B. H. (1970). Opacity, Coreference, and Pronouns. *Synthese*, *21*(3/4), 359-385.

Perret, M. (2000). Quelques remarques sur l'anaphore nominale aux XIVe et XVe siècles. *L'information grammaticale*, *87*(1), 17-23. https://doi.org/10.3406/igram.2000.2740

Poesio, M. (2016). Linguistic and Cognitive Evidence About Anaphora. In M. Poesio, R. Stuckardt, & Y. Versley (Éds.), *Anaphora Resolution : Algorithms, Resources, and Applications* (p. 23-54). Springer. https://doi.org/10.1007/978-3-662-47909-4_2

Poesio, M., Stuckardt, R., & Versley, Y. (2016). *Anaphora Resolution : Algorithms, Resources, and Applications*. https://doi.org/10.1007/978-3-662-47909-4

Pons, M. (2019). *Étude des incipit de productions écrites d'élèves : Comparaison des relations de référence dans des textes rédigés par des élèves de CE2 et de 3e* (p. 130).

Ponton, C. (2018, juin). Constituer et analyser un corpus scolaire. L'approche Scoledit. *3ème école de Printemps Petale*. https://hal.univ-grenoble-alpes.fr/hal-01910629

Ponton, C., Gutiérrez-Cáceres, R., Teruggi, L., Farina, E., Brissaud, C., & Wolfarth, C. (2021). SCOLINTER : Un corpus trilingue. L'exemple de la segmentation en mots. *Langue francaise*, *211*(3), 37-50.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza : A Python Natural Language Processing Toolkit for Many Human Languages. *arXiv:2003.07082 [cs]*. http://arxiv.org/abs/2003.07082

Quignard, M., Mene, M. le, & Landragin, F. (2021). Elaboration du corpus Democrat : Procedures d'annotation et d'evaluation. *Langages*, *224*, 25-46.

Recasens, M. (2010). *Coreference : Theory, Annotation, Resolution and Evaluation* [Universitat de Barcelona]. stel.ub.edu/cba2010/phd/phd.pdf

Recasens, M., & Hovy, E. (2011). BLANC : Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, *17*(4), 485-510. https://doi.org/10.1017/S135132491000029X

Revelli, L. (2011). *Scritture scolastiche dall'unità d'Italia ai giorni nostri : Studi e testimonianze*. Aracne.

Rondelli, F. (2010). La cohérence textuelle : Pratiques des enseignants et théories de référence. *Pratiques. Linguistique, littérature, didactique*, *145-146*, 55-84. https://doi.org/10.4000/pratiques.1505

Rosenblatt, F. (1957). *The perceptron—A perceiving and recognizing automaton* (Nº 85-460-1). Cornell Aeronautical Laboratory.

Savy, R., Alfano, I., Crocco, C., & Capitanio, S. (2012). *Corpus DILS (Dialoghi in Italiano Lingua Straniera)*. https://parlaritaliano.studiumdipsum.it/index.php/it/corpora-di-parlato/794-corpus-dils-dialoghi-in-italiano-lingua-straniera

Schnedecker, C. (1997). *Nom propre et chaînes de référence* (Klincksieck, Éd.). Librairie KLINCKSIECK. https://hal.archives-ouvertes.fr/hal-00808797

Schnedecker, C. (2005). Les chaînes de référence dans les portraits journalistiques : Éléments de description. *Travaux De Linguistique*, *51*. https://doi.org/10.3917/tl.051.0085

Schnedecker, C. (2014). Chaînes de référence et variations selon le genre. *Langages*, *195*(3), 23-42.

Schnedecker, C. (2019). De l'intérêt de la notion de chaîne de référence par rapport à celles d'anaphore et de coréférence. *Cahiers de praxématique*, *72*, Article 72. https://doi.org/10.4000/praxematique.5339

Schnedecker, C., & Longo, L. (2012). Impact des genres sur la composition des chaînes de référence : Le cas des faits divers. In *SHS Web of Conferences* (Vol. 1). https://doi.org/10.1051/shsconf/20120100061

Stuckardt, R. (2016). Introduction. In M. Poesio, R. Stuckardt, & Y. Versley (Éds.), *Anaphora Resolution : Algorithms, Resources, and Applications* (p. 1-19). Springer. https://doi.org/10.1007/978-3-662-47909-4_1

Studi e ricerche del progetto CoDiSSc. (s. d.). Consulté 28 décembre 2021, à l'adresse http://www.codissc.it/studi-codissc-ricerche-codissc-pubblicazioni-scientifiche

Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2020). Anaphora and coreference resolution : A review. *Information Fusion*, *59*, 139-162. https://doi.org/10.1016/j.inffus.2020.01.010

Sun, X., Yang, D., Li, X., Zhang, T., Meng, Y., Qiu, H., Wang, G., Hovy, E., & Li, J. (2021). Interpreting Deep Learning Models in Natural Language Processing : A Review. *arXiv:2110.10470 [cs]*. http://arxiv.org/abs/2110.10470

Tauveron, C. (1995). *Le Personnage : Une clef pour la didactique du récit à l'école élémentaire* (Delachaux&Niestlé).

The ARKref Noun Phrase Coreference System. (s. d.). Consulté 16 avril 2022, à l'adresse http://www.cs.cmu.edu/~ark/ARKref/

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L. (1995). A Model-Theoretic Coreference Scoring Scheme. *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*. MUC 1995. https://aclanthology.org/M95-1005

Weischedel, Ralph, Palmer, Martha, Marcus, Mitchell, Hovy, Eduard, Pradhan, Sameer, Ramshaw, Lance, Xue, Nianwen, Taylor, Ann, Kaufman, Jeff, Franchini, Michelle, El-Bachouti, Mohammed, Belvin, Robert, & Houston, Ann. (2013). *OntoNotes Release 5.0* (p. 2806280 KB) [Data set]. Linguistic Data Consortium. https://doi.org/10.35111/XMHB-2B84

Widlöcher, A., & Mathet, Y. (2012). The Glozz Platform : A Corpus Annotation and Mining Tool. *Proceedings of the 2012 ACM Symposium on Document Engineering*, 171-180.

Wilkens, R., Oberle, B., Landragin, F., & Todirascu, A. (2020). French Coreference for Spoken and Written Language. *Proceedings of the 12th Language Resources and Evaluation Conference*, 80-89. https://aclanthology.org/2020.lrec-1.10

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, *3*(1), 191-191. https://doi.org/10.1016/0010-0285(72)90002-3

Wolfarth, C. (2015). *Apport du TAL à la constitution et l'exploitation d'un corpus scolaire de cours préparatoire*.

Wolfarth, C. (2019). *Apport du TAL à l'exploitation linguistique d'un corpus scolaire longitudinal* [Phdthesis, Université Grenoble Alpes]. https://tel.archives-ouvertes.fr/tel-02517320

Wolfarth, C., Brissaud, C., & Ponton, C. (2018). *Transcrire et normer un corpus scolaire, pour quelles analyses ?* 25.

