Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

Étude comparative des expressions polylexicales verbales en français et en chinois : éléments linguistiques, statistiques et TAL

Master

Traitement Automatique des Langues

Parcours :

Ingénierie Multilingue

par

Jianying Liu

Directeur de mémoire :

Damien Nouvel

Encadrant :

Agata Savary
Jean-Yves Antoine
Anaïs Lefeuvre-Halftermeyer

Année universitaire 2020/2021

Remerciements

En préambule de ce mémoire, je souhaite adresser ici tous mes remerciements aux personnes qui m'ont apporté leur aide et qui ont ainsi contribué à l'élaboration de ce mémoire.

Tout d'abord, je voudrais remercier mon directeur de mémoire, Monsieur Damien Nouvel, pour son implication dans mes recherches, son aide lors de l'élaboration du plan de mémoire, sa relecture et ses conseils détaillés sur la langue française dans ce travail.

Je tiens à remercier particulièrement mes tuteur.rices Mme Agata Savary, Monsieur Jean-Yves Antoine et Mme Anaïs Lefeuvre-Halftermeyer, pour leur précieuse aide à la relecture de mon mémoire, ainsi que leurs encadrements pendant le stage.

Je souhaite exprimer ma gratitude envers Anaëlle Pierredon, avec qui j'ai passé un stage plaisant.

J'adresse mes plus sincères remerciements à mes parents et mes amies WANG Qi et XU Yiqing, qui m'ont toujours soutenue et encouragée au cours de la réalisation de ce mémoire.

Enfin, je n'oserais oublier de remercier tout le corps professoral du pluriTAL, pour leur formation de haute qualité et leurs aides au cours de mes études.

Merci à toutes et à tous.

Résumé

Ce travail porte sur une étude comparative des expressions polylexicales verbales en français et en chinois. Nous essayons d'abord de valider l'hypothèse sur la non-compositionnalité des expressions polylexicales et les chaînes de coréférence, selon laquelle les composants d'une expression polylexicale verbale sont peu susceptibles d'être repris dans une chaîne de coréférence. Ensuite, en se basant sur les fautes observées pendant l'annotation manuelle, les analyses statistiques des traits et nos connaissances sur la langue chinoise, nous supposons que l'hypothèse se comporte de manière similaire en chinois, et proposons des points à affiner dans l'hypothèse, ainsi que des pistes possibles pour améliorer l'identification automatique des expressions polylexicales verbales en français et en chinois.

Mots clés : expression polylexicale verbale, non-compositionnalité, coréférence, français, chinois

Table des matières

Liste des figures

Liste des tableaux

Introduction

Présentation générale

Les expressions polylexicales (multiword expression en anglais, ci-après EP) existent dans toutes les langues, et leur traitement reste un défi important pour le traitement automatique des langues. Parmi toutes leurs caractéristiques, la non-compositionnalité sémantique désigne le fait qu'il est impossible de déduire le sens global d'une expression polylexicale à partir du sens de leurs composants. Par ailleurs, les chaînes de coréférence, qui relèvent d'un phénomène linguistique courant, regroupent des mentions lorsqu'elles désignent la même entité dans le monde du discours. Ce chaînage demande donc une homogénéité sémantique entre les mentions, mais la différence entre la signification globale et le sens interne d'une expression polylexicale est une difficulté qui empêche des liaisons entre les substantifs intérieurs et extérieurs. De ce fait, ce travail examine l'hypothèse selon laquelle les composants nominaux des expressions polylexicales verbales sont des éléments qui se prêtent peu à la construction des chaînes de coréférence.

Dans ce mémoire, nous proposons une vérification quantitative de cette hypothèse. Nous utilisons d'abord deux logiciels existant – Seen2Seen et OFCORS – pour détecter les expressions polylexicales et les chaînes de coréférence. Après leur annotation automatique, nous fusionnons les résultats et vérifions manuellement les annotations avant d'effectuer une analyse quantitative et qualitative sur les croisements détectés. Nos expérimentations ont également permis de tester la fiabilité des outils et d'éprouver l'hypothèse sur des cas réels rencontrés en corpus.

En plus de cette vérification de l'hypothèse dans la langue française, nous élargissons le champ de notre étude à une langue éloignée du français, la langue chinoise. Nous résumons les caractéristiques principales du chinois en nous basant sur notre connaissance de cette langue, et illustrons les différences des expressions polylexicales verbales en chinois par rapport au français en les comparant statistiquement. En considérant les spécificités en chinois, nous discutons de difficultés spécifiques pour l'identification des expressions polylexicales en chinois, ainsi que les résultats obtenus pour la même hypothèse.

