Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

Étude du comportement des composants d'expressions polylexicales verbales dans les chaînes de coréférence

MASTER

TRAITEMENT AUTOMATIQUE DES LANGUES

Parcours :

Ingénierie Multilingue

par

Anaëlle PIERREDON

Directeur de mémoire :

Loïc Grobol

Encadrants :

Jean-Yves Antoine, Anaïs Lefeuvre-Halftermeyer et Agata Savary

Année universitaire 2020/2021

Remerciements

Je tiens à remercier mon directeur de mémoire, Loïc Grobol, ainsi que mes encadrants de stage Agata Savary, Jean-Yves Antoine et Anaïs Lefeuvre-Halftermeyer pour leur disponibilité et leur soutien au cours du stage et de la rédaction de ce mémoire.

Je remercie également Jianying Liu, avec qui j'ai eu le plaisir d'effectuer mon stage.

Résumé

La coréférence et les expressions polylexicales sont deux phénomènes linguistiques importants en traitement automatique des langues et notamment dans des tâches comme la traduction automatique ou encore la fouille de texte. Au cours de ce mémoire, nous chercherons à valider l'hypothèse selon laquelle les composants d'expressions polylexicales ne sont que très peu susceptibles d'être repris dans des chaînes de coréférence, et nous proposerons une façon d'utiliser ces résultats pour tenter d'améliorer les systèmes de résolution de coréférence.

Mots clés : coréférence, expressions polylexicales, MWE, français

TABLE DES MATIÈRES

Liste des figures

Liste des tableaux

INTRODUCTION

Présentation générale

La coréférence et les expressions polylexicales sont deux phénomènes linguistiques importants en traitement automatique des langues et notamment dans des tâches comme la traduction automatique ou encore la fouille de texte. Dans le cadre d'un stage co-encadré par le LIFO (Laboratoire d'Informatique Fondamentale d'Orléans) et le LIFAT (Laboratoire d'Informatique Fondamentale et Appliquée de Tours), nous avons cherché à étudier le comportement des composants d'expressions polylexicales et leur possible reprise dans des chaînes de coréférence. Au cours de ce mémoire nous chercherons à vérifier l'hypothèse selon laquelle les composants d'une expression polylexicale ne peuvent que rarement être repris dans des chaînes de coréférence. L'ensemble des travaux réalisés durant ce stage sont disponibles sur GitHub [1].

Plan de lecture

Ce mémoire est organisé en trois parties :

— L'état de l'art, où nous définirons les phénomènes étudiés et présenterons les récents travaux réalisés sur ces derniers. Cette partie se compose de trois chapitres s'attardant sur la coréférence, les expressions polylexicales et enfin l'hypothèse à vérifier.

— Les expérimentations, où nous décrirons les expériences réalisées et discuterons des résultats obtenus. Cette partie se compose de quatre chapitres où nous présenterons les outils utilisés, le choix des corpus, les traitements appliqués et enfin, les résultats.

— Les perspectives, où nous réfléchirons à la manière d'utiliser ces résultats pour améliorer les performances d'un système de résolution de coréférence.

1. GitHub du stage - anaelle-p/MWE_coref [Lien]

Première partie

ÉTAT DE L'ART

CHAPITRE 1

LA CORÉFÉRENCE

Sommaire

La coréférence est un phénomène linguistique important dans les tâches de traitement automatique des langues comme le résumé automatique, la fouille de texte ou encore la traduction automatique. Nous commencerons par définir ce phénomène, puis nous présenterons les corpus français annotés manuellement en coréférence et enfin, nous verrons les méthodes de résolution de coréférence et les travaux récents.

1.1 Définitions

La coréférence est un procédé linguistique dans lequel plusieurs éléments d'un discours réfèrent à une même entité du monde du discours. Par exemple :

« *Anna a ouvert la porte et l'a refermée.* »

La porte et *l'* sont deux mentions, c'est-à-dire des éléments qui réfèrent à une entité du monde du discours. Ces mentions font partie de la même chaîne de coréférence car elles réfèrent à une même entité, elles sont donc coréférentes. La plupart des mentions sont des groupes nominaux et des pronoms, les pronoms étant généralement coréférents avec un groupe nominal proche ([Grobol, 2020]).

Anna est également une mention mais, si l'on considère la phrase d'exemple comme un document complet, elle n'est coréférente avec aucune autre mention. C'est donc une mention que l'on appelle un singleton. Certains systèmes de reconnaissance de chaînes de coréférence ne considèrent pas ces entités comme des mentions, car elles n'appartiennent à aucune chaîne. D'autres considèrent qu'elles réfèrent tout de même à une entité du monde du discours, et forment une chaîne de coréférence ne contenant qu'une seule mention.

Dans la terminologie d'ANCOR (voir 1.2), on appelle antécédent la première mention nominale faisant référence à l'entité de la chaîne de coréférence. Nous utiliserons également cette terminologie dans la suite de ce mémoire. Par exemple *la porte* est l'antécédent de *l'*. L'antécédent est généralement la première mention du texte qui fait référence à une nouvelle entité du discours, mais il arrive qu'un pronom soit ren-

contré avant l'antécédent. Ce phénomène est appelé cataphore (ex : « *Quand il était petit, Paul jouait au foot.* » [Oberle, 2019]).

Une anaphore est une relation asymétrique entre un antécédent et une mention dont l'interprétation n'est possible qu'à partir de l'antécédent ([Désoyer et al., 2015]). Dans notre exemple, il y a une relation anaphorique entre *la porte* et *l'*. En effet, il est impossible d'interpréter *l'* sans son antécédent.

1.2 Corpus annotés en coréférence en français

Il y a peu de corpus de taille suffisamment importante annotés en coréférence en français disponibles gratuitement, pourtant indispensables à la mise en place de systèmes d'identification de chaînes de coréférence avec des techniques d'apprentissage automatique. Nous allons donc présenter ANCOR et DEMOCRAT, des corpus annotés manuellement en coréférence.

1.2.1 ANCOR

ANCOR[1] (Anaphore et Coréférence dans les Corpus ORaux) est le premier corpus en français de taille importante qui soit annoté en coréférence et disponible gratuitement, ainsi que le plus gros corpus oral annoté en relation anaphoriques [Muzerelle et al., 2014]. Il contient 488 000 mots et est composé de quatre sous-corpus contenant des entretiens, des discussions récupérées depuis un standard téléphonique et des dialogues.

Avec 115 672 mentions et 51 494 relations ([Désoyer et al., 2015]), on peut considérer ANCOR comme étant représentatif de tous les aspects de la coréférence orale. Seuls les groupes nominaux sont annotés comme des mentions, et les pronoms déictiques[2] sont toujours considérés comme des singletons lorsqu'ils réfèrent aux interlocuteurs ([Désoyer et al., 2015]).

Chaque mention est décrite par différents traits comme le genre, le nombre, la partie du discours, la définitude (défini, indéfini, démonstratif ou explétif), l'inclusion dans une proposition prépositionnelle, le type d'entité nommée ou encore une indication permettant de savoir si la mention est la première rencontrée faisant référence à cette entité. Il existe différentes façon de représenter les chaînes de coréférence et il été fait le choix dans ANCOR de relier les mentions coréférentes à la première mention rencontrée, mais il est également possible de relier chaque mention coréférente à la précédente ou encore de travailler avec des ensembles.

Il y a cinq relations possibles dans ANCOR :

— Les coréférences directes, où la mention et son antécédent ont la même tête.
« *je trouve par exemple qu' il est plus facile de dire parking que parking ou parc même ou* *parquage* *parquage*[3] *?* » (ESLO_ANCOR)

— Les coréférences indirectes, où la mention et son antécédent ont des têtes différentes.

1. Corpus ANCOR disponible sur la plateforme ORTOLANG [Lien]
2. Pronoms dont le sens ne peut être interprété qu'à partir de la situtation dans laquelle ils sont employés.
3. Les mentions seront surlignées en vert tout au long du mémoire.

« *que pensez- vous monsieur du franglais ?*
...
c' est le parler français anglais ? » (ESLO_ANCOR)

— Les anaphores pronominales, où la reprise est un pronom.
« *que pensez- vous monsieur du franglais ?*
...
c' est le parler français anglais ? » (ESLO_ANCOR)

— Les anaphores associatives nominales, où les mentions ne sont pas coréférentes mais dont l'interprétation de l'une dépend de l'autre.
« *oh oui y a -t -il des différences selon vous dans la façon de parler français entre les gens appartenant à des milieux sociaux différents* ?
...
sur la prononciation de euh des mots euh » (ESLO_ANCOR)

— Les anaphores associatives pronominales, où la reprise associative est pronominale.
« *j' ai des amis à Salbris là où on va ils parlent très bien français lui était géomètre euh sa femme était institutrice* » (ESLO_ANCOR)

Afin de prouver la fiabilité des annotations du corpus ANCOR, des accords inter-annotateurs et intra-annotateurs ont été calculés pour les tâches d'identification des relations et de typage de ces dernières. Les résultats ont montré que la tâche d'identification est subjective, mais que les annotateurs restent cohérents avec eux-mêmes et que l'accord est plutôt bon pour la tâche de typage ([Muzerelle et al., 2014]). Les annotations d'ANCOR sont donc fiables.

1.2.2 DEMOCRAT

DEMOCRAT [4] (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 corpus en français écrit annoté en coréférence. Le but de ce corpus est d'étudier la variation des chaînes de coréférence en fonction des genres discursifs et des époques. En effet, ce corpus contient des textes de genres variés (roman, pamphlet, fables, texte juridique...) écrits entre le 12ème et le 21ème siècle ([Landragin, 2020]). Il contient 689 000 mots et 198 000 expressions référentielles.

1.3 Méthodes de résolution de coréférence

La résolution de coréférence se compose de deux étapes : la détection des mentions et l'identification des chaînes de coréférence. La coréférence étant un phénomène hétérogène avec plusieurs mécanismes, sa résolution n'est pas une tâche aisée. Pour pouvoir identifier les mentions et les chaînes de coréférence, il est donc indispensable de faire des choix qui peuvent être motivés par les données (suivre les choix du corpus utilisé) ou avoir pour but de simplifier la tâche. Certains systèmes peuvent préférer se concentrer sur une seule partie du phénomène en ne souhaitant reconnaître qu'un seul type de mention ou qu'un seul type de relation par exemple. Les systèmes aspirant à représenter la coréférence dans son ensemble, doivent également définir

4. Corpus DEMOCRAT disponible sur la plateforme ORTOLANG [Lien]

précisément les annotations qu'ils réalisent afin de cadrer leur système et d'assurer une cohérence.

Les systèmes entraînés et évalués sur ANCOR comme CROC (Coreference Resolution for Oral Corpus, [Désoyer et al., 2015]) ou DECOFRE (Detecting Coreferences for Oral French, [Grobol, 2019]) doivent suivre les décisions prises au moment de l'annotation de ce corpus et ne prennent donc pas en compte les pronoms déictiques qui réfèrent aux interlocuteurs. Les systèmes réalisés lors de la tâche CoNLL-2012 [5] se basent sur des données pour lesquelles les singletons ne sont pas annotés comme des mentions. C'est d'ailleurs une des raisons pour lesquelles un des modèles créés dans le cadre de cette tâche partagée a été repris et modifié, afin de créer un système, DECOFRE, qui serait plus facilement applicable à d'autres jeux de données et qui prendrait en compte les singletons ([Grobol, 2019]).

Les premiers systèmes de résolution de coréférence étaient à base de règles, mais ces dernières années ce sont plutôt les méthodes statistiques ([Désoyer et al., 2015]) et plus récemment neuronales ([Grobol, 2019]) qui sont utilisées. Le système peut considérer que les mentions et les informations y étant associées sont déjà connues et ne se concentrer que sur la deuxième étape : la détection des relations ([Désoyer et al., 2015]), ou réaliser les deux étapes (systèmes end-to-end [Grobol, 2019], [Oberle, 2019]). Ces deux étapes peuvent utiliser des méthodes différentes comme [Oberle, 2019] qui emploie des méthodes statistiques pour la détection de mentions et des règles pour identifier les chaînes de coréférence.

Les systèmes qui utilisent l'apprentissage automatique peuvent être basés sur différents modèles ([Désoyer et al., 2015]) :

— Les modèles mention-pair ou pairwise, classification de paires constituées d'une mention et d'un antécédent, afin de déterminer si elles sont coréférentes. Suite à cette classification, un filtrage permet de ne garder qu'un antécédent par mention.

— Les modèles twin-candidate, classification d'instances composées d'une mention et de deux antécédents potentiels dont le but est de déterminer lequel de ces deux antécédents est coréférent avec la mention.

— Les modèles mention-ranking, classement des antécédents possibles pour une mention donnée.