Wolfarth, C., Ponton, C., & Totereau, C. (2017). Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire au travers du développement d'un outil d'annotation orthographique. *Corpus*, *16*. https://doi.org/10.4000/corpus.2796

Zhang, Y., Tiňo, P., Leonardis, A., & Tang, K. (2021). A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *5*(5), 726-742. https://doi.org/10.1109/TETCI.2021.3100641

Annexes

Annexe 1. Consigne utilisée pour le corpus RésolCo

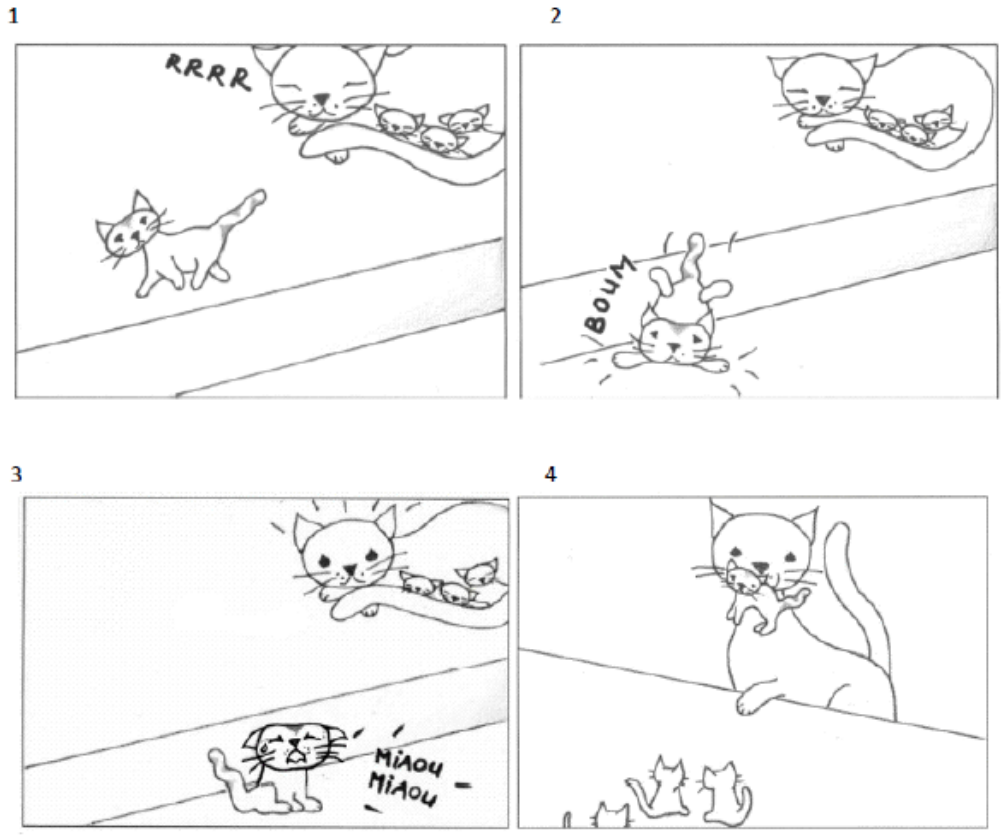
Raconte une histoire dans laquelle tu insèreras séparément et dans l'ordre donné les trois phrases suivantes :

P1 - Elle habitait dans cette maison depuis longtemps.

P2 - Il se retourna en entendant ce grand bruit.

P3 - Depuis cette aventure, les enfants ne sortent plus la nuit.

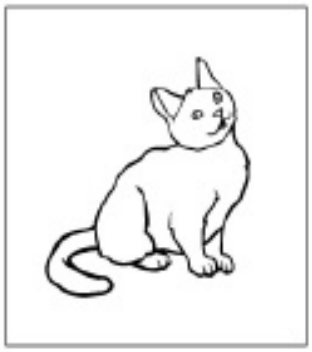
Annexe 2. Consignes utilisées pour le corpus Scoledit : images utilisées pour la production écrite en CP.



Annexe 3. Consignes utilisées pour le corpus Scoledit : images présentées aux élèves lors de la production écrite en CE1, CE2, CM1 et CM2.

1

2



3

4

Voici 4 personnages. Choisis un ou deux personnages et raconte une histoire.

Entoure le ou les personnages que tu as choisis.

Annexe 4. Sortie du programme DeCorScol sur les 50 textes du corpus de travail de CE2 (format de visualisation web)

Production NORM-EC-CE2-2016-102-D1-S212

Il était une fois une sorcière . elle avait un chat magique, le chat avait le pouvoir de changer de couleur. La maison de la sorcière Gigi était noire donc vu que le chat Victor pouvait changer de couleur, Gigi tomba mais un jour elle a dit « j'en ai marre oust. » Un jour un robot est entré dans la maison de Gigi et rentra dans l'ordre. Victor n'avait plus de pouvoir, Gigi arrêtait de tomber et le robot s'en alla.

Production NORM-EC-CE2-2016-102-D1-S214

Maitre chat et son robot nettoyeur Maitre chat a un robot, il nettoie tout, même le grenier. Le robot s'appela Bobis. Et un jour Bobis est mort. Puis le chat devint homme de ménage puis il travailla pour d'autres foyers.

Production NORM-EC-CE2-2016-104-D1-S831

Le chat des bois Il était une fois, à la forêt de "Touffubranche", un bûcheron, sa femme et leur chat . Ils vivaient assez bien, dans une petite chaumière à la lisière de la forêt. Ils se nourrissaient de fruits sauvages, de plantes. Ils avaient même réussi à travailler un peu de terre à côté de leur maison afin de pouvoir y planter des légumes. Ils étaient heureux. Mais vint un hiver, où la misère s'abattit : il faisait trop froid et il n'y eut plus de nourriture dans la forêt. Et pour le comble du malheur les légumes pourrirent. Ils n'avaient même plus de quoi nourrir leur chat . Le plus dur était arrivé : il fallait abandonner leur matou pour survivre. Le pauvre animal, tremblant de froid, réussit heureusement à trouver un trou de chouette pour passer la nuit. Il s'y endormit. Le lendemain, il sortit de son trou et partit à la recherche de nourriture. Il vit une maison ; la porte était ouverte. Il y entra. C'était la maison d'un riche très avare. Mais il se glaça d'horreur : ses pattes cédaient sous lui! Il était tombé à la cave! Et elle était remplie de diamants! Il en prit dans sa gueule et les rapporta à ses maîtres. Ils purent enfin vivre tranquilles. Fin.

Production NORM-EC-CE2-2016-104-D1-S835

Il était une fois une sorcière et un chat . le chat venait toujours chez la sorcière un moment. la sorcière en avait marre. du coup la sorcière lui a fait peur avec sa baguette mais le sort s'est fait sans faire exprès. le lendemain matin, le chat est devenu un loup . il est du coup chez la sorcière mais la sorcière n'était pas là. il attend, il attend, il attend. la sorcière était revenue. le loup rentra chez elle. la sorcière lança le sort. il devint un robot . encore un sort et il est redevenu un chat . tout est bien qui finit bien. Fin.