Finalement, à partir de notre constat pendant l'expérimentation, nous proposons certaines pistes afin d'améliorer l'outil et poursuivre l'étude sur le degré de compositionnalité des expressions polylexicales verbales.

Plan de lecture

Ce mémoire s'organise de manière suivante. Dans la première partie, nous nous concentrons sur l'aspect linguistique. Nous commençons par une présentation générale des expressions polylexicales dans le premier chapitre, en donnant la définition, la typologie et les caractéristiques. Le deuxième chapitre se concentre sur les spécificités de la langue chinoise et de ses expressions polylexicales, nous discutons également des impacts sur le traitement des expressions polylexicales en nous basant sur les traits relevés par la statis-

tique. Ensuite, dans le troisième chapitre, nous introduisons précisément notre hypothèse avec des exemples.

La deuxième partie explique notre vérification de l'hypothèse et les résultats obtenus. Le quatrième chapitre présente les outils utilisés ainsi que les deux domaines étudiés en TAL : la résolution de coréférence et surtout l'identification des expressions polylexicales. Nous continuons en précisant les corpus utilisés et la méthodologie suivie dans le cinquième chapitre. Le sixième chapitre explique notre méthode pour classifier et annoter les croisements détectés, nous finissons ce chapitre par l'analyse statistique des résultats d'annotation. Cette analyse est approfondie par une analyse du degré de compositionnalité des expressions polylexicales verbales dans le septième chapitre.

Nous finissons le mémoire en discutant les perspectives dans le huitième chapitre, telles que les points à affiner dans l'hypothèse, les pistes possibles pour poursuivre l'étude et des modifications possibles du système Seen2Seen pour améliorer sa performance en français et en chinois.

Première partie

Étude linguistique des expressions polylexicales

CHAPITRE 1

Présentation générale des expressions polylexicales

Sommaire

Dans ce chapitre, nous présentons l'état de l'art sur les expressions polylexicales d'un point de vue linguistique. Nous illustrons au fur et à mesure les attributs qui les qualifient comme expression polylexicale ainsi qu'une typologie générale des expressions polylexicales et une classification plus spécifique sur les expressions polylexicales verbales. Nous finissons ce chapitre par une synthèse des caractéristiques qui engendrent des difficultés ou des opportunités pour les tâches de TAL concernées.

1.1 Définition selon les attributs et enjeux

Il est possible d'analyser les langues humaines comme des constructions à l'aide de briques atomiques en respectant des principes de cohérence et de solidité de cette construction. On considère en général que les phrases sont formées à partir d'unités sémantiques plus petites, que l'on sépare en utilisant la notion de mot et en suivant des règles grammaticales. Néanmoins, les « briques langagières » ne sont pas toujours faciles à interpréter et à combiner. La segmentation est déjà une problématique importante. La plupart du temps, les mots sont délimités par des espaces ; pourtant, il existe un cas spécial mais incontournable dans toutes les langues : les expressions polylexicales. Dans ce cas, l'utilisation conjointe de plusieurs mots exprime une signification spécifique qui est différente voire éloignée de la simple combinaison des mots qui la composent. Ou, pour le cas de collocations, leur cooccurrence est habituellement figée donc la substitution d'un composant par son synonyme sera conventionnellement incorrect.

En fait, le concept d'expression polylexicale est tellement complexe qu'il est difficile d'en donner une définition simple. [Moon, 1998] a même dit que ce n'est pas un phénomène unifié à décrire, mais un ensemble complexe des traits qui interagissent de manière diverse (souvent en désordre) et représentent un continuum entre des groupes lexicaux idiomatiques et des groupes lexicaux compositionnels. Néanmoins, beaucoup de chercheurs ont quand même essayé de décrire EP à partir de ses traits : [Calzolari et al., 2002] le définissent comme une séquence de mots qui se comporte comme une seule unité pour l'analyse

linguistique ; [Carpuat and Diab, 2010] décrivent les expressions polylexicales comme les unités polylexicales et les collocations des mots qui se caractérisent par leur cooccurrence plus fréquente ; [Sag et al., 2002] mettent l'accent sur son interprétation sémantique particulière qui dépasse les frontières des mots.