— Les modèles entity-mention, déterminer la probabilité qu'une mention réfère à une entité précédemment repérée.

L'outil que nous utiliserons dans nos expériences, OFCORS (voir 4.2), se base sur le modèle de détection de mention de DECOFRE et utilise des méthodes statistiques pairwise pour la résolution des relations (voir 4.2).

Enfin, les deux étapes de la tâche de résolution de coréférence sont évaluées séparément. L'étape de détection de mentions est évaluée avec les mesures de précision, rappel et F-mesure et la seule difficulté réside dans le fait de déterminer ce qu'est une mention correcte (maximal et minimal span). En revanche, l'évaluation de l'étape d'identification des chaînes de coréférence n'est pour le moment pas standardisée. Il y a eu de nombreuses propositions (ex : MUC, B3, CEAF, CoNLL, BLANC, ...) dont chacune présente des avantages et des inconvénients. Cette

5. Tâche partagée CoNLL-2012 [Lien]

absence de mesure standard peut parfois rendre difficile la comparaison entre les systèmes ([Grobol, 2020]).

La coréférence est donc un phénomène complexe et hétérogène pour lequel le développement de systèmes basés sur des méthodes statistiques et neuronales pour le français a été rendu possibles grâce à la publication en 2014 et en 2020 de deux corpus de taille suffisante annotés manuellement en chaînes de coréférences.

CHAPITRE 2

LES EXPRESSIONS POLYLEXICALES

Sommaire

Les expressions polylexicales (*multiword expression*, *MWE* en anglais) sont fréquemment utilisées dans les discours et représentent un défi dans les tâches de traitement automatique des langues comme la fouille de texte, la traduction automatique ou encore la recherche d'informations. Dans la suite de ce mémoire, nous pourrons utiliser le terme *expression* pour désigner les expressions polylexicales. Nous commencerons par définir ce phénomène, puis nous présenterons quelques travaux récents d'identification d'expressions polylexicales verbales.

2.1 Définitions

Les expressions polylexicales sont des termes complexes composés de plusieurs mots tels que *blanc d'œuf*, *mémoire vive*, *prendre une pause* ou encore *prendre le taureau par les cornes*. Elles présentent diverses propriétés comme la non-compositionnalité sémantique, ce qui signifie que le sens de l'expression ne peut pas être obtenu de manière régulière à partir des sens de ses composants (ex : *cordon bleu* [1]) ou encore l'ambiguïté, les expressions pouvant être utilisées de façon littérale (ex : *Il a retourné sa veste avant de la mettre à la machine à laver*). [Constant et al., 2017] soulignent également d'autres propriétés pouvant présenter un challenge lors de traitements informatiques, comme la discontinuité, des éléments pouvant s'intercaler entre les composants de l'expression (ex : *C'est l'arbre qui cache la forêt.* / *Ne serait-ce pas l'arbre qui cache la forêt ?*) ou encore leur variabilité et notamment celle des expressions verbales.

Il existe de nombreux types d'expressions polylexicales et il est difficile de les classer. Nous allons nous concentrer au cours de ce mémoire sur les expressions polylexicales verbales du français, mais il existe également des expressions polylexicales nominales (*bouc émissaire, gros mots*...) ou encore adverbiales (*en long, en large et en travers, en revanche*...). [Laporte, 2018] souligne l'importance de choisir des critères de classification basés sur des faits linguistiques (ex : la partie du discours) plutôt que des critères se basant sur l'intuition qui pourraient varier d'un annotateur à un autre (ex : le poids sémantique d'un verbe).

1. Les expressions polylexicales seront soulignées dans les exemples tout au long du mémoire.

Il y a différents types d'expressions polylexicales verbales en français. Dans ce mémoire nous allons reprendre les catégories proposées par PARSEME [2] (PARSing and Multiword Expressions) :

- LVC (Light Verb Construction) : formé d'un verbe et d'un nom qui dépend directement du verbe ou qui est introduit par une préposition. Ce nom fait référence à un événement ou à un état. Il y a deux types de LVC :
  - LVC.full, le verbe n'a pas de valeur sémantique et n'apporte que des informations grammaticales (personne, temps, aspect . . .). Le sujet du verbe doit être l'argument sémantique du nom.
    Faire une présentation, procéder à une analyse, faire une visite
  - LVC.cause, le verbe indique que son sujet est la cause ou la source de l'événement ou de l'état exprimé par le nom. Le sujet du verbe ne doit pas être un argument sémantique du nom.
    Donner le vertige, provoquer un accident, donner le droit
- IRV (Inherently Reflexive Verb) : verbes pronominaux qui ne sont jamais utilisés sans leur pronom réfléchi ou pour lesquels la version sans pronom a un sens différent de la version avec le pronom ou un cadre de sous-catégorisation [3] différent.
  s'apercevoir, s'évanouir
- MVC (Multi-Verb Construction) : formé d'une séquence de deux verbes adjacents : un verbe gouverneur et un verbe dépendant, avec une inflexibilité lexicale sur le verbe dépendant.
  laisser tomber, vouloir dire
- VID (Verbal Idiom) : formé d'au moins deux composants lexicalisés : le verbe tête et au moins un de ses dépendants (sujet, objet, compléments circonstanciels. . .). Il se caractérise par une inflexibilité lexicale, morphologique ou syntaxique.
  l'emporter, se faire des idées, court-circuiter

Les annotations du corpus PARSEME se basent sur des arbres de décisions.

Nous considérons comme composant d'une expression tout token annoté comme appartenant à une expression polylexicale.

« *J'étais prêt à partir, malgré mon attachement à l'ASNL et à la ville, car à 32 ans, il s'agissait sans doute de ma dernière chance de signer un nouveau contrat.* » (AnnodisER)

Dans cette phrase, *contrat* est un composant de l'expression polylexicale *signer contrat*.

2. Réseau scientifique consacré au rôle des expressions polylexicales dans l'analyse syntaxique. [Lien] - Guide d'annotation de PARSEME : version 1.2 [Lien]

3. Manière dont les arguments syntaxiques du verbes sont introduits (prépositions, postpositions, marqueurs).

2.2 Méthodes d'identification des expressions polylexicales verbales

L'identification d'expressions polylexicales correspond à l'annotation automatique d'expressions dans un texte, en les associant à un type d'expression existant ([Constant et al., 2017]).

Lors de la tâche partagée de PARSEME 1.2, plusieurs équipes ont proposé des systèmes d'identification d'expressions polylexicales en travaillant sur des données couvrant 14 langues ([Ramisch et al., 2020]). L'objectif de cette édition était de proposer des systèmes semi-supervisés qui peuvent identifier les expressions polylexicales qui ont déjà été annotées dans le corpus d'entraînement et celles inconnues.

Certains ont proposé des systèmes neuronaux basés sur les modèles pré-entraînés de BERT [4] en considérant la tâche comme une classification de tokens ([Taslimipoor et al., 2020], [Kurfali, 2020]). MTLB-STRUCT ([Taslimipoor et al., 2020]) utilise le modèle multilingue de BERT pré-entraîné sur 104 langues, tandis que TRAVIS ([Kurfali, 2020]) est proposé en deux versions : une utilisant le modèle multilingue (TRAVIS-multi) et une autre utilisant un modèle spécifique pour chaque langue prise en compte (TRAVIS-mono). La F-mesure [5] globale de MTLB-STRUCT est de 0,74 ce qui en fait le meilleur système proposé pour cette tâche et celles de TRAVIS-multi et TRAVIS-mono sont de 0,65 et 0,50 ([Ramisch et al., 2020]). Les résultats détaillés montrent cependant que TRAVIS-mono est plus efficace que la version multilingue pour les langues qu'il couvre ([Kurfali, 2020]).

D'autres systèmes sont à base de règles, comme [Pasquer et al., 2020] avec Seen2Seen. La stratégie de Seen2Seen de faire un système à base de règles le rend facilement interprétable, car il est basé sur des filtres morphosyntaxiques. Alors que ce système n'a pas été conçu pour trouver les expressions inconnues, sa F-mesure globale pour la tâche d'identification des expressions déjà vues et inconnues est de 0,66. Ce résultat montre que l'absence de traitement pour les expressions inconnues est compensée par de bonnes capacités sur l'identification de celles déjà connues ([Pasquer et al., 2020]). En effet, les expressions polylexicales inconnues sont bien moins nombreuses dans les corpus.

Pour réaliser notre tâche nous allons utiliser l'outil Seen2Seen qui a de bons résultats tout en étant interprétable. Nous considérons que le corpus d'entraînement de cet outil est suffisamment fourni en expressions polylexicales pour compenser l'absence de traitement des expressions inconnues.

Les expressions polylexicales sont donc des éléments importants à prendre en compte dans les traitements informatiques et les récents travaux permettent d'obtenir de bonnes performances pour leur identification.

4. Bidirectional Encoder Representations from Transformers : modèle de langage basé sur l'apprentissage profond (utilisant plusieurs couches de neurones artificiels) et adoptant la technique de l'auto-attention (privilégiant certaines connexions entre neurones de la même couche).

5. Mesure dite token-based, basée sur le nombre de tokens correctement reconnus et non sur les expressions complètes. Par exemple, si le système reconnait uniquement *prendre le taureau* dans l'expression *prendre le taureau par les cornes*, il comptera 3 vrais positifs (*prendre*, *le*, *taureau*) et 3 faux négatifs (*par*, *les*, *cornes*).

CHAPITRE 3

L'HYPOTHÈSE

L'hypothèse autour de laquelle nous allons orienter nos réflexions au cours de ce mémoire consiste à supposer que les composants d'une expression polylexicale ne sont que rarement susceptibles d'appartenir à des chaînes de coréférence de plus d'un élément (non singletons).

« *Il a retourné sa veste avant de la mettre à la machine à laver.* »

Dans cet exemple, nous ne sommes pas en présence de l'expression polylexicale *retourner sa veste*. Si tel avait été le cas, la non-compositionnalité sémantique de celle-ci aurait empêché le composant *sa veste* d'être accessible à une coréférence. Or, dans cette phrase *sa veste* et *la* sont coréférents. [Laporte, 2018] a étudié les caractéristiques linguistiques des expressions polylexicales afin de proposer des critères communs permettant de les reconnaître. Parmi ces critères, il propose notamment l'impossibilité d'appartenir à une chaîne de coréférence. Il illustre son point avec les exemples suivants :

(1) « *Kathy avait une posture fière . Cette posture a été commentée.* »

(2) *« *Kathy était en mauvaise posture . Cette posture aurait pu être évitée.* »

Dans l'exemple (1), *posture fière* est accessible à la coréférence alors que dans l'exemple (2), il n'est pas possible de reprendre *mauvaise posture* par *cette posture*. Cette exemple illustre la théorie de [Laporte, 2018] car seule *mauvaise posture* est une expression polylexicale. L'impossibilité de coréférence pourrait donc permettre de reconnaître les expressions polylexicales.

La reprise de composant d'une expression étant un phénomène supposé rare, voire impossible dans certains cas, la vérification de cette hypothèse devra se faire sur des corpus larges et de thèmes ou de genres variés.

Deuxième partie

EXPÉRIMENTATIONS

CHAPITRE 4

OUTILS UTILISÉS

Sommaire

Lors de nos expériences nous avons principalement utilisé deux outils, Seen2Seen pour l'identification des expressions polylexicales et OFCORS pour la résolution de chaînes de coréférence. Nous commencerons par présenter Seen2Seen puis OFCORS.

4.1 Seen2Seen

Seen2Seen [1] est un outil d'identification des expressions polylexicales verbales dans différentes langues, développé principalement par Caroline Pasquer. C'est un outil à base de règles qui couvre les langues suivantes : le bulgare, le basque, le français, le polonais, le portugais brésilien, le roumain, l'allemand, le grec, l'hébreu, l'italien, le turc, l'irlandais, le hindi et le suéduois. Pour les travaux présentés dans ce mémoire, nous n'avons utilisé cet outil que sur la langue française.