Production NORM-EC-CE2-2016-104-D1-S841

C'est l'histoire d'un chat avec sa maitresse qu'il aimait beaucoup. Un jour ils se baladaient, le petit chat partit explorer les environs. il sautait, grimpait aux arbres et galopait si loin qu'il s'était perdu. à un moment donné il vit une grotte il rentra. tout d'un coup ce petit chat perdu vit une sorcière laide mais très gentille et elle lui dit " Que fais-tu là mon ami? ". Le chat répondit tremblant " Je me suis perdu. pouvez-vous m'aider s'il vous plait ". " Oui bien sûr tu es si mignon ". " Merci ". Ils étaient devenus les meilleurs amis du

monde. Ils cherchaient, cherchaient. le petit chaton perdu était désespéré. La sorcière lui dit " Ne t'inquiète pas nous allons trouver mais tout de suite rentrons. il commence à faire nuit et à pleuvoir. Le lendemain ils cherchaient encore et encore, tout d'un coup le petit chat qui avait une bonne vue vit sa maison, par la fenêtre il voyait sa maitresse pleurer. il rentra et elle sauta de joie. la sorcière était triste alors on lui dit de venir vivre ici. Fin

Production NORM-EC-CE2-2016-104-D1-S855

Dans une forêt il y avait une maison où habitaient une sorcière et son chat . Son chat était noir avec des yeux perçants qui faisaient peur à un chien. Et la sorcière ce n'était pas mieux, elle attrapait un enfant et " hop " elle le mettait directement dans une cage avec des serpents, des asticots... et elle fermait la cage à double tour. comme ça l'enfant ne pouvait pas s'échapper. sinon elle le mettait directement dans la marmite. Mais un jour, elle attrapa un enfant et elle le mit dans la cage. elle voulait aller chercher un enfant à la ville qui était très très très loin. elle était bien pressée, du coup elle oublia de fermer la cage et du coup elle avait demandé à son chat de surveiller la cage mais le chat a vu une souris passer, du coup le chat la poursuit. le chat ouvre la porte et poursuit la souris. du coup le petit enfant sort de sa cage et comme la sorcière avait laissé les clés le petit enfant ferme la porte à double tour et la sorcière avait laissé sa baguette, elle jette un sort qui bloque la porte. Et ensuite la sorcière revient avec 3 enfants. Mais elle ne peut pas rentrer. du coup elle pose le sac où il y a l'enfant et les 3 enfants s'échappent et le petit enfant qui est dans la maison appelle les 3 autres pour venir. Puis la sorcière voit son chat revenir avec une souris alors la sorcière dit à son chat « t'aurais dû la surveiller, j'avais laissé la cage ouverte ».

Production NORM-EC-CE2-2016-108-D1-S2972

Il était une fois une petite chatte qui s'appelait Boubou, on l'avait appelée Boubou parce qu'elle adorait faire des blagues. Mais un jour, cette petite farceuse a eu l'idée de dire une blague méchante aux policiers. Mais en plus qu'elle était toute petite elle ne savait pas qu'elle pouvait aller en prison. Mais quand elle vit les habits du monsieur tomber elle vit que c'était un robot qui lui dit « C'est carnaval aujourd'hui, viens je vais te déguiser ». Et ils firent la fête toute la journée, et depuis ce jour le chat et le robot ne se quittèrent plus. « FIN »

Production NORM-EC-CE2-2016-108-D1-S2976

Il était une fois un chat et une sorcière qui vivaient dans une maison noire et le chat quand il ferma les yeux la sorcière Crapouille s'assit sur lui et il faisait miaou miaou. Crapouille se levait vite, elle dit : « pardon pardon, je suis désolée petit chat ».

Production NORM-EC-CE2-2016-108-D1-S2978

La sorcière et son chat Il y avait une fois une sorcière appelée Camille qui vivait heureuse avec son chat mignon, qui s'ennuyait. La petite magicienne eut un jour une visite imprévue qui parlait d'école pour sorcière, Camille très mécontente eut la mauvaise idée de répondre non. Le jeune homme très fâché, alors qu'il ne croyait pas avoir une réponse pareille préféra partir. Camille entièrement fâchée jetait un sort à son adorable minou qui doit maintenant supporter des poils bleus. Si la visite n'aurait pas eu lieu le pauvre mignon ne serait pas en colère contre sa maitresse. Trois jours plus tard la magicienne en bon état rendait sa couleur à son chaton (marron clair, les yeux bleus). Depuis ce jour le chat et la sorcière sont heureux. FIN.

Production NORM-EC-CE2-2016-108-D1-S2981

Il était une fois une sorcière qui habitait seule avec son balai qui faisait le ménage. Un jour elle eut l'idée de construire un robot pour l'accompagner. Mais ce jour-là Bot le robot n'arrêtait pas de tout casser. Plus tard elle jeta un sort pour que l'engin mécanique parte d'ici mais ça n'avait pas marché. Alors elle l'a pris son ami et l'a balancé dehors. Le lendemain le robot cassa la porte pour tout réparer ses bêtises. Et il alla réveiller Marianne la sorcière pour se faire pardonner. Depuis ce jour ils faisaient des jeux de cartes et des blagues, même une fois le balai a pris une photo de Bot et la vieille en train de cuisiner.

Production NORM-EC-CE2-2016-108-D1-S2986

Il était une fois une sorcière qui vivait avec un chat qui s'appelait Noisaitou et la sorcière s'appelait Rutabagae . Un jour elle s'entraina à faire plein de tours de magie mais elle ne retenait pas les formules qu'elle avait apprises. tout le temps elle essayait mais elle faisait tout exploser. le lendemain elle alla chez la maitresse sorcière qui lui avait appris les formules. elle lui faisait répéter, elle lui faisait voir des tours de magie, elle lui faisait recommencer les tours de magie qu'elle a faits. Le matin elle essaya et elle réussit. elle voulut que son chat vole « abracadabra que tu voles tout de suite » et le chat vole et elle dit la formule et le chat tomba dans les bras de la sorcière et elle était heureuse d'avoir des pouvoirs.

Production NORM-EC-CE2-2016-108-D1-S2988

Il était une fois une sorcière et son chat qui vivaient dans la ville de Pastise où il n'y avait que des sorciers et des sorcières . un jour le voisin de la vieille va transformer le chat en petit minou méchant et du coup le petit minou griffa la vieille. c'est devenu un gros désastre alors la petite dame fait une potion pour tuer son chat . elle attendait une semaine pour pouvoir tuer son chat . le minou fut mort et la dame fut triste mais elle est si vieille qu'elle meurt vieille et triste.

Production NORM-EC-CE2-2016-115-D1-S1345

Il était une fois une sorcière et son chat qui vivaient dans la forêt, tous les jours la sorcière demandait à son chat d'aller chercher des myrtilles dans la forêt pour faire un gâteau à la myrtille. Un jour le chat alla chercher des myrtilles, il fallait traverser la rivière, le chat traversa la rivière. quand il fut arrivé de l'autre côté un monstre l'attrapa. Au bout d'une heure la sorcière commença à s'inquiéter donc elle alla le chercher, elle suivit ses empreintes jusqu'à la rivière. Le chat avait très peur du monstre car il voulait le manger. La sorcière arriva devant la maison du monstre, elle entra dans la maison puis elle vit son chat dans une cage. la sorcière savait que le chat n'aimait pas les cages. Pendant que le monstre n'était pas là la sorcière sortit le chat de la cage puis ils rentrèrent tous les deux ensemble à la maison. Le monstre prit une colère horrible.