[Baldwin and Kim, 2010] ont essayé de conclure ces traits en classifiant comme expression polylexicale les éléments lexicaux qui 1) peuvent être décomposés en multiple lexèmes et 2) présentent des qualités idiomatiques du point de vue lexical, syntaxique, sémantique, pragmatique et statistique. À noter que cette définition de multi-lexèmes est également applicable pour la langue chinoise, qui n'est pas segmentée par les espaces. S'agissant des idiomatismes, plus précisément, l'idiomatisme lexical se produit quand un ou plusieurs composants de l'expression ne font pas partie du lexique conventionnel de la langue parlée [1], autrement dit ces composants ne sont pas utilisables tout seul dans la langue parlée, nous remarquons souvent cet idiomatisme dans les traces latines comme ***ad hoc*** [2] ; la spécificité syntaxique est manifestée par les anomalies syntaxiques des composants, telles que « **drôle de question** » et (en) ***by and large*** 'généralement', deux structures syntaxiquement fausses ; l'idiomatisme sémantique, autrement-dit la non-compositionnalité, signifie le sens global non prédictible à partir de ses composants, comme (en) ***kick the bucket*** 'casser sa pipe' ou « **avoir beau** » ; l'idiomatisme pragmatique représente le fait que l'expression est liée à un contexte particulier ou à un ensemble des situations fixées, par exemple, « **à table** » pour un contexte de repas, « **bon courage** » afin de donner un souhait, ou (en) ***good morning*** 'bonjour' dans le cas de la salutation ; et finalement, l'idiomatisme statistique révèle : 1) la forte fréquence d'une combinaison des mots par rapport à ses composants ou à des expressions alternatives, comme (en) ***good morning*** 'bonjour', ou « ***ad hominem*** », ou même le fait que certaines combinaisons apparaissent probablement plus que d'autres ((en) « immaculate house » plus fréquente que « immaculate logic ») ; 2) la forte fréquence d'un ordre fixé des binômes, par exemple, la combinaison inversée de l'expression « **noir et blanc** » ne soit plus correcte conventionnellement, mais cet ordre pourrait être différent dans d'autres langues.

À part ces idiomatismes, dont la présence d'au moins un parmi eux est obligatoire pour caractériser un EP, [Baldwin and Kim, 2010] présentent d'autres attributs particuliers d'EP, qui, au contraire, ne sont ni nécessaires ni évidents. Premièrement, certaines expressions utilisent les figures de styles telles que la métaphore (« **prendre le taureau par les cornes** »), la métonymie (« verser des torrents de larmes ») ou l'hyperbole (« **pleuvoir comme vache qui pisse** » et (en) ***not worth the paper it's printed on*** (lit. 'ne vaut pas le papier sur lequel il est écrit') . Ces méthodes enrichissent le sens littéral de l'expression. Cet indicateur est fortement lié à la non-compositionnalité présentée avant. À part cela, certaines EP sont une paraphrase d'un mot simple, tel que l'expression « **leave out** » pour le mot « omit » en anglais. Finalement, certaines EP ont une prosodie particulière, telle que l'accentuation du composant plus important dans l'expression : dans (en) ***soft spot*** 'point faible', le mot « *soft* » est accenté. Cet aspect prosodique est fortement présent dans les idiomes chinois, mais d'une autre manière. Du fait que la langue chinoise est une langue tonale et monosyllabique, c'est-à-dire chaque sinogramme a une syllabe unique, pour une beauté prosodique, les poèmes les plus anciens sont composés des vers de 4 caractères qui fixent un ton précis pour certaines positions. Cette tradition dans le chinois classique est conservée dans le chinois moderne par une grande partie des idiotismes chinois – *chengyu*, qui se compose de 4 caractères figés sous une forme souvent

1. Dans l'article, il se concentre sur la langue anglaise, nous l'élargissons ici à toutes les langues éventuellement parlées.

2. Voir l'annexe E.1 et E.2 pour les notations utilisées dans les exemples

concise pour exprimer un sens plus compliqué et enrichi. Par exemple, (zh) 一 箭 双 雕 (lit. 'd'une flèche deux aigles') 'd'une pierre deux coups'.

Les caractéristiques des EP décrites ci-dessus conduisent à plusieurs problèmes en TAL, surtout pour les tâches au niveau sémantique. Puisque les EP remettent en question les frontières de mots, les tokens délimités par les espaces ne sont plus fortement des mots indépendamment présentés dans le lexique, en y ajoutant l'anomalie syntaxique, l'étiquetage des catégories morpho-syntaxiques est moins facile ; de plus, la variation de son niveau compositionnel en fonction de l'expression précise engendre des difficultés dans le regroupement des mots par le sens et dans l'analyse des relations, l'analyse syntaxique d'un texte rencontra de nombreux cas discutables ; finalement, trouver la bonne interprétation parmi les deux possibilités, soit idiomatique ou compositionnelle, est un des enjeux principaux de la désambiguïsation sémantique. La non-compositionnalité et l'ambiguïté des EP pénalisent la bonne compréhension de sa signification, qui est également une source importante des erreurs de la traduction automatique de nos jours.