FIGURE 4.1 – Extrait d'un fichier cupt avant l'annotation par Seen2Seen

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC PARSEME:MWE
# newdoc = 1AG0154.txt
# source_sent_id = http://my/newcorpus/uri 1AG0154.txt 1
# text = je peux m' occuper de quelqu'un
1   je  il  PRON    _   Number=Sing|Person=1|PronType=Prs   2   nsubj   _   _   _
2   peux    pouvoir VERB    _   Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin   0   root    _   _   _
3   m   m   NOUN    _   Gender=Masc|Number=Plur 2   obj _   SpaceAfter=No   _
4   '   '   PUNCT   _   _   3   punct   _   _   _
5   occuper occuper VERB    _   VerbForm=Inf   3   acl _   _   _
6   de  de  ADP _   _   7   case    _   _   _
7   quelqu'un   quelqu'un   PRON    _   Gender=Masc|Number=Sing 3   nmod    _   SpacesAfter=\s\n\n  _
```

Les fichiers d'entrée et de sortie sont au format cupt (figure 4.1). Ce sont des fichiers CoNLL-U avec une colonne supplémentaire pour l'annotation en expressions. Seen2Seen prend donc en entrée un fichier annoté en morphosyntaxe et y ajoute les annotations en expressions polylexicale.

L'identification des expressions polylexicales verbales est une tâche difficile, notamment en raison de la variabilité des verbes, des changements possibles dans l'ordre des mots ou encore des discontinuités. Seen2Seen s'appuie sur les lemmes pour reconnaître une expression polylexicale à partir de celles annotées dans le cor-

1. Seen2Seen [Lien]

pus d'entraînement. Sa méthode se base sur quatre hypothèses linguistiques détaillées dans [Pasquer et al., 2020].

La première étape de l'identification d'expressions polylexicales par Seen2Seen est l'extraction des candidats correspondant potentiellement à une expression déjà rencontrée. Cette première phase privilégie le rappel aux dépens de la précision afin de ne manquer aucune expression potentielle. La deuxième phase a pour but d'augmenter la précision à l'aide de huit filtres. Ces filtres vont, entre autres, s'appuyer sur les parties du discours des composants, la longueur des discontinuités, les relations syntaxiques ou encore les imbrications.

Pour chaque langue, l'outil détermine durant l'entraînement la meilleure combinaison de filtres à activer sur le corpus d'évaluation, en testant les 256 possibilités sur le corpus de développement. La combinaison conservée est celle pour laquelle la F-mesure est la plus élevée. La meilleure combinaison de filtres pour le français, et donc celle que nous avons utilisé, est la configuration 48 où six filtres sont activés sur les huit existants, avec 0,79 de F-mesure.

Enfin, il faut noter que cet outil n'annote que les parties fixes de l'expression :

(1) « *Le 15 mars, le général donne l'ordre de mise de jugement directe des 24 hommes désignés.* » (FRWIKI)

(2) « *Plus tard, le bombardement des tranchées françaises fera l'objet d'une polémique, à la suite d'un témoignage : Le général Réveilhac, ordonnateur de l'attaque, aurait demandé à l'artillerie de pilonner les positions française pour obliger les soldats à sortir de leurs tranchées.* »

Dans l'exemple (1), l'expression annotée est *donner ordre* car il est possible de *donner des ordres*, *donner un ordre* ou encore *donner l'ordre*. En revanche dans l'exemple (2), l'expression annotée est *fera l'objet* car aucun de ces tokens ne peut être remplacé.

4.2 OFCORS

OFCORS [2] (Oral French COreference Resolution System) est un outil end-to-end de résolution de coréférence du français développé par le LIFAT dans le cadre du projet TALAD [3] par Théo Azzouza. Il identifie les mentions à partir du modèle de DECOFRE ([Grobol, 2019]) entraîné sur ANCOR et chaîne ensuite ces mentions avec des méthodes statistiques. Il utilise un modèle de type pairwise, ou mention-pair (voir 1.3). Plusieurs options sont disponibles et on peut notamment choisir le tokeniseur (*stanza*, *spacy* ou *spacy-stanza*), le mode de pairage (*consecutive*, *permutations*, *combinations* ou *window*), le classifieur ou encore le mode de chaînage (*closest first*, *best first* ou *graph*).

Nos traitements fonctionnent quelque soit le tokeniseur choisi, mais nous avons tout de même cherché à utiliser celui dont la tokenisation était la plus proche de celle de nos corpus déjà tokenisés. Après avoir comparé la tokenisation sur des mots tels que *week-end*, *Paris.* ou encore *du* (que *spacy-stanza* décontracte en *de* et *le*), nous avons décidé d'utiliser *stanza* pour nos traitements.

2. OFCORS [Lien]

3. Projet ANR visant à renforcer les travaux en collaboration entre les équipes de Traitement Automatique des Langues et d'Analyse du Discours.

Concernant le pairage, nous avons comparé les résultats des modes *combinations* (toutes les paires de mentions parmi l'ensemble des mentions) et *window* (les paires comprises dans un certain nombre de mentions vers l'avant selon la taille de la fenêtre) sur un extrait de fichier. Le mode *combinations* n'apportait que du bruit dans notre étude. Nous avons de la même manière comparé quelques tailles de fenêtre (5, 8 et 10) et déterminé que la taille 8 était la plus adaptée. Elle permet de ne pas louper certaines paires éloignées sans apporter trop de bruit. Nous avons donc utilisé OFCORS avec le mode *window* et une fenêtre de taille 8.

Nous avons laissé le mode par défaut pour le chaînage, *closest_first*, ce qui signifie que parmi les antécédents possibles d'une mention c'est celui le plus proche qui est retenu. Le mode *best_first* signifie que c'est l'antécédent ayant la plus forte probabilité d'être coréférent avec la mention qui est conservé.

OFCORS prend des fichiers en texte brut en entrée et produit plusieurs fichiers de sorties au format JSON dont le texte tokenisé, les mentions détectées et les chaînes de coréférence (en utilisant les indices des mentions).

FIGURE 4.2 – Extrait d'un fichier de sortie d'OFCORS indiquant les chaînes de coréférence


```
{"type": "clusters", "clusters": {"0": ["217", "208", "207", "201", "209"],
    "1": ["96", "97", "93"], "2": ["141", "135"], "3": ["58", "54", "50"], "
    4": ["113", "106", "112", "111", "101", "105", "103", "104"], "5": ["210
    ", "195", "189", "203", "206", "191", "215", "205", "184", "183", "194",
     "199", "213", "193"], "6": ["83", "78", "81", "75", "77", "76", "68"],
    "7": ["18", "10", "17"], "8": ["169", "172", "176", "171", "170"], "9":
    ["163", "168", "161", "160", "162"], "10": ["232", "237", "235"], "11":
    ["115", "120"], "12": ["124", "125"], "13": ["60", "66"], "14": ["164",
    "156"], "15": ["219", "226", "233"], "16": ["52", "49", "67", "42", "59"
    , "53"], "17": ["218", "220", "222"], "18": ["90", "91"], "19": ["45", "
    37"], "20": ["155", "149", "150", "146"], "21": ["131", "136"], "22": ["
    158", "157"], "23": ["40", "34"], "24": ["182", "174"], "25": ["224", "
    227"], "26": ["126", "134"], "27": ["196", "198"], "28": ["187", "192"],
     "29": ["9", "15"], "30": ["118", "110"], "31": ["159", "154"]}}
```


Les chaînes de coréférence correspondent toutes à un identifiant et leurs composants sont indiqués à l'aide des identifiants de mentions. Par exemple, la chaîne 2 de la figure 4.2 est composée des mentions 141 et 135. On peut ensuite retrouver les tokens qui correspondent à ces mentions à l'aide du fichier sur les mentions détectées. Les singletons ne sont pas présentés comme des chaînes par OFCORS.

Nous noterons également qu'OFCORS a des performances correctes mais que nous travaillons sur un phénomène rare et qu'il est donc attendu que la plupart des résultats susceptibles de correspondre à ce que l'on cherche soient de faux positifs.

Ces deux outils n'utilisant pas les mêmes formats d'entrée et de sorties, des conversions seront nécessaires afin de pouvoir les utiliser sur les mêmes corpus.

CHAPITRE 5

CHOIX DES CORPUS

Sommaire

Le choix des corpus est une étape importante, car ils doivent être adaptés autant à la résolution de coréférence qu'à l'identification des expressions polylexicales. Nous allons présenter nos critères de sélection ainsi que les corpus utilisés.

5.1 Critères de sélection

Les corpus utilisés doivent répondre à certains critères dont les plus importants sont le respect de l'ordre des phrases du document original et l'existence de frontières entre les textes. En effet, la coréférence ne peut être traitée qu'à l'intérieur d'une unité discursive.

Il faut également prendre en compte le fait que les outils d'identification d'expression polylexicales et de résolution de coréférence ne sont pas parfaits et font des erreurs. Par conséquent, il sera plus simple de travailler avec des corpus annotés manuellement pour au moins un des phénomènes étudiés.

Enfin, la reprise coréférentielle d'un élément d'une expression polylexicale étant un phénomène que nous supposons rare voire impossible, nous devons disposer de corpus larges et ayant des thèmes ou des genres variés afin de vérifier notre hypothèse de départ.

5.2 Corpus utilisés

5.2.1 PARSEME

Le premier corpus que nous avons utilisé est le corpus PARSEME 1.2 [1] qui a servi à l'entraînement de Seen2Seen. Il est annoté manuellement en expressions polylexicales selon le guide PARSEME et contient environ 20 961 phrases et 5 654 mentions annotées pour le français. Cependant, il n'est possible de retrouver l'ordre des phrases

1. Corpus PARSEME 1.2 [Lien]

que pour certains sous-corpus de SEQUOIA, SEQUOIA [2] étant un des sous-corpus qui composent le corpus PARSEME avec GSD, PARTUT et PUD.

Le sous-corpus SEQUOIA contient 3 099 phrases qui sont réparties en quatre sous-corpus (Europar, annodisER, FRWIKI et EMEA), et nous avons pu retrouver l'ordre des phrases pour trois d'entre eux. Les différents documents présents dans ces sous-corpus pour lesquels l'ordre des phrases correspondait à celui du document initial ont donc été isolés :

— 2 rapports de l'agence européenne du médicament (EMEA, 1000 phrases environ),

— 19 articles Wikipédia sur des affaires sociales ou politiques (FRWIKI, 1000 phrases environ),

— 36 articles de journaux courts avec des thèmes variés (annodisER, 500 phrases environ).

FIGURE 5.1 – Extrait du corpus SEQUOIA

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC PARSEME:MWE
# source_sent_id = http://hdl.handle.net/11234/1-3105 UD_French-Sequoia/fr_sequoia-ud-dev.conllu annodis.er_00228
# text = L'EBM s'inclina devant Joeuf par 70-61.
1   L'  le  DET _   Definite=Def|Number=Sing|PronType=Art   2   det _   SpaceAfter=No   *
2   EBM EBM PROPN   _   Number=Sing 4   nsubj   _   _   *
3   s'  se  PRON    _   Person=3|Reflex=Yes 4   expl:comp    _   SpaceAfter=No   1:IRV
4   inclina incliner    VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   0   root    _   _   1
5   devant  devant  ADP _   _   6   case    _   _   *
6   Joeuf   Joeuf   PROPN   _   Gender=Masc|Number=Sing 4   obl:mod _   _   *
7   par par ADP _   _   8   case    _   _   *
8   70-61   70-61   NOUN    _   NumType=Card    4   obl:mod _   SpaceAfter=No   *
9   .   .   PUNCT   _   _   4   punct   _   _   *

# source_sent_id = http://hdl.handle.net/11234/1-3105 UD_French-Sequoia/fr_sequoia-ud-train.conllu annodis.er_00229
# text = La finale se joua entre Utrecht et Soleuvre.
1   La  le  DET _   Definite=Def|Gender=Fem|Number=Sing|PronType=Art    2   det _   _   *
2   finale  finale  NOUN    _   Gender=Fem|Number=Sing  4   nsubj   _   _   1:LVC.full
3   se  se  PRON    _   Person=3|Reflex=Yes 4   expl:pass    _   _   *
4   joua    jouer   VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   0   root    _   _   1
5   entre   entre   ADP _   _   6   case    _   _   *
6   Utrecht Utrecht PROPN   _   _   4   obl:arg _   _   *
7   et  et  CCONJ   _   _   8   cc  _   _   *
8   Soleuvre    Soleuvre    PROPN   _   _   6   conj    _   SpaceAfter=No   *
9   .   .   PUNCT   _   _   4   punct   _   _   *
```

Ces documents récupérés sont au format cupt et contiennent les informations suivantes : l'identifiant du token dans la phrase (ID), le token (FORM), le lemme du token (LEMMA), les parties du discours (UPOS et XPOS), des informations morphologiques (FEATS), la tête du token (HEAD), ses relations de dépendance (DEPREL, DEPS) ainsi que l'annotation en expressions polylexicales (PARSEME :MWE). Dans la figure 5.1 on observe qu'un IRV a été annoté dans la première phrase (*se incliner*) et une LVC.full dans la seconde phrase (*finale jouer*).

5.2.2 ANCOR

Nous avons ensuite travaillé sur le corpus ANCOR, un corpus d'oral transcrit annoté manuellement en mentions et chaînes de coréférence ayant servi à l'entraînement de DECOFRE, utilisé par OFCORS. Il contient 488 000 mots et est composé des quatre sous-corpus suivants ([Muzerelle et al., 2014]) :

— ESLO_ANCOR : des entretiens divisés en sous dialogues thématiques cohérents (25 000 phrases environ).