Production NORM-EC-CE2-2016-117-D1-S3003

La sorcière maléfique Il était une fois une sorcière qui s'appelait Carabistouie. Aujourd'hui la sorcière faisait sa soupe quand tout d'un coup elle entendit de la porte un bruit : « toc-toc-toc ». La sorcière répondit : « Entrez » un robot entre dans la maison et lui demande : « Bonjour madame, aurez-vous besoin d'une personne pour nettoyer votre

chaudière? » la sorcière dit « Oh oui! Mets-toi au travail tout de suite, pendant ce temps je vais faire un gros dodo » et le robot dit « D'après ce que j'ai vu et entendu c'est une sorcière, il faut que je m'en débarrasse et demain plus de sorcière ». Le lendemain matin la sorcière continua sa soupe : « bisacarbo lata noutou pourtar ext » le robot qui nettoyait le sol se rapprochait tout doucement et il la poussa brusquement dans la soupe et dit : « Voilà vilaine sorcière, plus de magie et surtout plus de sorcellerie et au moins tu ne me traiteras plus comme une boniche et à moi la maison ». 1 an plus tard le robot s'était bien installé dans la maison et vit que la sorcière était très riche, le robot avait trouvé la cachette où il y avait plein de pièces donc le robot alla à l'école des robots pour s'améliorer dans sa conjugaison et il se fit plein d'amis et une amoureuse. Fin.

Production NORM-EC-CE2-2016-120-D1-S2439

Il était une fois un loup et un robot qui étaient les meilleurs amis au monde. Un jour le loup avait très très faim. Le robot avait une faim de loup, mais le loup lui a dit : « Tu ne veux pas me manger quand même! » « Non, pas du tout! » « Ouf! » Ils partirent aller chercher de la viande dans une boucherie. Ils ont pris : du jambon et de la mortadelle. Ils ont pris aussi un dessert : une glace au caramel. Ils ont mangé et encore mangé. Le lendemain matin, le robot avait encore faim. Le loup était encore en train de dormir. Alors le robot s'est faufilé dans sa chambre, et lui a mangé ses deux oreilles, et puis ensuite le nez, les bras, les pieds, les jambes, le ventre... Et puis après... la tête! Le robot était vraiment triste d'avoir mangé son meilleur ami. Maintenant le robot est le seul au monde.

Production NORM-EC-CE2-2016-120-D1-S2453

La Sorcière et son chat Une sorcière veut que son chat parle. Elle fait la potion mais il lui manque 1 ingrédient. Donc elle demande à son chat d'aller le chercher et elle a dit : « si tu ne le ramèneras pas avant 2 jours je te tuerai! » Donc le chat se mit tout de suite. Il ne le ramène pas tout de suite et la sorcière allait tuer son chat . Mais elle se dit « Comment elle va faire pour aller chercher les ingrédients? »

Production NORM-EC-CE2-2016-126-D1-S1826

La sorcière et son chat qui parle Aujourd'hui la sorcière et son chat Griboille préparent une potion. La potion terminée la sorcière goûta la potion. « Goûte-moi ça Graboille. » Il était si apeuré qu'il sauta du haut de son arbre à chat sur l'étagère. Elle grince crac! crac! puis tombe. Plouf dans le chaudron. En sortant il dit : « Miaou » la sorcière entend « C'est chaud » « il parle » la sorcière le replonge dedans, il se tait. Depuis il guette à tout sans exception jusqu'à sa mort.

Production NORM-EC-CE2-2016-126-D1-S1830

La sorcière et le chat Il était une fois une sorcière qui s'appelait Mademoiselle Cascou et un chat qui s'appelait Minou. Un jour Minou se promenait tout seul dans la rue et il rencontra Mademoiselle Cascou. ils étaient pires ennemis parce que Mademoiselle Cascou transformait toujours Minou en grenouille, en fourmi, en abeille, en cafard... et plein d'autres choses. Heureusement après il arrivait à se remettre en chat . mais un jour il n'arrivait pas à se remettre en chat, il était resté en fourmi donc Minou alla voir Mademoiselle Cascou

et il lui demanda « est-ce que tu peux me remettre en chat s'il te plait? » Mademoiselle Cascou répondit « non » Minou dit « s'il te plait, s'il te plait » Mademoiselle Cascou répondit « non non non et non » et la nuit tomba donc Minou devait dormir en fourmi. le lendemain Minou alla revoir Mademoiselle Cascou, Mademoiselle Cascou dit à Minou « tu as dormi en chat » Minou répondit « oui » Mademoiselle Cascou répondit « mon pauvre je vais te remettre en chat » « merci » répondit Minou et ils deviennent les meilleurs amis du monde.

Production NORM-EC-CE2-2016-130-D1-S1125

Il était une fois un chat qui alla dans la forêt. Il rencontra le loup noir. Le chat dit : « Qui es-tu? » Le loup noir lui dit : « Je suis le plus méchant loup de la planète et de la forêt ». Le chat s'en alla discrètement mais le loup l'avait vu. Le loup entendit un bruit mais avant que le loup aille voir d'où venait le bruit. Beee le loup enferma le chat dans une cage. Ensuite il alla voir le bruit, tout était calme mais tout à coup le loup entendit le bruit d'un mouton, il l'attaqua et le dévorait et tout à coup il entendit un bruit de chèvre un peu plus loin dans la colline mais il se disait : « non, je n'irai pas, je dois aller voir le chat ». Plus loin le chat ne réussit pas à s'en aller de la cage. Ensuite à plusieurs reprises de coup de griffes le chat réussit à sortir de la cage mais le loup venait d'arriver. le chat se cacha vite sous un grand arbre si grand que le loup ne l'avait pas vu. le loup cria « non » parce que le chat avait réussi à sortir de la cage et que le loup voulait manger le chat pour le diner. Le chat s'en alla discrètement et rentra chez lui. Mais il rencontra un chien abandonné et le ramena chez lui. Ensuite il sortit dehors pour jouer avec lui dans le jardin mais le loup arriva et ensuite le chat et le chien rentrent. le loup ne les voyait pas, il repartit dans la forêt. Mais le soir il revint mais il ne vit personne dans la maison et il resta 1h puis il partit fatigué puis il revint le jour pour attraper le chat mais alors le chat avec le chien avaient déménagé à cause du loup . le loup désespéré attrapa un aigle pour le dîner. Il repartit dans la forêt fatigué. Il avait trouvé sa mère qui arrivait dans la forêt avec son petit frère et sa petite soeur.

Production NORM-EC-CE2-2016-130-D1-S2531

Il était une fois un chat qui se promenait dans la forêt et qui ramassait des fleurs. Pendant un moment un loup le guettait. Au bout d'un moment il se jetait sur le pauvre petit chat et il lui demandait : « Où vas-tu?. ». Alors le petit chat lui disait : « Dans ma cabane où il y a ma poupée- sorcière ». Et le loup allait tout de suite vers la cabane. Quand il était arrivé il se mit à dévorer la poupée-sorcière du chat. Et quand le chat était arrivé lui aussi dans la cabane il ne voyait plus la grand-mère mais que le loup. Alors il décidait d'appeler la police. Quand la police arrivait vers la cabane du chat ils descendaient de leurs voitures. Tout à coup le loup avait entendu un bruit dehors alors il sortait. Mais quand il sortait la police le capturait. Quand il fut attrapé et mis en prison pour loup on n'entendit parler que de lui dans les journaux. Mais la police avait oublié la poupée sorcière du petit chat. Alors ils décidaient d'ouvrir le ventre du loup. Quand ils avaient ouvert le ventre du loup et attrapé la poupée-sorcière du petit chat. Et pour finir le petit chat qui était content fit miaou, miaou, miaou, miaou mais de plus en plus fort jusqu'à la fin de la journée. Mais ça énervait ses parents alors ses parents décidaient de lui mettre du scotch sur la bouche. Mais le chat l'avait enlevé alors ils le laissaient faire maintenant jusqu'à la fin de la journée.