1.2 Typologie des EP

1.2.1 Typologie générale des EP

[Constant et al., 2017] proposent une typologie non-exhaustive des expressions polylexicales en général :

- **Idiome** (en anglais *idiom*) : un groupe de lexèmes dont la signification provient de la convention et qui n'est pas déductible de celles de ses lexèmes. Ce sont les expressions polylexicales les moins compositionnelles. Par exemple, si on dit « quelqu'un est **une langue de vipère** », cela signifie en fait que cette personne ne cesse de tenir des propos méchants.
- **Construction à verbe support** (en anglais *light-verb construction*) : elle se compose d'une tête verbale contenant peu de signification, mais accompagnée d'un dépendant prédicatif nominal qui spécifie son sens sémantique, tel que « **poser question** » en français, et (zh) 提供 服务 'offrir service' en chinois.
- **Verbe à particule** (en anglais *verb-particle construction* ou *phrasal verb*) : se compose d'un verbe et une particule (qui est souvent une préposition ou un adverbe), dont le dernier modifie la signification du verbe et ne se trouve pas forcément à côté du verbe, par exemple (en) ***give up*** 'renoncer' et (zh) 提 出 'proposer'. À noter que cette catégorie n'existe pas pour le français.
- **Mot composé** (en anglais *compound*) : c'est une catégorie hétérogène qui demande que les composants soient juxtaposés, et ses composants peuvent avoir des flexions morphologiques. À noter que ces composants peuvent être nominaux ou verbaux ou un mélange de ces deux, et la présence d'espace ou de tiret n'est pas obligatoire. De plus, le mot composé est un phénomène linguistique plus général, donc seulement ceux qui présentent un idiomatisme statistique (« **carte de crédit** ») ou sémantique (« **carte verte** ») sont des EP.
- **Mot composé fonctionnel** (en anglais *complex function word*) : c'est le mot fonctionnel ayant plus d'un lexème, y compris des conjonctions (« **bien que** »), des prépositions/locutions prépositives (« **à force de** », « **à l'aide de** ») ou des locutions adverbiales (« **tout de suite** », « **au fur et à mesure** »).
- **Entité nommée polylexicale** (en anglais *multiword named entity*) : une expression polylexicale qui désigne strictement une entité dans le monde, qui peut être une personne (*Charles de Gaulle*), une organisation (*Académie française*) ou une

localisation (*Les Pays-Bas*). En fait, de très nombreuses entités nommées sont polylexicales.

— **Terme complexe** (en anglais *multiword term*) : c'est une sous catégorie du mot composé, qui contient des désignations d'un concept général dans un domaine spécifique, tel que « tibia péroné » en médecine.

S'agissant de leur compositionnalité, sauf les idiomes qui sont plus idiomatiques et les constructions à verbe support qui sont plus compositionnelles, toutes les autres catégories ont des degrés très variés en fonction de leurs particularités.

1.2.2 Typologie des expressions polylexicales verbales

Parmi toutes ces catégories, nous voyons clairement que les expressions polylexicales verbales sont les plus difficiles à définir et à retrouver à cause de leur variation morphologique et syntaxique, qui est aussi la plus grande difficulté à résoudre pour les EP. De ce fait, notre tâche se concentre sur les expressions polylexicales verbales (en anglais *verbal MWE*, ci-après **EPV**). Il faut clarifier certaines notions d'abord pour mieux les caractériser et les détecter par la suite.

Le programme de recherche PARSEME comporte une tâche centrale qui se concentre sur l'identification des EPV dans les textes courants. Un corpus multilingue est établi dans le cadre de ce programme (mentionné ci-après le corpus PARSEME). Son guide d'annotation [3] essaie de diviser les EPV en 5 grandes catégories [4] de manière claire et expérimentale en utilisant les arbres de décisions. Les 5 catégories, soit LVC, VID, IRV, VPC, MVC, ont été définies pour être utilisées dans des tâches multilingues. Parmi eux, VPC n'est pas présent en français et IRV n'existe pas en chinois. Nous présentons ci-dessous rapidement ces acronymes.

— **LVC** (*light-verb construction* en anglais) : cette catégorie déjà présentée plus haut, est subdivisée en 2 sous-catégories : **LVC.full** et **LVC.cause**. Dans le premier cas, la sémantique du verbe est perdue. Dans le second cas, le verbe ajoute quand même un aspect causal au nom. Voici deux exemples de LVC.full [5] :

(1) Ils **ont** la **capacité** de chélater plusieurs minéraux du sol et de favoriser leur absorption par les plantes. (fr)

(2) 他 在 会议 上 对 事故 情况 做
tā zài huìyì shàng duì shìgù qíngkuàng **zuò**
il pendant conférence sur vers accident situation faire
了 说明。
le **shuōmíng**
(part.achèvement) présentation (zh)
lit. 'Il a fait un présentation sur l'actualité de l'accident lors de la conférence' | 'Il présente l'actualité de l'accident lors de la conférence.'