— ESLO_CO2 : trois entretiens complets (2 500 phrases environ).

2. Corpus SEQUOIA [Lien]

— OTG : des dialogues interactifs entre des individus et le personnel d'accueil de l'office de tourisme de Grenoble (2 800 phrases environ).

— UBS : des dialogues interactifs par téléphone recueillis auprès du standard téléphonique d'une université (700 phrases environ).

FIGURE 5.2 – Extrait du corpus ANCOR

```
      <tei:w xml:id="s1.u1.w8">il</tei:w>
      <tei:w xml:id="s1.u1.w9">vous</tei:w>
      <tei:w xml:id="s1.u1.w10">plaît</tei:w>
    </tei:u>
    <tei:u xml:id="s1.u2" tei:who="#spk1" tei:start="#s1.u2.t0" tei:end="#s1.u2.t2">
      <tei:w xml:id="s1.u2.w0" tei:join="right">j'</tei:w>
      <tei:w xml:id="s1.u2.w1">en</tei:w>
      <tei:w xml:id="s1.u2.w2">ai</tei:w>
      <tei:w xml:id="s1.u2.w3">plus</tei:w>
      <tei:w xml:id="s1.u2.w4">il</tei:w>
      <tei:w xml:id="s1.u2.w5">faut</tei:w>
      <tei:w xml:id="s1.u2.w6">attendre</tei:w>
      <tei:w xml:id="s1.u2.w7">1</tei:w>
      <tei:w xml:id="s1.u2.w8">mois</tei:w>
      <tei:w xml:id="s1.u2.w9">et</tei:w>
      <tei:w xml:id="s1.u2.w10">demi</tei:w>
      <tei:w xml:id="s1.u2.w11">ils</tei:w>
      <tei:w xml:id="s1.u2.w12">sont</tei:w>
      <tei:w xml:id="s1.u2.w13">en</tei:w>
      <tei:w xml:id="s1.u2.w14">train</tei:w>
      <tei:w xml:id="s1.u2.w15">de</tei:w>
      <tei:w xml:id="s1.u2.w16">les</tei:w>
      <tei:w xml:id="s1.u2.w17">rééditer</tei:w>
    </tei:u>
  </tei:div>
 </tei:body>
</tei:text>
<tei:standOff>
 <tei:annotation tei:type="coreference">
  <tei:spanGrp tei:type="unit" tei:subtype="mention">
   <tei:span xml:id="u-MENTION-sduchon_1329158313025" tei:from="#s1.u2.w0" tei:to="#s1.u2.w0" tei:ana="#u-MENTION-sduchon_1329158313025-fs"/>
   <tei:span xml:id="u-MENTION-sduchon_1329158318664" tei:from="#s1.u1.w0" tei:to="#s1.u1.w0" tei:ana="#u-MENTION-sduchon_1329158318664-fs"/>
   <tei:span xml:id="u-MENTION-sduchon_1329158365230" tei:from="#s1.u2.w11" tei:to="#s1.u2.w11" tei:ana="#u-MENTION-sduchon_1329158365230-fs"/>
   <tei:span xml:id="u-MENTION-sduchon_1329158519147" tei:from="#s1.u2.w16" tei:to="#s1.u2.w16" tei:ana="#u-MENTION-sduchon_1329158519147-fs"/>
   <tei:span xml:id="u-MENTION-sduchon_1329158554534" tei:from="#s1.u1.w4" tei:to="#s1.u1.w6" tei:ana="#u-MENTION-sduchon_1329158554534-fs"/>
   <tei:span xml:id="u-MENTION-sduchon_1329158629380" tei:from="#s1.u2.w7" tei:to="#s1.u2.w10" tei:ana="#u-MENTION-sduchon_1329158629380-fs"/>
   <tei:span xml:id="u-MENTION-sduchon_1329158736190" tei:from="#s1.u2.w1" tei:to="#s1.u2.w1" tei:ana="#u-MENTION-sduchon_1329158736190-fs"/>
   <tei:span xml:id="u-MENTION-sduchon_1329158541546" tei:from="#s1.u1.w2" tei:to="#s1.u1.w6" tei:ana="#u-MENTION-sduchon_1329158541546-fs"/>
  </tei:spanGrp>
```

L'ordre des phrases n'a pas été modifié dans ce corpus, et les fichiers sont au format XML TEI. La figure 5.2 présente un extrait de l'annotation d'un dialogue dans ANCOR dont les différents tour de paroles sont découpés en unités (s1.u1, s1.u2 ...). L'annotation des mentions utilise les identifiant des mots (s1.u2.w0, s1.u1.w0 ...) pour indiquer les tokens de début et de fin des mentions.

5.2.3 Est Républicain

Enfin, nous avons également travaillé avec le corpus Est Républicain [3], qui réuni les articles du journal régional Est Républicain parus en 1999, 2002 et les deux premiers mois de 2003. Ce corpus est de taille importante : environ 36 millions de mots pour l'année 1999, 98 millions de mots pour 2002 et 15 millions de mots pour 2003. Son traitement complet aurait demandé trop de temps et de puissance de calcul. Nous nous sommes donc concentrés sur les 100 premiers articles ayant plus de 300 mots de l'année 2003 (3000 phrases environ). Les fichiers de ce corpus n'ont aucune annotation manuelle.

3. Corpus EstRépublicain [Lien original, Lien vers REDAC]

FIGURE 5.3 – Extrait du corpus Est Républicain

```
<estRepublicain date="2003-01-14">

<head>Pages spécifiques à l'édition de Bar le Duc</head>
En ne poussant pas hier au dépôt de bilan, le gouvernement a donné une dernière chance de 48 heures à
Air Lib. Demain soir, la compagnie devra se remettre à payer ses charges courantes. Le ministre de
Robien, pessimiste sur la viabilité du plan de restructuration, souligne cependant que l'investisseur
néerlandais IMCA présidée par Erik De Vlieger a promis de renouveler la flotte de la compagnie et
d'apporter des capitaux frais.

<head>Fains-Véel : échec aux cambrioleurs dans une grande surface</head>
En Meuse

<head>Le NANCie au bord du gouffre financier</head>
En Région, l'article de Patrice COSTA

<head>Seillière candidat à sa réelection</head>
En Economie, l'article de Jean-Louis DENES
```

La figure 5.3 montre un extrait du corpus Est Républicain. La date de publication des articles se trouve au début du document (*2003-01-14*) et les titres sont indiqués par des balises (<*head*>).

Nos traitements ont donc été appliqués sur environ 36 500 phrases en tout.

CHAPITRE 6

TRAITEMENTS

Sommaire

Le but de nos expériences est de vérifier que les composants d'une expression polylexicale ne peuvent que rarement être repris dans des chaînes de coréférence. Pour ce faire, nous disposons de deux outils : Seen2Seen pour annoter automatiquement en expressions polylexicales et OFCORS pour annoter automatiquement en chaînes de coréférence. Les deux outils n'ont pas les mêmes formats d'entrée et de sortie. Nous disposons également de trois corpus avec des spécificités différentes. Nous allons à présent décrire notre chaîne de traitements et définir les cas possibles d'intersection entre une mention et une expression.

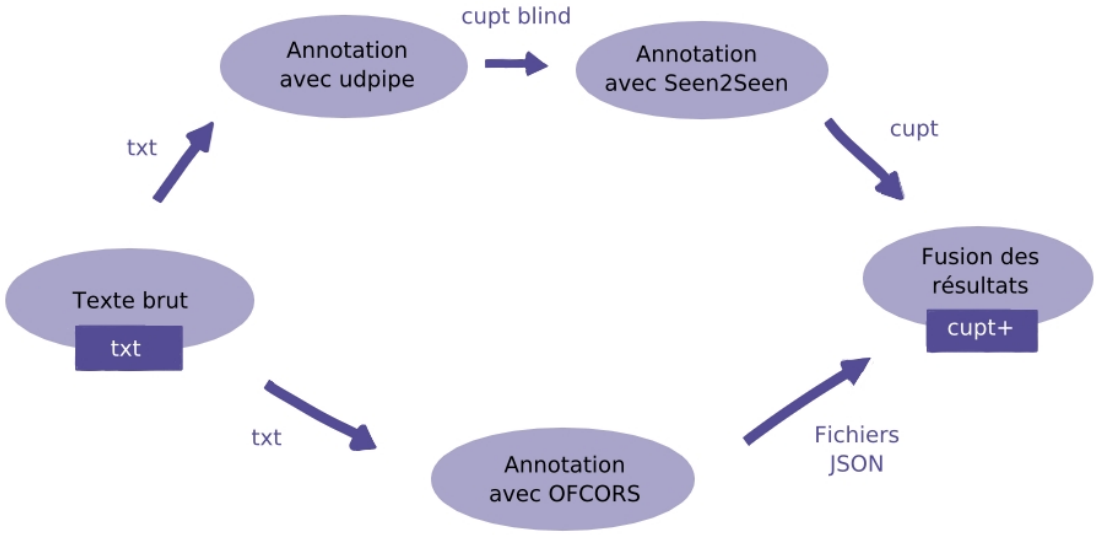
6.1 La chaîne de traitements

Nous commençons par annoter les corpus avec Seen2Seen. Cette étape n'est pas nécessaire pour les sous-corpus de PARSEME qui sont annotés manuellement en expressions polylexicales. En sortie nous récupérons donc des fichiers au format cupt.

Ensuite nous utilisons OFCORS pour récupérer les mentions et les chaînes de coréférence des corpus. Nous récupérons également un fichier contenant le texte tokenisé qui nous servira à aligner les fichiers de sorties des deux outils. Pour ANCOR, l'outil OFCORS ne doit pas être utilisé car le corpus est déjà annoté en coréférence. Cependant, ces informations étant stockées au format TEI, nous avons écrit un script pour les récupérer dans des fichiers au même format que ceux d'OFCORS afin de simplifier la fusion des résultats.

Une fois que l'on dispose d'un fichier cupt annoté en expressions polylexicales et des trois fichiers JSON contenant les indices des mentions, des chaînes de coréférence et de chaque token, on utilise le script principal (*lanceur.sh*) pour fusionner les différentes annotations. Au cours de la fusion, un fichier cupt contenant deux colonnes supplémentaires est alors créé (appelé cupt+ sur la figure 6.1). La première colonne ajoutée est celle des mentions : si le token appartient à une mention alors l'identifiant de cette mention y est renseigné, une étoile sinon. La seconde colonne

FIGURE 6.1 – Schéma représentant la chaîne de traitement



ajoutée est celle de la coréférence : si le token appartient à une chaîne de coréférence alors on indique l'indice de la chaîne et de la mention concernée (un token pouvant appartenir à plusieurs mentions). Il est également possible qu'un token soit contenu dans plusieurs chaînes de coréférence. Les colonnes qui nous permettront de tirer des conclusions dans notre étude sont donc les trois dernières (expressions polylexicales, mentions et coréférence).