Production NORM-EC-CE2-2016-17-D1-S562

Il était une fois un loup et son fidèle serviteur le robot. Le robot faisait tout ce que le loup demandait mais un jour le robot dit au loup : «-Je m'en vais. » Le loup dit « :-Pourquoi

pars-tu? » Le robot dit : «-Je prends grève, je suis fatigué » dit le robot . «-Mais un robot est fait de fer, il n'est jamais fatigué, » dit le loup«-Mais je prends grève quand même. Je prends grève de 100 ans. » Le loup reste bouche bée devant le robot qui s'en va. Fin.

Production NORM-EC-CE2-2016-17-D1-S572

Bonjour, c'est Bob le robot. Je vais vous raconter la fois où je marchais dans les bois : je marchais dans les bois et là je découvris une vieille mais très vieille maison, elle était louche parce que dans le jardin poussaient des plantes bizarres comme une aubergine bleue. je rentrai dans la maison. il y avait des potions de toutes les couleurs. Mais j'entendis du bruit et me cachai derrière la poubelle. elle était là. je levai la tête et vis un chapeau. je me levai et vis une robote très jolie. Et voila, au revoir.

Production NORM-EC-CE2-2016-17-D1-S576

La sorcière et sa chatte . La sorcière et sa chatte font une potion magique avec des plantes et des serpents et des myrtilles tellement elles étaient les plus gentilles et des fois étaient des fois méchantes. La sorcière s'appelle Justine et sa chatte s'appelle Chillie .

Production NORM-EC-CE2-2016-17-D1-S582

Le chat et la sorcière Le chat se promenait dans la forêt. Et il vit une sorcière qui transforma les chevaliers en grenouille. Et la sorcière vit le chat et essaye de le transformer le chat en grenouille. La sorcière transforma le chat en grenouille, et le mit dans un bocal. Mais la grenouille attrapa la et se délivra et reprit sa taille.

Production NORM-EC-CE2-2016-17-D1-S584

Le chat et la sorcière C'est l'histoire d'un chat très malin qui vit dans un château, tous les soirs le chat sort pour aller au village. un soir, il se perd dans la forêt croit qu'il allait au village, il prend peur, court dans tous les sens. à un moment il trouve une cabane il entrait. se coucha sur le canapé. le lendemain il tombe nez à nez avec la sorcière du village, la sorcière prend le chat et le transforma en chat noir et elle dit « joue avec moi » le chat étonné s'approche d'elle et la caresse. maintenant chaque soir ils se retrouvent dans la forêt pour jouer. Fin.

Production NORM-EC-CE2-2016-19-D1-S1596

Il était une fois un chat et une sorcière . le chat était sur son tapis pendant que la sorcière est en train de travailler, et d'un coup le chat miaule, la sorcière lui donne de la pâté de chat . le chat miaule toujours donc la sorcière met le chat dehors. le chat miaule, miaule par la fenêtre et d'un coup la sorcière crie et le chat miaule encore plus fort donc la sorcière refait rerentrer le chat et le chat ne miaule plus et le chat s'endort et quand le chat se réveille la sorcière était avec lui. pour le petit-déjeuner le chat mange des croquettes et la sorcière mange un bol de lait et deux tartines de beurre et confiture. la sorcière et le chat sortent dehors. la sorcière donne la balle au chat et le chat fait la passe à la sorcière et ainsi de suite, la sorcière dit au chat « c'est l'heure de manger mon petit chat » elle lui a préparé des croquettes.

Production NORM-EC-CE2-2016-19-D1-S3051

Il était une fois une sorcière qui préparait une potion magique. il survint un homme qui se transformait en chat. la sorcière essaya de le transformer en humain mais la sorcière n'arrivera jamais à le retransformer en humain. le chat restera pour toujours. la sorcière sera morte dans une prison parce qu'elle s'est suicidée. le chat meurt aussi sur les roues d'une voiture.

Production NORM-EC-CE2-2016-19-D1-S3052

Il était une fois une sorcière qui rencontre un robot et le robot lui demande « je veux devenir en un loup » et elle lui dit « d'accord » parce que je vais te transformer. il est en loup. Je lui dis « merci » et le loup a mangé la sorcière, et le village était sauvé de la méchante sorcière.

Production NORM-EC-CE2-2016-20-D1-S107

Il était une fois une sorcière et son petit chat. Ils vivaient heureux et s'aimaient beaucoup. La sorcière nourrissait son chat et s'occupait de lui. Ils faisaient des jeux et s'amusaient à attraper des souris et des araignées. Mais un jour un méchant loup s'approcha du chat et... il sauta sur lui et le dévora en une bouchée. La sorcière était très triste mais un jour elle eut une idée. Avec sa magie elle allait retrouver le loup et lui jeta un sort! Grâce à ce sort le chat sortit du ventre et ils rentraient chez eux. Et ils étaient encore en train de jouer et la sorcière se sentit de nouveau heureuse. FIN.

Production NORM-EC-CE2-2016-20-D1-S93

Il était une fois, dans une maison un petit chat qui aimait jouer. Il voulait voir à quoi ressemblait la maison des voisins. Il se demande si les croquettes sont meilleures que chez lui. Il dit : « Salut les amies! » crie le chat « Salut, tu fais quoi? » dit l'un « Ben je vais voir si les croquettes sont meilleures chez les voisins! » dit le chat étonné. Et il continue sa route. Il arriva devant la maison des voisins. Malpoli par la fenêtre il entra. « Mais où sont-ils passés? » s'étonna le chat. Juste devant lui sont servies des croquettes. En revenant à la maison, il avait mangé tellement de croquettes qu'on l'appela Croquette. FIN.

Production NORM-EC-CE2-2016-20-D1-S94

Il était une fois ... un loup qui voulait manger un beau et gros chat mais il était dans la jungle alors il pensait : « J'ai mangé beaucoup d'animaux alors forcément je trouverai un chat sauvage... » Soudainement il entendit un bruit. il alla se cacher dans les buissons en faisant un peu de bruit. Mais le chat a entendu ce bruit et il le regarda d'un air fixe et disait : « Je suis sûr que vous voulez me manger » « Non Non, je visitais les lieux. » « C'est ça! Vous dites n'importe quoi. » « Je vous préviens... » « MIAOU!!! » Le chat fit un saut périlleux mortel en arrière et lui a donné des claques, a sorti ses griffes! « Mais arrête... » « Jamais! » À ce moment précis il l'a avalé tout cru et pour finir un chasseur est venu et a tout vu alors il tua le loup et le chat est sauvé. THE END.