— **VID** (*verbal idiom* en anglais) : une expression de cette catégorie a au moins 2 composants lexicalisés qui se composent d'une tête verbale et d'au moins un de ses dépendants. Le(s) dépendant(s) peut(peuvent) être le sujet, l'objet direct ou même le complément adverbial. Par exemple :

(3) Si vous travaillez dans le quartier et que vous **en avez marre** de manger à la cantine de l'entreprise, allez vous restaurer aux Jardins d'Epicure. (fr)

3. Voir https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=home
4. Nous ignorons ici leur catégorie expérimentale IAV (inherently adpositional verbs).
5. Voir l'annexe E.3 pour l'explication des abréviations utilisées dans la glose des exemples chinois.

(4) 他 吃 了 个 闭 门 羹。 (zh)
tā **chī** le ge **bì** **mén** **gēng**
il manger (part.achèvement) un fermer porte soupe
lit. 'Il a mangé un soupe de porte fermée.' | 'Quelqu'un lui claque la porte au nez.'

— **IRV** (*inherently reflexive verb* en anglais) : ce sont les verbes intrinsèquement réflexifs, d'un point de vue sémantique. En français, cette catégorie ne concerne que les verbes pronominaux. Néanmoins, tous les verbes réflexifs en français n'entrent pas directement dans cette catégorie, elle prend en compte seulement ceux qui n'existent que sous cette forme tel que « s'évanouir » ou ceux qui expriment une signification différente de la version non réflexive tel que « s'apercevoir » par rapport à « apercevoir ». Un contre-exemple est « se traduire » dans « Le mot se traduit comme… ». Ici, le verbe promonimal n'est pas IRV.
— **VPC** (*verb-particle construction / phrasal verb / phrasal-prepositional verb* en anglais) : comme ce qui est présenté avant, l'expression dans cette catégorie contient une tête verbale lexicalisée et une particule lexicalisée comme son dépendant. Nous subdivisions cette catégorie en 2 parties selon le niveau de compositionnalité de leur signification : **VPC.full** pour les expressions totalement non-compositionnelles, autrement dit le nouveau sens est radicalement différent de la combinaison des sens du verbe et celui de la particule ; et **VPC.semi** pour les semi-non-compositionnelles, dont la préposition ajoute un sens non-spatial mais partiellement prédictible. Il n'existe que VPC.semi en chinois. Voici un VPC.full en anglais et un VPC.semi en chinois :

(5) The airplane has **token off**. (en)
Le avion a pris absent
lit. 'L'avion a pris absent.' | 'L'avion a décollé.'

(6) 他 提 出 不同 意见。 (zh)
tā **tí** **chū** bùtóng yìjiàn
il lever sortir différent opinion
'Il soulève de différentes opinions.'

— **MVC** (*multi-verb construction* en anglais) : L'expression de cette catégorie se compose d'une séquence de 2 verbes adjacents, dont l'un fonctionne comme verbe gouverneur (v-gov) et l'autre comme verbe dépendant (v-dep). Voici certaines de leurs caractéristiques : 1) ils ont souvent le même sujet ; 2) ils désignent des actions fortement connectées qui pourraient être considérés même comme des parties d'un même événement ; 3) ils fonctionnent ensemble comme un seul prédicat ; 4) ils peuvent être idiomatiques ou indiquent la succession des événements ; 5) le v-gov est sémantiquement délexicalisé et le v-dep constitue l'essentiel du sens global de l'expression, donc leurs positions pourraient être inversées dans l'arbre de dépendance syntaxique. À noter que les MVC varient fortement en fonction des langues, donc les arbres de décisions pour les annoter sont établis spécifiquement pour différentes langues. Voici deux exemples en français et en chinois :

(7) Il a **laissé tomber** la partie la plus difficile. (fr)

(8) 他 变 成 一只 青蛙。 (zh)
tā **biàn** **chéng** yìzhī qīngwā
il changer en(devenir,v.prep.) une grenouille
lit. 'Il devient grenouille.' | 'Il se transforme en une grenouille.'

(Remarque : en fait, le chinois ne distingue pas « devenir » et « se transformer en », on utilise toujours ce même terme 变 成.)