FIGURE 6.2 – Extrait d'un fichier obtenu après fusion des annotations (EMEA)

```
32  entama  entamer VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   0   root    _   _   *   *   *
33  un  un  DET _   Definite=Ind|Gender=Masc|Number=Sing|PronType=Art   34  det     _   _   *   219 60:219
34  combat  combat  NOUN    _   Gender=Masc|Number=Sing 32  obj _   _   *   219 60:219
35  pour    pour    ADP _   37  case    _   _   *   *   *
36  la  le  DET _   Definite=Def|Gender=Fem|Number=Sing|PronType=Art    37  det _   _   *   220 25:220
37  réhabilitation  réhabilitation  NOUN    _   Gender=Fem|Number=Sing  34  nmod    _   _   *   220 25:220
38  de  de  ADP _   40  case    _   _   *   220 25:220
39  son son DET _   Number=Sing|Poss=Yes    40  det _   _   *   220;221 25:220
40  époux   époux   NOUN    _   Gender=Masc|Number=Sing 37  nmod    _   _   *   220;221 25:220
41  et  et  CCONJ   _   _   45  cc  _   _   *   *   *
42-43   des _   _   _   _   _   _   _   *   *   *
42  de  de  ADP _   45  case    _   _   *   222 *
43  les le  DET _   Definite=Def|Number=Plur|PronType=Art   45  det _   _   *   222 *
44  autres  autre   ADJ _   Number=Plur 45  amod    _   _   *   222 *
45  caporaux    caporal NOUN    _   Gender=Masc|Number=Plur 40  conj    _   _   *   222 *
46  fusillés    fusiller    VERB    _   Gender=Masc|Number=Plur|Tense=Past|VerbForm=Part    45  acl _   _   *   222 *
47  de  de  ADP _   48  case    _   _   *   *   *
48  Souain  Souain  PROPN   _   45  nmod    _   SpaceAfter=No   *   223 *
49  ;   ;   PUNCT   _   _   32  punct   _   _   *   *   *

# source_sent_id = http://hdl.handle.net/11234/1-3105 UD_French-Sequoia/fr_sequoia-ud-train.conllu frwiki_50.1000_00027
# text = combat contre les institutions, mené sans relâche, qui dura près de deux décennies et qui, en dehors de son activité
d'institutrice, l'occupa à plein temps.
1   combat  combat  NOUN    _   Gender=Masc|Number=Sing 0   root    _   _   1:LVC.full  224 60:224
2   contre  contre  ADP _   4   case    _   _   *   *   *
3   les le  DET _   Definite=Def|Number=Plur|PronType=Art   4   det _   _   *   225;226 *
4   institutions    institution NOUN    _   Gender=Fem|Number=Plur  1   nmod    _   SpaceAfter=No   *   225;226 *
5   ,   ,   PUNCT   _   _   1   punct   _   _   *   226 *
6   mené    mener   VERB    _   Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part    1   acl _   _   1   *   *
7   sans    sans    ADP _   8   case    _   _   *   *   *
8   relâche relâche NOUN    _   Number=Sing 6   obl:mod    SpaceAfter=No   *   227;228 3:228
```

On peut voir sur la figure 6.2 que l'on a une LVC.full dans la deuxième phrase (*combat mené*). Celle ci contient la mention 224 (*combat*) qui fait partie de la chaîne de coréférence 60 (*60 :224*). Le composant *combat* de l'expression polylexicale *combat mené* est donc une mention qui fait partie de la chaîne de coréférence 60 avec la mention 219 (*combat*) de la phrase précédente.

Le fichier cupt précédemment créé est ensuite utilisé pour ne récupérer que les

résultats qui nous intéressent : les phrases contenant une expression dont un des composants est repris dans une chaîne de coréférence selon nos outils. Un fichier JSON est créé (voir figure 6.3), organisé selon les types d'expressions polylexicales (VID, LVC.full, LVC.cause, MVC, IRV). Pour chaque type est renseigné le nombre total d'expressions trouvées appartenant à cette catégorie, le nombre d'expressions dont un des composants est repris selon les outils, et la liste de ces dernières. On voit donc dans la figure 6.3 que 77 LVC.full ont été trouvées dans le corpus AnnodisER et que 28 d'entre elles appartiennent possiblement à une chaîne de coréférence. Pour chacune des expressions est indiqué le fichier dans lequel elle se trouve, la phrase contenant l'expression, ses tokens et leurs lemmes, les indices de sa chaîne de coréférence et de sa mention, le ou les tokens concernés par la reprise, le cas d'intersection entre l'expression et la mention (voir 6.2) et la chaîne entière.

FIGURE 6.3 – Extrait du fichier de résultats du corpus AnnodisER


```
{
    "LVC.full": {
        "TYPE": "LVC.full",
        "COREF": "28/77",
        "MWES": [
            {
                "FICHIER": "annodisER_10_mwe_coref.cupt",
                "PHRASE": "Il est à noter que le couple a eu la chance que son accident soit vu par des
                    témoins qui ont alerté les secours.",
                "TOKENS": "['eu', 'chance']",
                "LEMME": [
                    "avoir",
                    "chance"
                ],
                "COREF": "['*', '1:42']",
                "CAS": "{'42': 4}",
                "CHAINE(S)": {
                    "1": "{'42': ['la', 'chance'], '48': ['la', 'rue', 'de', 'Danjoutin'], '52': ['la',
                        'sortie', 'de', 'route']}"
                }
            },
```


Cependant, ces résultats doivent ensuite être vérifiés manuellement afin de ne conserver que les expressions dont un composant est effectivement repris. Il faut s'assurer que le composant de l'expression qui est repris a le même référent qu'au moins un des éléments de la chaîne de coréférence et vérifier la validité des expressions polylexicales à l'aide du guide d'annotation de PARSEME. Cette annotation manuelle a été réalisée lors de réunions, les annotateurs n'étaient donc pas indépendants les uns des autres. Il y a eu entre deux et sept annotateurs.

Les erreurs humaines étant plus rares que les erreurs dues aux annotations automatiques, nous avons principalement vérifié les chaînes de coréférence pour le corpus PARSEME et les expressions polylexicales pour le corpus ANCOR. En revanche, le corpus Est Républicain n'ayant aucune annotation manuelle, nous avons vérifié les deux phénomènes. Lors de cette annotation, nous ajoutons trois champs au fichier JSON :

- L'état de la validation, avec cinq valeurs possibles :
  - Vrai, lorsque le composant détecté dans l'expression polylexicale se trouve effectivement dans une chaîne de coréférence.
  - Faux, dans le cas contraire.
  - Non concerné, lorsque ce n'est pas un composant de l'expression polylexicale qui est repris mais l'expression entière.
  - Discutable, lorsque l'exemple peut être vrai ou faux selon l'interprétation.
  - Répétitions, lorsque la reprise du composant est due à des disfluences ou à

des répétitions complètes des expressions (notamment dans le corpus ANCOR).

— Le degré de compositionnalité, à partir de la réponse de quatre annotateurs indépendants à un test adapté aux expressions polylexicales verbales inspiré de [Cordeiro et al., 2019]. Trois valeurs sont possibles : faible, moyen et fort.

— La source d'erreur, lorsqu'il y en a une. Cinq sources d'erreurs sont possibles :
  — MWE incorrecte, lorsque l'expression détectée n'est pas une vraie expression polylexicale.
  — MWE littérale, lorsque l'expression détectée correspond à une lecture littérale (sous-cas de l'erreur précédente).
  — MWE type incorrect, lorsque l'expression détectée est correcte, mais qu'elle n'est pas correctement classée.
  — Chaîne incorrecte, lorsque aucune des mentions de la chaîne n'est coréférente avec la mention de l'expression.
  — Mention incorrecte, lorsque la mention utilisée dans la chaîne détectée est incorrecte, mais que la chaîne serait correcte si la mention prenait en compte plus ou moins de tokens.

— Les lemmes ont ensuite été ajoutés (automatiquement) afin de faciliter l'utilisation des résultats.

En tout, 1 311 résultats ont été annotés manuellement.

FIGURE 6.4 – Extrait du fichier de résultat du corpus AnnodisER annoté manuellement


```
"LVC.full": {
    "TYPE": "LVC.full",
    "COREF": "28/77",
    "MWES": [
        {
            "FICHIER": "annodisER_34_mwe_coref.cupt",
            "PHRASE": "J'étais prêt à partir, malgré mon attachement à l'ASNL et à la ville, car à 32
                ans, il s'agissait sans doute de ma dernière chance de signer un nouveau contrat\".",
            "TOKENS": "['signer', 'contrat']",
            "COREF": "['*', '4:36']",
            "CAS": "{'36': 4}",
            "CHAINE(S)": {
                "4": "{'36': ['un', 'nouveau', 'contrat'], '42': ['un', 'an'], '43': ['il'], '44':
                    ['à', 'le', 'souhait', 'de', 'le', 'joueur'], '45': ['de', 'le', 'joueur'], '46':
                    ['qui'], '48': ['\"', 'Notre', 'objectif'], '49': ['Notre', 'objectif'], '50':
                    ['le', 'maintien', '.'], '51': ['Je']}"
            },
            "VALIDATION": "faux",
            "DEGRE DE COMPOSITIONNALITE": "",
            "SOURCE D'ERREUR": [
                "chaîne incorrecte"
            ],
            "LEMMES": [
                "signer",
                "contrat"
            ]
        },
```


6.2 Cas d'intersection entre une mention et une expression polylexicale

Il y a plusieurs cas possibles d'intersection entre une mention et une expression polylexicale. Il est important de pouvoir les différencier, car ils ne sont pas tous

concernés par notre étude. En effet, nous nous intéressons à la reprise des composants des expressions polylexicales et non à la reprise de l'expression complète.

6.2.1 Inclusion de l'expression dans la mention

Le premier cas correspond à l'inclusion de l'expression polylexicale dans la mention. Par exemple, dans la phrase *L'histologie osseuse a été évaluée 6 mois après le traitement par 5mg d'acide zolédronique chez 7 patients atteints de la maladie de Paget* tirée du corpus EMEA, les tokens de l'expression *atteint maladie* sont inclus dans la mention *7 patients atteints de la maladie de Paget.* (voir tableau 6.1).

TOKENS	EXPRESSION	MENTION
chez	*	*
7	*	1
patients	*	1
atteints	1	1
de	*	1
la	*	1
maladie	1	1
de	*	1
Paget	*	1

TABLE 6.1 – Exemple de cas d'inclusion de l'expression dans la mention tiré du corpus EMEA

Ce cas ne nous concerne pas car l'expression complète se trouve dans une mention englobante, et non un des composants.

6.2.2 Correspondance exacte entre l'expression et la mention

Le deuxième cas est celui où l'expression polylexicale est formée exactement des mêmes tokens que ceux de la mention. Par exemple, si on avait une phrase contenant l'expression *mise en évidence* et la mention *mise en évidence*, on serait en présence d'un cas de correspondance exacte (voir tableau 6.2).

TOKENS	EXPRESSION	MENTION
...	*	*
mise	1	1
en	1	1
évidence	1	1
...	*	*

TABLE 6.2 – Exemple de cas de correspondance exacte entre l'expression et la mention

Ce cas ne nous intéresse pas non plus, car l'expression complète serait reprise.

6.2.3 Inclusion de la mention dans l'expression

Le troisième cas correspond à l'inclusion de la mention dans l'expression polylexicale. Par exemple, dans la phrase *De nombreux protagonistes ont trouvé la mort depuis la signature du contrat.* tirée du corpus FRWIKI, la mention *la mort* est incluse dans l'expression *trouvé la mort* (voir tableau 6.3).

TOKENS	EXPRESSION	MENTION
de	*	*
nombreux	*	*
protagonistes	*	*
ont	*	*
trouvé	1	*
la	1	1
mort	1	1
depuis	*	*

TABLE 6.3 – Exemple de cas d'inclusion de la mention dans l'expression tiré du corpus FRWIKI

6.2.4 Chevauchement entre la mention et l'expression

Enfin, le quatrième cas correspond au chevauchement entre la mention et l'expression polylexicale. La mention et l'expression ont donc une partie en commun et une partie propre à chacune. Par exemple, dans la phrase *Ce dernier est pris en flagrant délit d'extorsion de fonds*., la mention *flagrant délit d'extorsion de fonds* et l'expression *pris en flagrant délit* ont une partie en commun (*flagrant délit*) et une partie propre à chacune (*pris en* et *d'extorsion de fonds*) (voir tableau 6.4).

TOKENS	EXPRESSION	MENTION
est	*	*
pris	1	*
en	1	*
flagrant	1	1
délit	1	1
d'	*	1
extorsion	*	1
de	*	1
fonds	*	1

TABLE 6.4 – Exemple de cas chevauchement entre la mention et l'expression tiré du corpus FRWIKI

Les parties annotées étant uniquement les parties fixes de l'expression polylexicale, il arrive souvent que le déterminant ne soit pas compris dans l'annotation mais fasse partie de la mention. Cette organisation correspond également à un chevauchement. Par exemple, dans la phrase *La commune a maintenu sa subvention de 150 euros qui est égale à celle des autres associations qui en ont fait la demande*., la mention *la demande* et l'expression *fait demande* ont une partie en commun (*demande*) et une partie propre à chacune (*fait* et *la*).

TOKENS	EXPRESSION	MENTION
en	*	*
ont	*	*
fait	1	*
la	*	1
demande	1	1

TABLE 6.5 – Exemple de cas 4 tiré du corpus Est Républicain

Les cas d'inclusion de la mention dans l'expression et de chevauchement nous intéressent car ce sont ceux pour lesquels la mention correspond à un composant d'une expression polylexicale.

Notre chaîne de traitement nous permet donc d'obtenir la liste des phrases contenant une expression polylexicale dont un des composant est repris dans une chaîne de coréférence, ainsi que d'autres informations utiles pour l'interprétation des résultats comme le type de l'expression, le nombre total d'expressions de ce type ou encore le cas d'intersection entre la mention et l'expression.

CHAPITRE 7

RÉSULTATS DES EXPÉRIENCES

Sommaire

Une fois les traitements et annotations manuelles réalisés, nous obtenons nos résultats qui nous permettrons de conclure sur la probabilité de reprise d'un composant d'une expression polylexicale.