Production NORM-EC-CE2-2016-20-D1-S96

Il était une fois un chat très malin et attentif à ce qu'il faisait, mais il y a le loup aussi, lui tout le contraire du chat distrait et pas du tout malin. Un jour le chat se promenait mais le loup était toujours à surveiller les chats, les lapins, et il tomba sur le chat. « Mmm c'est un bon repas que je vois là, je vais l'attraper et le manger pour le repas de ce soir » dit le loup.

Mais le lapin très attentif regarda le loup d'un air malin alors se faufila entre deux buissons pour écouter le loup parler. Pendant ce temps... « Ha! Ha je vais préparer ma bouillotte ». Le loup s'imagina le petit film quand il va manger le lapin. Mais le lapin est très malin, il ne sait pas ce qu'il va se passer au loup ! FIN!

Production NORM-EC-CE2-2016-20-D1-S98

Il était une fois une sorcière qui voulait manger un chat très futé et qui faisait rater toutes ses potions. Un jour elle a essayé de lui jeter un sort qui rend instantanément aveugle mais il se cacha derrière un miroir et la sorcière devint aveugle. Après tous ses échecs elle a voulu jouer loyal. Alors elle proposa un combat. Le chat accepta. La sorcière utilisa son balai sans succès. Alors elle fonça pour l'attraper, il évita habilement. La sorcière se retenait pour ne pas tomber. Le chat attaqua ses mains et elle tomba dans le puits. Depuis ce jour on n'entendit plus jamais parler d'elle.

Production NORM-EC-CE2-2016-28-D1-S1173

Il était une fois une sorcière, et son chat noir. Un bon matin commence. Elle prépara une potion magique. La sorcière mit une grenouille, des cailloux et une pièce de tout en 1. Puis le chat noir s'enfuit car la sorcière s'enfuit.

Production NORM-EC-CE2-2016-37-D1-S1560

Il était une fois une sorcière terrifiante à la voix lugubre, elle avait un chat noir comme une nuit sans étoile et gros, il était aussi très sale. La sorcière partit souvent à la chasse sur son balai magique, un jour qu'elle partait à la chasse elle rencontra un loup, comme elle en avait grand besoin pour ses potions, elle fit boire à Cupidon une potion dont elle ne se séparait jamais, et qui faisait qu'elle pouvait mettre ses idées dans la tête de Cupidon, ainsi elle pouvait lui faire faire ce qu'elle voulait. elle put lui faire planter une flèche dans les fesses du loup pour qu'il soit amoureux d'elle. Elle put l'emporter chez elle pour s'en servir pour ses potions magiques. Fin.

Production NORM-EC-CE2-2016-47-D1-S2937

Autrefois le chat et le chien étaient des grands amis. Un beau matin le chat allait voir son ami mais il rencontra l'écureuil et le chat demanda « Tu n'as pas vu mon ami le chien? » « Si, il m'a dit qu'il ne voulait plus être ton ami » et il dit qu'il était chez lui. Le chat était très en colère et le chat allait lui crier dessus. le chien était très colère de lui avoir crié dessus. Et c'est depuis ce jour-là que le chat a peur du chien.

Production NORM-EC-CE2-2016-6-D1-S2000

Le robot martien Un jour une navette spatiale s'écrase à Chatville, le pays des chats quand descendit un drôle de bonhomme fait en un métal et en clous. Bien sûr les humains savent qu'elle était un robot, mais les chats .... non. Alors le robot sortit de la navette. les chats restèrent muets mais quand il en descendit on pouvait entendre à des kilomètres : « Va t'en! » « Ce n'est pas ta planète ! » « Tu n'es pas comme nous, pars! » « File! » Alors le robot s'enfuit mais, dans le sens inverse de sa navette, mais vers la maison de l'inventeur. Il ouvrit la porte et là se dressait un autre robot . Alors tous deux devinrent amis

mais quand l'inventeur chat se réveilla et il les chassait. Alors avec la navette il retournait sur la planète du robot où il était heureux. Fin.

Production NORM-EC-CE2-2016-6-D1-S2035

La magie puissante Il y avait une sorcière qui s'appelait Sorcièla et un chat qui s'appelait Mignon. Il était une fois une sorcière et elle avait plein de puissance et de magie avec son balai volant et sa baguette magique. Sorcièla habitait dans un château hanté avec plein d'araignées collées au plafond et plein de chauve-souris volantes partout. Mignon, lui était un chat mignon qui habitait dans la forêt, abandonné et qui cherchait souvent un refuge pour s'abriter. Mais un jour il tomba sur le château hanté de Sorcièla. Et d'un coup de baguette le chat disparut et Sorcièla avec sa magie de plus en plus puissante arriva à contrôler le monde entier et les empoisonnait à jamais. Fin.

Production NORM-EC-CE2-2016-85-D1-S1981

La sorcière prépare une marmite et prend son balai et après elle part chercher de la farine pour la marmite et invite sa famille et dit « c'est bon. Ça c'est très très très bon, au revoir, à demain, je t'aime très fort » « moi aussi maman, à demain ». « Arrête papa, j'ai une fille, venez on va manger. » « ok encore à demain, je viendrai » et soudain ma maman me crie dessus. papa dit des grossièretés. Et la soeur vient manger chez moi le mercredi. Et sa grand-mère arrive à la maison, elle voit un loup très très méchant « à l'aide grand-mère. Je suis perdue dans la forêt. à l'aide » « je suis là pour vous aider madame, je suis un pompier » « merci beaucoup et quand j'ai peur ça tremble ». Le chat « bonjour je suis perdu dans le magasin. Je veux de l'aide madame et monsieur. Je suis trop perdu dans le magasin, j'ai fini de manger mon gouter monsieur et mademoiselle. Je viens à la recherche d'un canard. Je vous prouvais-je suis un chat perdu. Et je parle pour un film, je prouvais, je suis très mignon madame. « C'est vrai, je vous garantis que ça va mal se passer » « très bien madame, arrêtez de me dire des gros mots » « ok arrêtez je vais appeler la police hein madame » arrêtez, j'appelle la police. »

Production NORM-EC-CE2-2016-91-D1-S1154

1 Il était une fois un loup qui vivait dans la forêt. Il mangeait les enfants qui n'étaient pas gentils. Un jour il dévora un petit garçon qui était méchant avec le loup . quand il a mangé le petit garçon le loup gonfla, gonfla, et gonfla tellement qu'il explosa et le loup disparait. 2 Il était une fois un petit chat qui était perdu. Il était triste et ses parents l'appelaient, eux aussi étaient tristes. Il était pris par des méchants garçons mais leurs parents l'appelaient très fort. 3 Il était une fois une sorcière méchante. Elle faisait croire aux enfants qu'elle faisait manger des enfants. En fait elle donna du poison aux enfants et les enfants meurent.

Production NORM-EC-CE2-2016-91-D1-S1160

Il était une fois un petit chaperon rouge qui avait une grand-mère. La grand-mère était très malade. La maman du petit chaperon rouge lui a dit : « Va chez ta grand-mère qui est malade et va lui apporter cette tarte et ce petit pot de beurre et fais attention à toi car il y a le loup dans la forêt. » Et la petite y alla. le loup arriva avant elle chez la grand-mère et la dévora et s'est déguisé en vieille dame. La jeune fille entra chez sa grand-mère et le loup l'avala.