Étant donné les variations syntaxiques riches des EPV et le besoin de l'analyse structurelle de l'expression pour l'évaluation, il nous faut définir une forme standard pour appliquer cette analyse – la **forme canonique**. Selon le guide d'annotation de PARSEME, la forme prototypique de chaque expression est sous voix active, qui a le verbe sous forme finie et les autres composants lexicalisés sont dépendants de ce verbe ou des autres composants. Si cette forme prototypique existe, nous la considérons comme forme canonique. En revanche, pour les EPV qui n'ont pas une forme prototypique possible, par exemple elles n'utilisent que la voix passive, la forme utilisée sera sa forme canonique. Prenons l'exemple de « **prendre** une **décision** », qui est la forme prototypique, ses variantes comprennent « avoir **pris** une **décision** », « les **décisions** que nous avons **prises** », « la **décision prise** » ou même « les **décisions** se **prennent** ». Mais pour l'expression « **les carottes sont cuites** », qui signifie que « la situation est irrémédiablement compromise, sans espoir », une autre formulation comme « On a cuit les carottes » ou « les carottes ont été cuites » n'a pas le même sens. Dans ce cas-là, la première forme passive au présent devient la forme canonique.

1.3 Caractéristiques

La partie précédente présente certains attributs principaux des expressions polylexicales d'un point de vue linguistique. À l'aspect TAL, le traitement des EP peut être résumé par 5 caractéristiques principales ([Constant et al., 2017]) : la cooccurrence arbitrairement prédominante, la discontinuité, la non-compositionnalité, l'ambiguïté et la variabilité, que nous décrivons en détails plus loin et qui constituent soit des difficultés essentielles, soit des opportunités de la tâche.

Premièrement, la **collocation ou cooccurrence arbitrairement prédominante**. Cette caractéristique remarquable des EP est beaucoup utilisée dans la découverte des EP, puisqu'il est facile de réaliser par des mesures d'association statistiques. Mais pour certaines expressions, dont le sens global a changé, considérer seulement cette particularité statistique pendant le traitement pourrait conduire à une traduction mot-à-mot incorrecte dans la langue cible, puisqu'elle suppose que les éléments de l'expression dans deux langues ne soient pas trop distantes.

Deuxièmement, la **discontinuité**, c'est la possibilité d'insérer des syntagmes extérieurs entre les composants de l'expression, ou le changement de l'ordre des composants des EPV en fonction de la formulation de la phrase. Par exemple, pour l'expression (en) ***turn on*** 'allumer', les deux formulations « **turn on** the TV » et « **turn** the TV **on** » sont toutes les deux correctes. Mais certaines discontinuités ne sont pas universelles. Prenons l'exemple du verbe à particule séparable, qui est fréquent en allemand mais presque inexistant dans les langues romanes, dont la particule est un préfixe à l'infinitif de l'expression, mais ce préfixe est séparé du verbe dans certains cas. Par exemple, le verbe à particule séparable (de) ***abfahren*** 'partir' provient du verbe simple « **fahren** » (conduire). Dans l'indicatif présent nous disons « (de) *Der Zug* ***fährt*** *um 9 :45 Uhr* ***ab***. 'Le train part à 9 :45' ».

Quant à la flexibilité de l'ordre, prenons l'exemple de l'expression « **poser question** » (forme canonique), nous voulons que la machine reconnaisse également « la **question** qui se **pose** » ou « la **question posée** » dont l'ordre de deux lexèmes est inversé par rapport à sa forme canonique.

Troisièmement, la **non-compositionnalité**. Nous la mentionnons plusieurs fois ci-dessus et elle est prototypique dans les idiomes. Des exemples concrets comme l'expression nominale « **fleur bleue** » signifie une personne sentimentale ou même naïve, ou la locution verbale « **avoir lieu** » exprime « se passer ». Leur signification est toujours obscure si l'interlocuteur ne connaît que le sens de chaque mot. D'un point de vue traductique, [Melamed and Melamed, 2001] spécifient pour les termes complexes la non-compositionnalité comme le cas où « le terme complexe de la langue cible n'est pas typiquement composé de la traduction des parties du terme de la langue source », par exemple la traduction anglaise de « curage axillaire » est « axillary dissection ». Nous pouvons l'élargir à toutes les sortes d'EP. Cette caractéristique est un grand enjeu pour la traduction automatique puisqu'une traduction littérale sera probablement fausse. De ce fait, il nous faut une stratégie d'identifier correctement les frontières des EP d'abord et puis associer la signification correcte à l'expression détectée. Certaines méthodes se servent de ressources externes pour la détection de l'expression et la recherche de l'équivalent.