7.1 Résultats globaux

Comme le montre le tableau ci-dessous, la reprise d'un composant d'une expression polylexicale dans une chaîne de coréférence est un phénomène rare (voir Annexe A pour les résultats détaillés), cependant cette observation est à nuancer selon le type de l'expression. Les résultats de la colonne *Reprises coréférentielles* du tableau 7.1 correspondent aux expressions pour lesquelles il a été verifié manuellement qu'un de leur composant était effectivement repris dans une chaîne de coréférence. Sur les 7 632 expressions de nos corpus, seules 276 ont donné lieu à une reprise coréférentielle, soit 3,6%.

TYPE	Reprises coréférentielles	Nombre total d'expressions	Pourcentage
VID	29	5 266	0,6%
LVC.full	245	1 726	14,2%
LVC.cause	1	18	5,6%
MVC	0	41	0%
IRV	1	996	0,1%
	276	**8 047**	**3,4%**

TABLE 7.1 – Résultats cumulés de tous les corpus

Parmi les quatre types annotés (LVC, VID, IRV et MVC), le nombre le plus important de chaînes de coréférences a été observé avec les LVC. Il n'y a aucune reprise d'un composant de MVC car elles sont seulement composées de verbes, et ne contiennent donc pas de mentions. Notre outil n'étant pas entraîné pour reconnaître les pronoms possessifs des IRV comme des mentions, on observe un seul cas de reprise pour ce type d'expression :

« *Lorsque vous êtes à l'hôpital : prévenez immédiatement votre médecin ou votre infirmière. Après votre sortie de l'hôpital : dirigez vous immédiatement au service des urgences de l'hôpital le plus proche* » (IRV - EMEA)

La première mention *vous* est reprise par la deuxième mention *vous*, qui fait partie de l'IRV *dirigez vous*.

Pour les LVC, on a en tout 246 chaînes qui correspondent à de vraies reprises. C'est un phénomène plutôt rare, mais ces observations nous indiquent que le composant reste accessible à la reprise. Les chaînes correctes ont principalement été observées sur des LVC.full, les LVC.cause étant plus rares.

« *Le 11 février 2004, l'ordonnance de renvoi devant le tribunal de 47 prévenus a été signé par le juge Armand Riberolles. Dans son ordonnance , Jacques Chirac est évoqué a plusieurs reprises, mais le juge ne peut pas poursuivre le président de la république qui est protégé par son immunité.* » (LVC.full - FRWIKI)

« *bah parce que celui qui reçoit une lettre qui est remplie de fautes d' orthographe ça donne mauvaise impression*

mauvaise impression oui hein hein et vous trouvez que c' est juste » (LVC.cause - ESLO_ANCOR)

Dans la première phrase, la mention *l'ordonnance de renvoi devant le tribunal* qui fait partie du LVC.full *ordonnance signé*, est reprise par la mention *son ordonnance*. Dans la seconde, la première mention *mauvaise impression* qui fait partie du LVC.cause *donne mauvaise impression*, est reprise par la deuxième mention *mauvaise impression*.

Concernant les VID, très peu de chaînes extraites correspondent à de vraies reprises : soit l'expression polylexicale est reconnue à tort, soit les chaînes sont mal construites. Et même en privilégiant le rappel pour s'assurer de ne pas louper un résultat intéressant, nous n'avons trouvé que très peu de cas de coréférence corrects parmi les nombreux VID annotés dans nos corpus. De plus, la très grande majorité (28/29) se trouvent dans le corpus ANCOR et le seul cas annoté comme *vrai* en dehors du corpus ANCOR est discutable (voir 7.2). Le composant de l'expression ne semble donc pas accessible pour ce type.

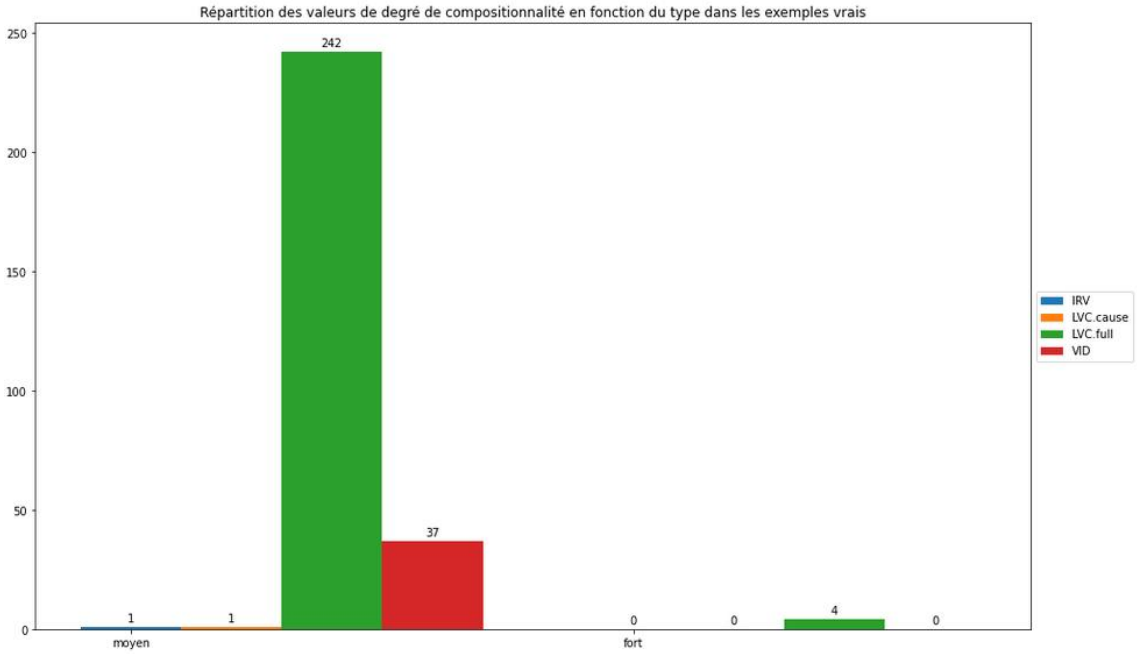
« *c'est pas les mêmes problèmes qui se posent à chacun* » (VID - ESLO_ANCOR)

La mention *les mêmes problèmes* qui fait partie du VID *problèmes posent* est reprise par la mention *qui*.

Nous avons annoté en compositionnalité les expressions polylexicales pour lesquelles un de leur composant a été retrouvé dans une chaîne de coréférence et 25 autres expressions dont les composants n'ont jamais été retrouvés dans des chaînes de coréférence au cours de nos expériences. La plupart ont été annotées comme ayant un degré de compositionnalité moyen.

Parmi les expressions polylexicales qui ont un de leur composant dans une chaîne de coréférence, 98,6% ont été annotées comme ayant un degré de compositionnalité moyen et 1,4% comme fort (voir figure 7.1). Aucune de ces expressions n'a un degré de compositionnalité faible. Seules quatre LVC.full ont été annotées comme fortement

FIGURE 7.1 – Graphique exposant la répartition des types d'expressions en fonction du degré de compositionnalité dans les exemples présentant une reprise coréférentielle



compositionnelles (*commettre crime* et trois occurrences d'*avoir idée*). Parmi les 25 expressions polylexicales dont les composants n'ont jamais été retrouvés dans des chaînes de coréférence, la plupart ont été annotées comme moyennement compositionnelles et trois comme faibles.

Bien qu'il soit difficile d'établir un test de compositionnalité fiable, nous pouvons néanmoins en conclure que la plupart des expressions polylexicales sont moyennement compositionnelles et que les composants de celles ayant un fort taux de compositionnalité ont plus de chance d'apparaitre dans des chaînes de coréférence. De plus, les composants des expressions ayant un faible taux de compositionnalité ne semblent pas pouvoir être repris.

7.2 Chaînes discutables

Parmi les résultats récupérés, certaines reprises potentielles de composants d'expressions polylexicales sont difficiles à annoter, car interprétables différemment selon les annotateurs. Certaines de ces chaînes ont été annotées comme *vrai* ou *faux* suite aux discussions et d'autres comme *discutables* lorsque aucun des deux points de vues ne semblait plus logique que l'autre.

« *Traitement de l'ostéoporose post-ménopausique l'ostéoporose masculine chez les patients à risque élevé de fractures, notamment chez les patients ayant eu une fracture de hanche récente secondaire à un traumatisme modéré.* » (EMEA)

La mention *de fractures* est reprise par *une fracture de hanche récente* qui fait partie de l'expression *eu fracture*. *de fracture* est donc une mention plus générale et abstraite que *une fracture de hanche récente*.

« *Créé par la Fédération nationale qui perpétue le souvenir de l'homme d'État meusien qui fut ministre de la Guerre et l'initiateur d'un système de défense qui porte son nom , le prix André-Maginot récompense des travaux liés au civisme et au devoir de mémoire.* » (FRWIKI)

Dans cette phrase, la mention *son nom* qui fait partie de l'expression *porte nom* est repris par *André-Maginot*. Cependant, *son nom* fait référence au nom de l'homme alors que la mention *André-Maginot* fait référence au nom du prix. Il est donc difficile de trancher sur la relation entre *son nom* et *André-Maginot*.

« *mais on on peut faire des bons manuels moi vous vous me direz que j' en reviens toujours à cette question mais enfin y a des bons métiers manuels et on euh un élément moyen peut faire un un bon manuel tout de même et il gagne sa vie* » (CO2_ANCOR)

La mention *en* de l'expression *en revenir* est annotée dans le corpus ANCOR comme faisant référence à la mention *cette question*. On peut cependant se demander si *en* fait vraiment référence à une quelconque entité que ce soit.

« *et vous avez le temps d' aller alors à au concerts ou au musée ou au théâtre*

j' ai pas le temps mais sinon j' irais bien » (ESLO_ANCOR)

Le corpus ANCOR étant de l'oral transcrit, on observe certaines spécificités du langage parlé qui peuvent favoriser la reprise de composants d'expressions. Par exemple, dans le cas ci-dessus la reprise du composant *le temps* est provoquée par le schéma de question/réponse.

Il y a également les disfluences et les répétitions d'expressions complètes qui favorisent la reprise dans ce corpus, mais ce type de chaînes de coréférence n'a pas été pris en compte dans notre étude et ces cas ont été annotés comme *répétitions* (voir 6.1).

Les exemples discutés dans cette partie concernent tous des composants de VID. La reprise de ces composants est donc rare (0,6% tous corpus confondus) et souvent discutable.

On peut en conclure que les composants d'expressions polylexicales ont peu de chance d'appartenir à des chaînes de coréférences, et encore moins si ce sont des VID.

Troisième partie

PERSPECTIVE

CHAPITRE 8

AMÉLIORATION DE LA RÉSOLUTION DE CORÉFÉRENCE

Sommaire

8.1 Méthode proposée

Dans cette partie nous proposerons une méthode qui pourrait permettre d'améliorer les résultats des systèmes de résolution de coréférence en nous appuyant sur les résultats obtenus et les observations réalisées au cours de ce mémoire.

Les résultats de nos expériences ont montré que les composants de VID ne sont que très peu accessibles à la reprise coréférentielle. En effet, parmi les VID correctement reconnus par Seen2Seen ou annotés manuellement (selon les corpus) seuls 0,6% ont un composant se trouvant dans une chaîne de coréférence. De plus, les VID sont les expressions les plus courantes dans nos corpus : environ 5 000 VID pour presque 2 000 LVC.full, environ 1 000 IRV, une cinquantaine de MVC et une vingtaine de LVC.cause.

Nous souhaitons donc proposer une méthode qui permettrait aux outils de résolution de coréférence de ne pas prendre en compte les mentions correspondant à des composants de VID. Pour illustrer cette méthode de manière simplifiée, nous allons utiliser une phrase extraite du corpus FRWIKI, que l'on considérera comme un document complet :

« *De nombreux protagonistes ont trouvé la mort depuis la signature du contrat.* » (FRWIKI)

Lorsque cette phrase est annotée par nos outils, la mention *la mort* qui fait partie de l'expression polylexicale *trouver la mort* est considérée comme coréférente avec la mention *la signature du contrat*.

« *De nombreux protagonistes ont trouvé la mort depuis la signature du contrat .* » (FRWIKI)

— **Annotation en expressions polylexicales**
« De nombreux protagonistes ont trouvé la mort depuis la signature du contrat. » (FRWIKI)
La première étape consisterait à annoter nos données en expressions polylexicales afin que le système de résolution de coréférence puisse utiliser ces informations.

— **Détection des mentions**
« [De nombreux protagonistes] ont trouvé [la mort] depuis [la signature [du contrat]] . » (FRWIKI)
Il y a quatre mentions dans ce document : *De nombreux protagonistes*, *la mort*, *la signature du contrat* et *du contrat*. Les crochets ont été ajoutés autour des mentions pour montrer les mentions imbriquées.