Production NORM-EC-CE2-2016-91-D1-S1164

Il était une fois une sorcière qui vivait dans un château affreux. La sorcière avait un nez laid et son chat était tout gris. Elle va dans la forêt et rencontre une jeune fille qui cueille des champignons. La sorcière voulait manger la fille. Elle s'est cachée derrière un arbre et quand la fille se retourne la sorcière lui saute dessus. Elle retourne dans son château et mange avec le chat .

Production NORM-EC-CE2-2016-91-D1-S663

Il était une fois un chat qui se promenait dans la rue. Et il a vu un robot alors le chat va vers le robot et lui parla. Et il lui dit « veux-tu jouer avec moi? » dit le chat . Le robot dit « oui je veux bien ». Alors ils font une partie de cache-cache. Le robot s'est caché dans une poubelle, le chat ne le trouve pas. Alors le chat le trouve dans une poubelle. Le chat se cache dans une caisse, le robot l'a vu.

Production NORM-EC-CE2-2016-91-D1-S949

La sorcière et son chat Il était une fois une sorcière qui avait un chat . Elle était méchante, elle voulait des enfants. Elle va toujours dans la rue pour piquer des enfants. Et aujourd'hui elle va dans la rue avec son chat . Elle a pris son chat pour que dans la rue elle le montre aux enfants pour qu'elle les prenne. Elle prit son balai, son chat puis elle grimpa sur son balai et elle vole, vole, vole. Arrivée dans la rue elle a vu un enfant, elle se posa et va le voir. elle lui dit « tu veux un bonbon? » et il lui dit « Oui ». Elle lui dit « D'accord, tiens » il tendit la main, elle le prend et elle l'emmène chez elle. Puis elle se rend compte que ce n'est pas bien puis elle lui dit « Tu veux être mon enfant? » il lui dit « Non » elle le relâche. Puis elle trouva un mari et elle se marie et ils ont plein d'enfants.

Production NORM-EC-CE2-2016-91-D1-S982

Le Robot maladroit Il était une fois un robot qui avait 3 frères très intelligents, qui s'appelaient Tibot, Toto, Sami. Tibot était le plus grand mais il y avait aussi le petit robot Jeanne qui était le plus maladroit, il ne faisait que des bêtises mais un jour une sorcière très gentille qui s'appelait Diabelle le rendit normal et puis les robots c'est des robots . Fin.

Production NORM-EC-CE2-2016-96-D1-S1927

La sorcière et le loup Il était une fois un loup qui se promenait. Un jour le loup la rencontra, il lui dit. « Bonjour la sorcière, comment allez-vous? » « Très très bien. » lui dit la sorcière . « Nous allons faire un marché. si tu grimpes cet arbre je te donne un morceau de viande et si tu n'y arrives pas tu seras une grenouilles, d'accord? » « OK si tu veux. » Alors il grimpa mais il tombe, heureusement que la sorcière était endormie, alors il recommence et il n'y arrive pas alors il décide de partir et la sorcière se réveille.- Elle lui dit « je vais te transformer » et lance le sort. Il esquive le sort et part en courant. La sorcière sort avec son balai. Et le loup trouve un morceau de miroir et lui jetait encore le sort et lui rejette le sort. Et la voilà grenouille. Mais elle ne lâche rien alors elle reprend sa baguette magique. Et elle lui jetait un autre sort. Mais elle s'était trompée de sort, c'était le sort d'un bout de viande. Et maintenant elle lui éjecte le sort de la grenouille mais il avait toujours son bout de miroir. Et elle devient un crapaud. Et elle lâche l'affaire et elle va chez elle et se

retransforme en sorcière et elle se maquille et devient une jeune fille et elle cherche le loup et le voit et lui dit bonjour. Et elle la baguette et elle le transforme en grenouille.

Production NORM-EC-CE2-2016-96-D1-S1931

La sorcière et le chat Il était une fois une sorcière, qui s'appelait Juliette. Elle n'aimait pas du tout ce nom, elle voulait s'appeler la sorcière de la rue Broca. Et elle avait un chat, elle ne l'aimait pas du tout, c'est pour ça qu'elle l'appelait Juliette. Ce chat n'aimait pas du tout ce nom et lui comparé à la sorcière, il aimait la sorcière . Un jour la sorcière le transforma en petite fille, et la sorcière trouva cette petite fille très gentille et très mignonne. La sorcière a appelé sa fille (son chat qu'elle a transformé en fille), la fille de la sorcière Broca et la sorcière acheta un nouveau chat et plein d'animaux pour sa fille. Mais sa fille voulait encore plus d'animaux mais tous ces animaux étaient méchants et la sorcière les transforma en petites filles, en petits garçons et en grandes filles et en grands garçons. Un chat apparait et la sorcière n'avait plus de magie et c'était un chat gentil donc la sorcière aimait ce chat et tous ces enfants aussi.

Production NORM-EC-CE2-2016-98-D1-S815

Un petit chat s'éloignait doucement de la ville. La nuit tombée, il décide de dormir dans la forêt. La pleine lune, les yeux du petit chat scintillèrent. Et d'un seul coup le petit chaton se transforma en un glouton de loup qui s'appela Nicolas François.