Quatrièmement, l'**ambiguïté**. Comme pour beaucoup d'autres tâches TAL, l'ambiguïté est aussi une des difficultés principales pour l'identification des EP. Dans ce cas, le choix est parmi l'interprétation idiomatique, l'interprétation littérale et la cooccurrence par hasard de la séquence rencontrée. Par exemple, pour les 3 phrases suivantes :

1) « Le président a retourné sa veste face aux nouvelles conditions. »
2) « Il a retourné sa veste et l'a suspendue dans l'armoire. »
3) « Dans la photo, Phillipe a sa tête retournée, sa veste bleue à la main. »

La première demande une compréhension idiomatique de l'expression polylexicale « **retourner sa veste** » qui signifie changer de camp ou d'opinion tout à coup ; quant à la deuxième, il faut la comprendre littéralement, « veste » ici désigne un vêtement réel ; finalement, dans la troisième phrase, le verbe « retourner » ne gouverne même pas le groupe nominal « sa veste », leur voisinage (après la virgule) n'est qu'une coïncidence. L'humain peut choisir assez facilement entre les interprétations idiomatique et littérale selon le contexte, ou remarquer à premier vue l'appartenance des mots aux syntagmes différents, mais ce n'est pas le cas pour les machines. Dans certains cas, le parsing pourrait aider à déterminer si la séquence apparue est une vrai EP, puisque le groupement de l'EP fautive provoque une erreur grammaticale.

Finalement, la **variabilité**, qui est la formulation flexible de l'EP. Elle existe dans 2 aspects : d'un côté, certains composants ne sont ni lexicalisés ni figés dans l'expression, même si leur présence est obligatoire, c'est le cas surtout pour les déterminants, comme « **faire** ses **études** » et « **faire** les **études** », qui appartiennent à la même expression. L'analyse syntaxique pourrait aider à résoudre ce problème. De l'autre côté, pour les langues morphologiquement riches comme le français, les flexions des mots, qui font l'accord ou expriment un temps, génèrent plusieurs ou même centaines formes à chercher dans le texte, telles que « il **pose** une **question** » et « nous **posons** des **questions** ». Face à ces variations qui causent de grandes difficultés dans l'identification de l'expression, l'analyse syntaxique et la lemmatisation pourraient diminuer ses impacts. Néanmoins, il faut faire attention quand nous utilisons les informations syntaxiques, parce que les formes variantes peuvent aussi modifier les relations syntaxiques relevées dans la forme canonique de l'expression.

En conclusion, les EP sont intrinsèquement décomposables en plusieurs lexèmes et présentent des qualités idiomatiques sous l'aspect lexical, syntaxique, sémantique, pragmatique et statistique. S'agissant des EPV, le projet PARSEME les divise pratiquement en 5 catégories : LVC, VID, IRV VPC et MVC. La discontinuité et la variabilité des EPV peuvent causer des difficultés pour leur détection dans les textes courants, et sa non-compositionnalité sera le centre d'intérêt dans notre mémoire.

CHAPITRE 2

Expressions polylexicales en chinois

Sommaire

Dans ce chapitre, nous présentons d'abord les principales spécificités de la langue chinoise par rapport à la langue française, et illustrons ses influences sur les EPV en chinois de manière statistique. Nous finissons ce chapitre en discutant les influences de ces différences sur l'identification des EPV en chinois.

2.1 Enjeux spécifiques au chinois

Les parties ci-dessus présentent les expressions polylexicales pour l'anglais et le français. Néanmoins, quant au chinois, une langue extrêmement éloignée de ces deux langues, à cause de différentes caractéristiques langagières, les difficultés principales se focalisent sur d'autres aspects.

Tout d'abord, d'un côté, une grande proportion des variations n'existe pas en chinois, ce qui facilite énormément l'identification des EP dans le texte courant. Le chinois ne dispose d'aucun comportement morphologique : les différents temps sont exprimés soit par les adverbiaux temporels, soit par les 3 caractères « 着 » « 了 » « 过 » qui signifient respectivement la progressivité, l'achèvement et le passé d'une action ; tandis que les adjectifs et les adverbes sont distingués par 2 caractères – « 的 » (pour les adjectifs) et « 地 » (pour les adverbes), qui relient les modifieurs avec les substantifs ou les verbes, mais il est possible d'omettre ces deux caractères quelquefois. De plus, la présence des déterminants comme les articles indéfinis et les articles définis n'est pas obligatoire. Prenons l'exemple suivant :

(1) 他 能 为 我们 提供 服务。 (zh)
tā néng wèi wǒmen **tígòng fúwù**
il pouvoir pour nous offrir service
'Il peut nous **offrir** des **services**.'

Pour l'expression « 提供 服务 » et « offrir service », en français, le déterminant « des » est une insertion externe dont la forme variée embarrasse le traitement, et l'absence de « des » avant « services » conduit à une faute grammaticale dans la pratique, mais cela est tout à fait acceptable en chinois. Ces deux particularités en chinois limitent la variation des EP aux aspects d'ordre et de discontinuité, donc réduit énormément les variantes possibles.