— **Création des paires**

[De nombreux protagonistes] [la mort] [la mort] [la signature du contrat]
[De nombreux protagonistes] [la signature du contrat] [la mort] [du contrat]
[De nombreux protagonistes] [du contrat]

[la signature du contrat] [du contrat]

Chaque mention est associée aux autres mentions du document pour former des paires. Notre document est très court pour simplifier les explications, mais dans un document plus grand le mode de pairage sélectionné pourrait changer le nombre de paires obtenues.

— **Filtrer les paires à présenter au classifieur**

- -
[De nombreux protagonistes] [la signature du contrat] -
[De nombreux protagonistes] [du contrat]

-

Les mentions incluses ne peuvent pas être coréférentes avec celles qui les englobent. La paire *[la signature du contrat] [du contrat]* est donc retirée. Nous proposons d'ajouter un filtre qui retire les paires contenant une mention correspondant à un composant d'un VID. Toutes les paires contenant *la mort* seraient alors ignorées pour la construction des chaînes de coréférence.

— **Classification pour trouver les paires coréférentes et filtrage pour ne garder qu'un antécédent par mention**

— **Construction des chaînes de coréférence**
« De nombreux protagonistes ont trouvé la mort depuis la signature du contrat. » (FRWIKI)

Il n'y a donc pas de chaîne de coréférence dans ce document. La prise en compte de l'appartenance ou non d'une mention à un VID pourrait permettre aux outils de

résolution de coréférence d'éviter certaines erreurs comme celle de lier *la mort* avec *la signature du contrat*.

8.2 Discussion

8.2.1 Avantages

Le but de notre méthode est de ne pas donner la possibilité aux systèmes de résolution de coréférence d'inclure les mentions appartenant à un VID dans des chaînes de coréférence. Dans la partie précédente, nous avons vu que cette méthode devrait nous permettre d'éviter des faux positifs.

Nous allons utiliser les annotation des sous-corpus de PARSEME afin de tenter d'évaluer les apports de cette méthode sur les performances d'un outil de résolution de coréférence. Les sous-corpus de PARSEME étant annotés manuellement en expressions polylexicales, les erreurs ne peuvent être dûes qu'à des fausses chaînes de coréférence.

CORPUS	**Nombre total de VID**	**Nombre de composants de VID dans des chaînes selon OFCORS**	**Nombre de composants de VID effectivement dans des chaînes**
FRWIKI (1000 phrases)	98	14	1
EMEA (1000 phrases)	32	2	0
AnnodisER (500 phrases)	74	18	0
TOTAL	**204**	**34**	**1**

TABLE 8.1 – Nombre de chaînes de coréférence contenant des composants de VID dans PARSEME

Ce tableau nous montre de nouveau que les reprises sont très rares, mais également que l'outil gagnerait en précision avec une perte minimale de rappel s'il ne prenait pas en compte les composants de VID pour la construction des chaînes. Les chaînes ne comprenant que deux éléments ne seraient alors pas créées, et un faux positif serait évité pour chacune d'entre elles. Par exemple :

« De nombreux protagonistes ont trouvé la mort depuis la signature du contrat . » (FRWIKI)

La chaîne reliant *la mort* à *la signature du contrat* selon nos outils, n'est composée que de deux éléments. En ignorant la mention *la mort*, la chaîne ne serait alors pas construite.

Cependant, les chaînes fausses avec plus de deux éléments perdraient simplement un de leur composant. La chaîne ensuite formée pourrait devenir vraie si tous les éléments sont désormais coréférents ou rester fausse :

« La coagulation sanguine peut poser un problème en cas de perturbation du flux sanguin. Angiox est un anticoagulant ; il empêche le sang de coaguler (formation

de caillots). » (EMEA)

Dans cet extrait, une chaîne erronée a été construite par nos outils avec les mentions : *un problème*, *un anticoagulant*, *il* et *le sang*. Même si la mention *un problème* qui fait partie d'un VID était retirée, la chaîne resterait fausse.

On ne peut donc pas affirmer que 33 faux positifs seraient évités.

CORPUS	2 éléments	>2 éléments (le reste de la chaîne est correct)	>2 éléments (le reste de la chaîne est incorrect)
FRWIKI	4	0	9
EMEA	1	0	1
AnnodisER	8	0	10
TOTAL	**13**	**0**	**20**

TABLE 8.2 – Types de chaînes existants parmi les 33 faux positifs dont une des mention fait partie d'un VID dans PARSEME

Parmi les 33 faux positifs, 13 chaînes ne sont composées que de deux mentions et 20 ont plus de deux mentions coréférentes. Aucune des chaînes de plus de deux éléments ne deviendrait correcte en retirant la mention appartenant à un VID. Avec la méthode proposée nous aurions donc évité 13 faux positifs sur les 33 recensés dans le cadre de cette analyse, et 1 faux négatif serait créé.

8.2.2 Difficultés

Cependant, comme on a pu le voir dans la première partie de ce mémoire, les outils d'identification d'expressions polylexicales ont de bonnes performances mais ne sont pas parfaits (0,66 de f-mesure pour Seen2seen, voir partie 2.2). Il y aura donc des erreurs de reconnaissance, qui influeront sur les performances de l'outil de résolution de coréférence.

Le corpus ANCOR étant annoté manuellement en chaînes de coréférences, les erreurs d'annotations ne peuvent être dûes qu'aux expressions polylexicales. Nous ne pouvons cependant analyser que les VID contenues dans des chaînes de coréférences selon nos outils, car ce sont celles qui ont été vérifiées manuellement. Cela peut donc nous permettre de constater les difficultés que pourrait rencontrer un système de résolution de coréférence qui utiliserait une annotation automatique en expressions. Notons que ces résultats ne sont pas représentatifs des performances des outils de reconnaissance d'expressions polylexicales, car si ces mentions se trouvent dans des chaînes de coréférence alors il y a de fortes chances que le VID n'ai pas été correctement reconnu.

Dans le tableau 8.3, seules les expressions annotées comme *faux* avec la valeur *Mention Incorrecte* ont été renseignées (voir 6.1). Ce tableau nous indique que bien que la méthode proposée dans la section 8.1 semble pouvoir théoriquement améliorer la résolution de coréférence, il peut y avoir des difficultés techniques et du bruit serait alors produit par plusieurs outils. En effet, sur les 578 VID reconnus dans le corpus ANCOR comme appartenant à une chaîne de coréférence, seuls 55 étaient effectivement des VID.

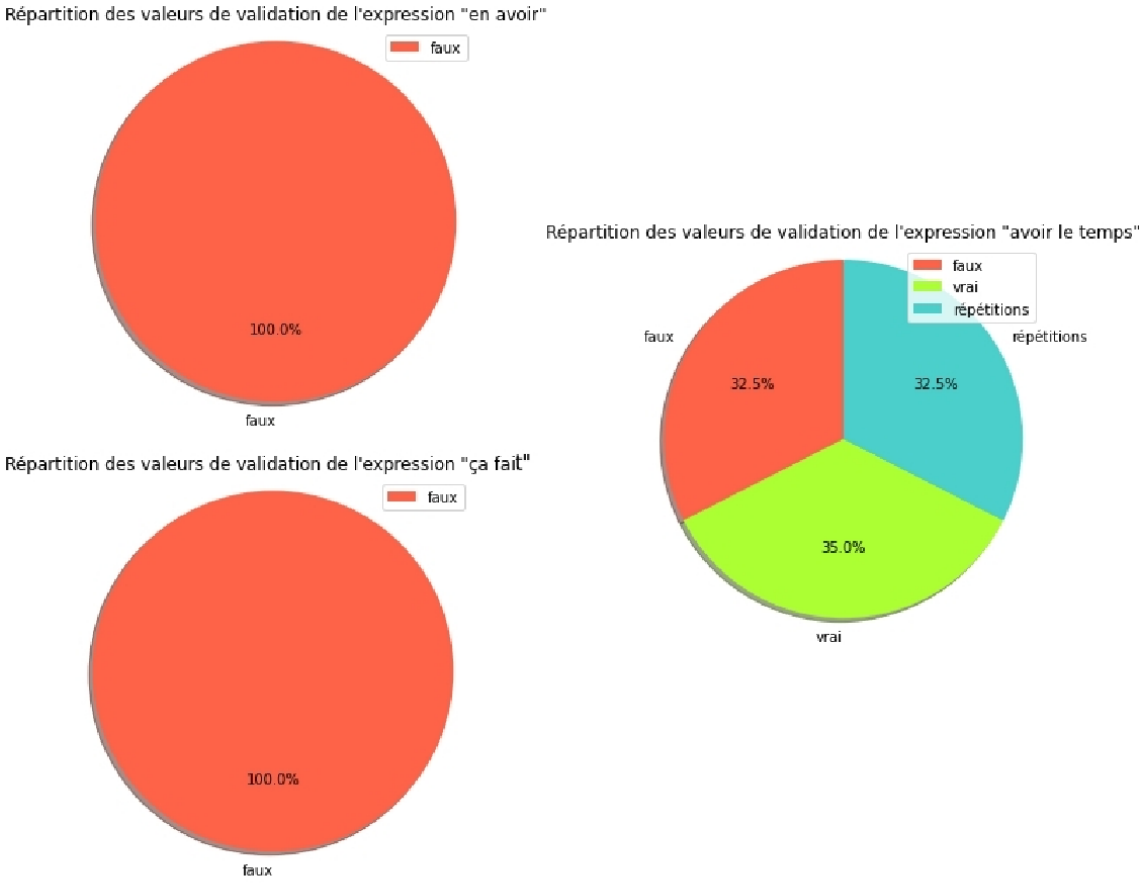
CORPUS	**Nombre TOTAL de VID reconnus**	**Nombre de VID reconnus dans une chaîne**	**Nombre réel de VID dans une chaîne**
ESLO_ANCOR (25 000 phrases)	4 007	478	47
ESLO_CO2 (2 500 phrases)	366	36	5
OTG (2 800 phrases)	338	56	3
UBS (700 phrases)	49	8	0
TOTAL	**4 760**	**578**	**55**

TABLE 8.3 – Nombre de composants de VID compris dans des chaînes correctement reconnues dans le corpus ANCOR

Les trois VID les plus courants parmi ceux analysés sont *en avoir* (279 occurrences), *ça fait* (74 occurrences) et *avoir le temps* (40 occurrences). On voit dans la figure 8.1 que *en avoir* et *ça fait* sont toujours annotés comme *faux*, ce qui signifie que lorsqu'ils sont reconnus dans des chaînes de coréférence, ils sont toujours reconnus à tort. *en avoir* et *ça fait* ont donc peu de chances d'appartenir à des chaînes de coréférence.

La méthode proposée pourrait donc permettre d'améliorer les performances d'un outil de résolution de coréférence à condition que l'annotation en expressions polylexicales soit fiable, mais les performances actuelles des outils d'identification d'expressions polylexicales ne semblent pas être suffisantes. On pourrait alors envisager d'autres méthodes comme l'ajout d'un trait indiquant si la mention correspond à un composant d'un VID dans les informations à présenter au classifieur, afin qu'il apprenne par lui même l'influence de ce trait sur les résultats.

FIGURE 8.1 – Graphiques de la répartition des valeurs d'annotation de *en avoir*, *ça fait* et *avoir le temps* dans ANCOR



Conclusion générale

Au cours de ce mémoire nous avons annoté des corpus de tailles importantes (environ 36 500 phrases en tout) et présentant des caractéristiques et des thèmes variés, en expressions polylexicales et en chaînes de coréférence. Ces expériences nous ont permis d'observer le comportement des composants des expressions polylexicales et notamment leur possible reprise dans des chaînes de coréférences. Nous pouvons donc conclure à partir des résultats obtenus, qu'il est peu probable que des composants d'expressions polylexicales verbales appartiennent à des chaînes de coréférences. Il y a cependant plusieurs critères qui viennent nuancer ce propos :

— Le comportement des composants varie selon le type d'expression polylexicale. En effet, les composants de LVC ont plus de chance d'être repris (environ 14% dans nos corpus) que ceux des VID (0.6% dans nos corpus). Nos annotations ne nous permettent cependant pas de tirer des conclusions sur les IRV, notre outil n'étant pas entraîné pour reconnaître les pronoms possessifs des IRV comme des mentions.
— L'accessibilité des composants à la reprise coréférentielle semble également varier selon le degré de compositionnalité sémantique des expressions. Les composants des expressions les plus compositionnelles étant plus susceptibles d'être repris.

Notons également que ces critères sont liés, les expressions annotées comme les plus compositionnelles étant les LVC.