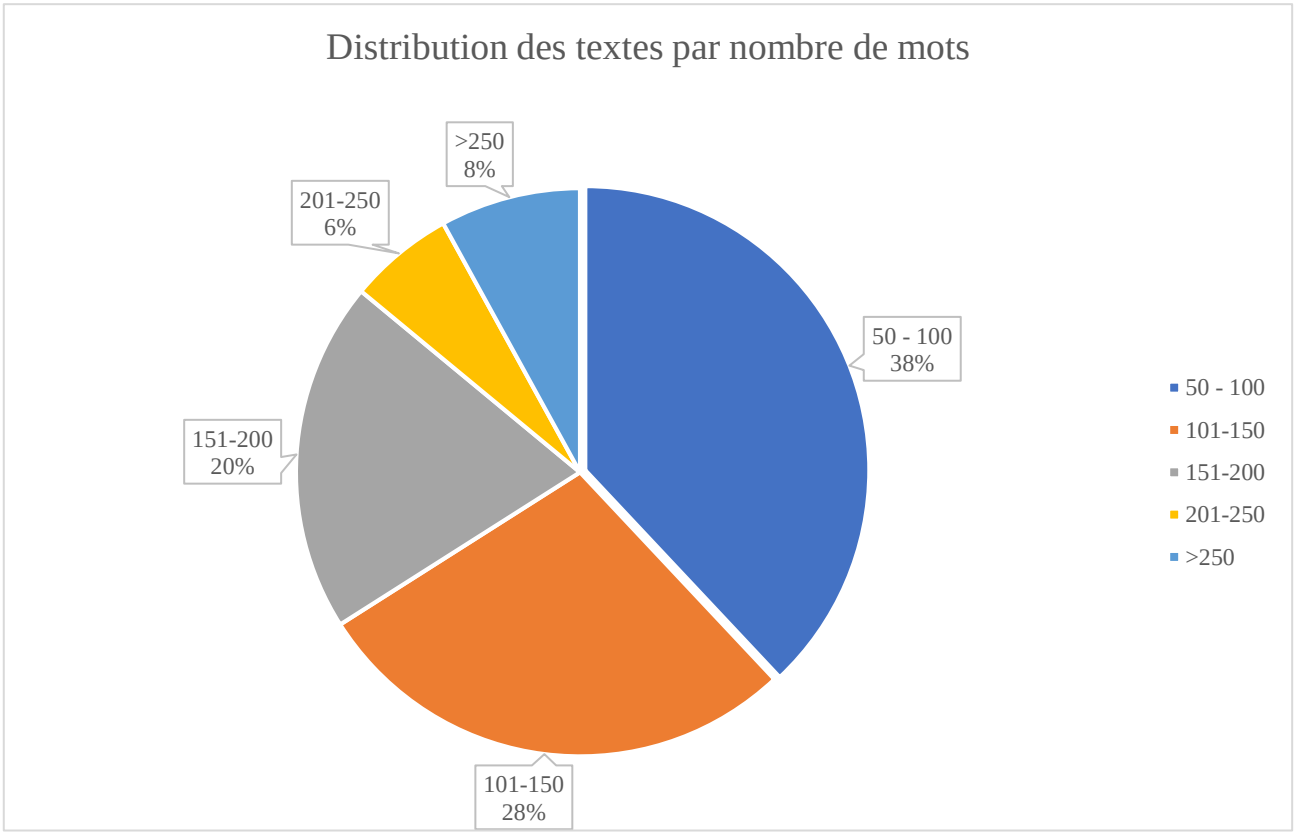
Production NORM-EC-CE2-2016-98-D1-S817

Chapitre 1 pas de nom Le robot Jen -pouleloul le destructeur de l'univers. Le loup Trololole qui trolle tout le monde. Et le Jen- pouleloul voit Trololole qui trolle Nicolas Sarkozy qui dit bonjour à François Hollande. Jen -pouleloul tue Nicolas Sarkozy. Chapitre 2 Minecrafte Jen -pouleloul se trouve dans Minecrafte et Trololole aussi. Jen- pouleloul pousse 1 0000 0000 9999 99999999999 TNT. Trololole allume les TNT et ça fait Boum! Jen- pouleloul a détruit Minecrafte. Chapitre 3 La destruction Mais Jen- pouleloul a gardé 99999999999999999999999999999999 TNT et Trololole a un briquet. Jen- pouleloul pousse toute sa TNT et Trololole enflammait Jen- pouleloul et Jen- pouleloul se jette dans la TNT et ça fait boum! Fin

Production NORM-EC-CE2-2016-98-D1-S936

« Bonjour le chat, je vais te déchiqueter. En un morceau, je ne vais pas te déchiqueter, c'est une blague » dit le chat « au revoir » « attends mon vieil ami, je ne suis pas vieux, tu n'as pas compris le mot. »

Annexe 5. Détail de la distribution des textes par nombre de mots



Annexe 6. Détail du calcul de la densité référentielle (avec longueur des textes et longueur des chaînes de coréférence pour chaque référent, len_référent) sur les 50 textes du corpus de travail de CE2

	Densité référentielle					Longueur chaînes			
texte	witch	cat	wolf	robot	Longueur texte	len_witch	len_cat	len_wolf	len_robot
102-D1-S212	0,07	0,05	0	0,02	86	6	4	0	2
102-D1-S214	0	0,09	0	0,07	44	0	4	0	3
104-D1-S831	0	0,06	0	0	223	0	13	0	0
104-D1-S835	0,12	0,04	0,02	0,01	107	13	4	2	1
104-D1-S841	0,04	0,07	0	0	191	7	13	0	0
104-D1-S855	0,07	0,03	0	0	276	18	8	0	0
108-D1-S2972	0	0,04	0	0,04	111	0	4	0	4
108-D1-S2976	0,1	0,1	0	0	50	5	5	0	0
108-D1-S2978	0,05	0,05	0	0	135	7	7	0	0
108-D1-S2981	0,02	0	0	0,05	128	3	0	0	6
108-D1-S2986	0,12	0,06	0	0	141	17	8	0	0
108-D1-S2988	0,05	0,07	0	0	98	5	7	0	0
115-D1-S1345	0,07	0,05	0	0	169	11	9	0	0
117-D1-S3003	0,06	0	0	0,04	239	14	0	0	10
120-D1-S2439	0	0	0,06	0,05	162	0	0	10	8
120-D1-S2453	0,06	0,1	0	0	86	5	9	0	0
126-D1-S1826	0,1	0,05	0	0	91	9	5	0	0
126-D1-S1830	0,04	0,05	0	0	188	8	9	0	0
130-D1-S1125	0	0,05	0,07	0	357	0	17	26	0
130-D1-S2531	0	0,06	0,04	0	270	1	15	11	0
17-D1-S562	0	0	0,05	0,08	92	0	0	5	7
17-D1-S572	0	0	0	0,01	99	0	0	0	1
17-D1-S576	0,1	0,1	0	0	49	5	5	0	0
17-D1-S582	0,06	0,09	0	0	66	4	6	0	0
17-D1-S584	0,05	0,11	0	0	120	6	13	0	0
19-D1-S1596	0,09	0,12	0	0	172	15	20	0	0
19-D1-S3051	0,07	0,06	0	0	68	5	4	0	0
19-D1-S3052	0,05	0	0,05	0,03	58	3	0	3	2
20-D1-S107	0,08	0,05	0,03	0	119	10	6	3	0
20-D1-S93	0	0,06	0	0	125	0	7	0	0
20-D1-S94	0	0,06	0,05	0	173	0	10	8	0
20-D1-S96	0	0,05	0,08	0	155	0	7	12	0
20-D1-S98	0,06	0,04	0	0	110	7	4	0	0
28-D1-S1173	0,06	0,06	0	0	48	3	3	0	0
37-D1-S1560	0,09	0,01	0,02	0	132	12	1	2	0
47-D1-S2937	0	0,12	0	0	102	0	12	0	0
6-D1-S2000	0	0,04	0	0,08	157	0	6	0	12
6-D1-S2035	0,03	0,04	0	0	128	4	5	0	0
85-D1-S1981	0,02	0,02	0,01	0	235	4	4	3	0
91-D1-S1154	0,05	0,03	0,06	0	126	6	4	8	0

91-D1-S1160	0	0	0,03	0	96	0	0	3	0
91-D1-S1164	0,1	0,03	0	0	73	7	2	0	0
91-D1-S663	0	0,08	0	0,07	89	0	7	0	6
91-D1-S949	0,12	0,03	0	0	177	21	5	0	0
91-D1-S982	0,04	0	0	0,08	71	3	0	0	6
96-D1-S1927	0,06	0	0,07	0	261	15	0	17	0
96-D1-S1931	0,1	0,05	0	0	191	19	9	0	0
98-D1-S815	0	0,08	0,02	0	49	0	4	1	0
98-D1-S817	0	0	0,05	0,08	119	0	0	6	10
98-D1-S936	0	0,05	0	0	42	0	2	0	0
Moyennes avec 0	0,0688	0,06	0,0444	0,0507	133,08	8,4242	7,175	7,5	5,5714
Moyennes sans 0	0,044	0,048	0,0142	0,0142	133,08	5,56	5,74	2,4	1,56

Dénsité référentielle sur les quatre personnages et sur les 50 textes du corpus de travail