Enfin, ces résultats nous permettent de dire que les annotations en expressions polylexicales pourraient aider les systèmes de résolution de coréférence à avoir de meilleures performances, mais cette méthode pourrait être difficile à mettre en place à cause du bruit produit par les annotations automatiques en expressions polylexicales. Il pourrait également être envisageable d'utiliser les annotations en chaînes de coréférence pour améliorer les systèmes de reconnaissance d'expressions polylexicales, en ignorant les VID reconnues dans des chaînes de coréférence. Afin d'éviter trop d'erreurs, on pourrait par exemple se limiter aux relations donnant les meilleurs scores (les reprises anaporiques directes).

BIBLIOGRAPHIE

[Constant et al., 2017] Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Multiword expression processing: A survey. Computational Linguistics, 43(4):837–892. – Cité pages 16 et 18.

[Cordeiro et al., 2019] Cordeiro, S., Villavicencio, A., Idiart, M., and Ramisch, C. (2019). Unsupervised compositionality prediction of nominal compounds. Computational Linguistics, 45(1):1–57. – Cité page 32.

[Désoyer et al., 2015] Désoyer, A., Landragin, F., Tellier, I., Lefeuvre, A., and Antoine, J.-Y. (2015). Les coréférences à l'oral: une expérience d'apprentissage automatique sur le corpus ancor. Traitement Automatique des Langues, 55(2):97–121. – Cité pages 12 et 14.

[Grobol, 2019] Grobol, L. (2019). Neural coreference resolution with limited lexical context and explicit mention detection for oral french. In Second Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC19). – Cité pages 14 et 23.

[Grobol, 2020] Grobol, L. (2020). Coreference resolution for spoken French. PhD thesis, Université Sorbonne Nouvelle-Paris 3. – Cité pages 11 et 15.

[Kurfali, 2020] Kurfali, M. (2020). Travis at parseme shared task 2020: How good is (m) bert at seeing the unseen? In International Conference on Computational Linguistics (COLING), Barcelona, Spain (Online), December 13, 2020, pages 136–141. – Cité page 18.

[Landragin, 2020] Landragin, F. (2020). Rapport final du projet ANR Democrat," Description et modélisation des chaînes de référence: outils pour l'annotation de corpus et le traitement automatique". PhD thesis, ANR (Agence Nationale de la Recherche-France). – Cité page 13.

[Laporte, 2018] Laporte, É. (2018). Choosing features for classifying multiword expressions. – Cité pages 16 et 19.

[Muzerelle et al., 2014] Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.-Y., Pelletier, A., Maurel, D., Eshkol, I., and Villaneau, J. (2014). Ancor_centre, a large free spoken french coreference corpus: description of the resource and reliability measures. In LREC'2014, 9th Language Resources and Evaluation Conference., pages 843–847. – Cité pages 12, 13 et 26.

[Oberle, 2019] Oberle, B. (2019). Détection automatique de chaînes de coréférence pour le français écrit: règles et ressources adaptées au repérage de phénomènes linguistiques spécifiques. In TALN-RECITAL-PFIA 2019, pages 499–512. ATALA. – Cité pages 12 et 14.

[Pasquer et al., 2020] Pasquer, C., Savary, A., Ramisch, C., and Antoine, J.-Y. (2020). Verbal multiword expression identification: Do we need a sledgehammer to crack

a nut? In Proceedings of the 28th International Conference on Computational Linguistics, pages 3333–3345. – Cité pages 18 et 23.

[Ramisch et al., 2020] Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Mititelu, V. B., Bhatia, A., Iñurrieta, U., Giouli, V., et al. (2020). Edition 1.2 of the parseme shared task on semi-supervised identification of verbal multiword expressions. In Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 107–118. – Cité page 18.

[Taslimipoor et al., 2020] Taslimipoor, S., Bahaadini, S., and Kochmar, E. (2020). Mtlb-struct@ parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. arXiv preprint arXiv:2011.02541. – Cité page 18.

ANNEXE A

TABLEAUX DE RÉSULTATS PAR CORPUS

TYPE	Reprises coréférentielles	Nombre total d'expressions	Pourcentage
VID	1	98	1%
LVC.full	2	79	2,5%
LVC.cause	0	1	0%
MVC	0	3	0%
IRV	0	45	0%
	3	**226**	**1,3%**

TABLE A.1 – Résultats - PARSEME FRWIKI

TYPE	Reprises coréférentielles	Nombre total d'expressions	Pourcentage
VID	0	32	0%
LVC.full	19	184	10,3%
LVC.cause	0	7	0%
MVC	0	0	0%
IRV	1	39	2,6%
	20	**262**	**7,6%**

TABLE A.2 – Résultats - PARSEME EMEA

TYPE	Reprises coréférentielles	Nombre total d'expressions	Pourcentage
VID	0	74	0%
LVC.full	1	77	1,3%
LVC.cause	0	1	0%
MVC	0	0	0%
IRV	0	48	0%
	1	**200**	**0,5%**

TABLE A.3 – Résultats - PARSEME AnnodisER

TYPE	Reprises coréférentielles	Nombre total d'expressions (selon Seen2seen)	Pourcentage
VID	0	302	0%
LVC.full	3	122	2,5%
LVC.cause	0	3	0%
MVC	0	2	0%
IRV	0	174	0%
	3	**603**	**0,5%**

TABLE A.4 – Résultats - Est Républicain

TYPE	Reprises coréférentielles	Nombre total d'expressions (selon Seen2seen)	Pourcentage
VID	26	4 007	0,6%
LVC.full	206	1 120	18,5%
LVC.cause	1	6	16,6%
MVC	0	33	0%
IRV	0	581	0%
	233	**5 747**	**4,1%**

TABLE A.5 – Résultats - ESLO_ANCOR

TYPE	Reprises coréférentielles	Nombre total d'expressions (selon Seen2seen)	Pourcentage
VID	2	366	0,5%
LVC.full	13	106	12,3%
LVC.cause	0	0	0%
MVC	0	1	0%
IRV	0	52	0%
	15	**525**	**2,9%**

TABLE A.6 – Résultats - ESLO_CO2

TYPE	Reprises coréférentielles	Nombre total d'expressions (selon Seen2seen)	Pourcentage
VID	0	338	0%
LVC.full	0	27	0%
LVC.cause	0	0	0%
MVC	0	2	0%
IRV	0	51	0%
	0	**418**	**0%**

TABLE A.7 – Résultats - OTG

TYPE	Reprises coréférentielles	Nombre total d'expressions (selon Seen2seen)	Pourcentage
VID	0	49	0%
LVC.full	1	11	9,1%
LVC.cause	0	0	0%
MVC	0	0	0%
IRV	0	6	0%
	1	**66**	**1,5%**

TABLE A.8 – Résultats - UBS

ANNEXE B

TABLEAUX DE RÉSULTATS PAR TYPES D'EXPRESSIONS POLYLEXICALES

TABLE B.1 – Résultats - VID

EXPRESSIONS	QUANTITÉ	EXEMPLES
avoir le temps	16	*est -ce que vous avez le temps* [1] *de faire des mots-croisés ? le temps ou la condition ? jamais bon* (ESLO_CO2)
poser problème	4	*c' est pas les mêmes problèmes qui se posent à chacun* (ESLO_ANCOR)
prendre le temps	2	*je suppose parce que voyez toutes mes soeurs elles elles sont en campagne bon bah celle qui est dans le commerce ici elle est comme moi elle a plus le temps de lire elle lit plus* *ah bon ?* *je crois que si j' étais encore en campagne peut-être que je je lirais mais quand même euh je j' aimais bien lire oh tiens maman lis- le et puis tu vas me l' expliquer je que j' avoue que j' étais un peu nerveuse pour ces points -là oh je dis j' ai rien compris ou je tu vas me me le réexpliquer hm oui enfin question surtout de s' arrêter de prendre le temps* (ESLO_ANCOR)
prendre sa place	2	*euh moi au départ c' était un travail de formation parce que je me rendais compte que la femme a une place à prendre dans la vie civique et sociale et au fond euh on n' est pas du tout préparé euh ni par les études euh si vous voulez jusqu' à un niveau de baccalauréat des études ordinaires on n' est pas du tout préparé à prendre notre place* (ESLO_ANCOR)

1. Les expressions polylexicales sont soulignées et les mentions sont surlignées en vert.

TABLE B.1 – Résultats - VID

EXPRESSIONS	QUANTITÉ	EXEMPLES
il est question	1	*à seize ans mais il y a des enfants qui n' aiment pas les études et à partir de ce moment -là il n' est pas question d' âge qu' on les fasse aller jusqu' à seize ans et qu' après on leur fasse faire autre chose puisque mais enfin je sais pas si c' est une question d' âge joue tellement enfin je sais que j' aurais des enfants personnellement je ferais n' importe quoi pour qu' ils fassent des études euh le plus longtemps possible à la condition que ça les intéresse s' ils ont envie de faire de la danse qu' ils fassent de la danse s' ils sont heureux en faisant de la danse.* (ESLO_ANCOR)
porter nom	1	*Créé par la Fédération nationale qui perpétue le souvenir de l'homme d'État meusien qui fut ministre de la Guerre et l'initiateur d'un système de défense qui porte son nom , le prix André-Maginot récompense des travaux liés au civisme et au devoir de mémoire.* (FRWIKI)
en revenir	1	*mais on on peut faire des bons manuels moi vous vous me direz que j' en reviens toujours à cette question mais enfin y a des bons métiers manuels et on euh un élément moyen peut faire un un bon manuel tout de même et il gagne sa vie* (ESLO_CO2)
faire plaisir	1	*oh j' ai les loisirs vous savez pour moi j' ai vu la télévision ça me fait bien plaisir ah autrefois j' aimais aller au théâtre* *[. . .]* *on n' entend plus des ténors on n' entend plus de barytons chanter on n' entend plus rien ce n' est plus de belles voix comme autrefois oh j' ai vu des pièces magnifiques alors là après la guerre de quatorze quand mon père était rentré en dix-neuf là j' ai pris du plaisir j' avais vingt euh vingt ans* (ESLO_ANCOR)
en savoir	1	*vos vos grands p- vos grands-parents euh euh qu' est -ce qu' ils avaient comme euh comme diplôme ? est -ce qu' ils sont toujours ils sont toujours en vie vos grands-parents ? non* *non ? qu' est -ce qu' ils avaient comme* *oh ça j' en sais rien* (ESLO_ANCOR)

TABLE B.2 – Résultats - IRV

EXPRESSIONS	QUANTITÉ	EXEMPLES
se diriger	1	*Lorsque vous êtes à l'hôpital : prévenez immédiatement votre médecin ou votre infirmière.*

		Après votre sortie de l'hôpital : dirigez vous immédiatement au service des urgences de l'hôpital le plus proche (EMEA)

TABLE B.3 – Résultats - LVC.cause

EXPRESSIONS	**QUANTITÉ**	**EXEMPLES**
donner impression	1	*bah parce que celui qui reçoit une lettre qui est remplie de fautes d' orthographe ça donne mauvaise impression mauvaise impression oui hein hein et vous trouvez que c' est juste* (ESLO_ANCOR)

TABLE B.4 – Résultats - LVC.full

EXPRESSIONS	QUANTITÉ		EXPRESSIONS	QUANTITÉ
faire étude	50		avoir religion	1
poser question	24		avoir relation	1
faire grêve	19		avoir rendement	1
prendre sanction	13		avoir responsabilité	1
avoir difficulté	12		avoir rôle	1
avoir problème	6		avoir vocation	1
avoir contact	5		commettre crime	1
avoir habitude	4		comporter risque	1
avoir question	4		dispenser enseignement	1
avoir rapport	4		donner concert	1
faire essai	4		donner conseil	1
passer vacances	4		donner cours	1
avoir fracture	3		donner ordre	1
avoir idée	3		entreprendre action	1
faire confiance	3		exercer activité	1
faire travail	3		faire demande	1
avoir activité	2		faire effort	1
avoir besoin	2		faire fête	1
avoir conséquence	2		faire guerre	1
avoir importance	2		faire recherche	1
avoir impression	2		faire service	1
avoir opinion	2		garder souvenir	1
avoir projet	2		mener combat	1
donner enseignement	2		prendre cours	1
donner réponse	2		prendre position	1
exercer contrôle	2		produire résultat	1
faire classe	2		présenter saignements	1
faire course	2		présenter symptômes	1
insuffisance atteint	2		recevoir perfusion	1
mener action	2		recevoir éducation	1
mener étude	2		réaliser étude	1
prendre décision	2		SCA atteint	1
prendre photo	2		signer ordonnance	1
subir traitement	2		souffrir de maladie	1
atteindre maladie	1		souffrir de syndrôme	1
avoir capacité	1		subir angioplastie	1
avoir connaissance	1		subir pontage	1
avoir formation	1		suivre cours	1
avoir influence	1		travail accomplir	1
avoir intention	1		avoir perception	1
avoir intérêt	1		avoir possibilité	1