Néanmoins, puisque différentes catégories morpho-syntaxiques (nous utilisons ci-après son acronyme anglais POS) pour une même signification sémantique partagent souvent la même forme (voir l'exemple suivant), c'est le cas surtout pour les verbes et les noms verbaux ou les adjectifs et les adverbes, la distinction entre certaines catégories devient plus vague, et l'étiquetage automatique des POS rencontre plus de difficultés et devient moins fiable. Par conséquent, si le modèle dépend trop des étiquettes POS annotées par la machine, il est probable que l'on soit mis en erreur par les catégories morpho-syntaxiques.

(2) 她 很 美 (zh)
elle vraiment **belle**
'Elle est belle.'

(3) 他 被 她的 美 吸引 了 (zh)
il (prép.passif) sa beauté attirer (part.changement)
'Il est attiré par sa **beauté**.'

(Remarque : ici, le caractère 了 signifie qu'« il » *devient* être attiré par la beauté. On souligne un changement de son état : de l'état précédent « non attiré » à l'état actuel « être attiré ».)

(4) 想 得 美! (zh)
penser (part.complément) **bien**
'Tu rêves !'

De l'autre côté, comme il n'y a pas d'espaces entre les mots en chinois, cela cause une difficulté majeure pour le TAL. En fait, la segmentation d'une phrase dépend fortement de l'interprétation humaine en contexte, qui est souvent une tâche de désambiguïsation difficile. Voici quelques exemples caractéristiques pour illustrer ce problème (les différentes tonalités [1] sont mises en gras dans les exemples) :

(5) 南京 市长 江 大桥 (zh)
nánjīng shì**zhǎng** jiāng dàqiáo
Nanjing maire JIANG Daqiao
'Le maire de Nanjing Daqiao JIANG'

(6) 南京 市 长 江 大 桥 (zh)
nánjīng shì **cháng** jiāng dà qiáo
Nanjing ville Yangzi fleuve grand pont
'Le pont de Yanzi de la ville de Nanjing'

Cette ambiguïté de la segmentation influence aussi l'identification des EP. Par exemple, les deux caractères 成为 peuvent être traité comme un MVC ou séparé en 2 mots différents selon la phrase considérée :

(7) 他 想 成为 一个 工程师 (zh)
tā xiǎng chéng**wéi** yígè gōngchéngshī
il vouloir devenir un ingénieur
'Il veut devenir ingénieur.'

1. En mandarin, il y a 5 tons différents, qui se manifestent par les diacritiques sur les pinyin (transcription des sinogrammes en alphabet latin) : premier ton (¯), deuxième ton(´), troisième ton (ˇ), quatrième ton (`) et un ton neutre (sans accent sur la lettre).

(8) 他 被 当成 为 国 捐躯 的 战士。 (zh)
tā bèi dāngchéng **wèi** guó juānqū de zhànshì
Il (prép.passif) considérer_comme pour pays mourir (part.nom) soldat
'Il est considéré comme un soldat qui se sacrifie pour le pays.'

En tant que langue tonale, le chinois utilise souvent de différentes tonalités pour exprimer des significations différentes. De ce fait, la désambiguïsation est souvent réalisée par l'accent tonique, qui peut également servir à la segmentation des phrases en mots. Cela est explicitement montré par les exemples présentés ici. Néanmoins, cette distinction marquée par la tonalité n'existe pas toujours et les informations de prononciation ne sont pas présentes dans les textes courants.

Enfin, un aspect spécifique à prendre en compte est qu'il existe différents systèmes d'écriture en chinois. En effet, pendant les années 50, pour faciliter l'alphabétisation de la population illettrée, la Chine a popularisé un système des sinogrammes simplifiés, qui est officiellement considéré comme l'écriture standard en Chine maintenant. Néanmoins, les sinogrammes traditionnels existent toujours et sont officiellement utilisés dans les régions comme Hong Kong, Macao et Taïwan [2]. La table 2.1 illustre cela par un exemple de cette différence. De ce fait, si le corpus d'entraînement est écrit en caractères traditionnels et que nous appliquons le modèle entraîné sur un corpus de caractères simplifiés, le système sera inopérant, et vice versa. Heureusement, il existe des outils de conversion performants, donc ce problème peut être résolu facilement par l'enrichissement du dictionnaire et du corpus d'entraînement par les formes d'un autre système.

Signification française	**Sinogramme traditionnel**	**Sinogramme simplifié**
avoir une question	有 疑問	有 疑问
devenir	成為	成为

Table 2.1 – Exemple des sinogrammes traditionnels et simplifiés

2.2 Comparaison statistique des spécificités

À partir du corpus que nous utilisons, le corpus PARSEME, nous menons une étude comparative des EPV apparues, qui nous permet de projeter notre connaissance du chinois sur le corpus et la tâche précise.

Le corpus PARSEME [Ramisch et al., 2020], qui sert à l'identification des EPV, est un corpus multilingue. Sa version 1.2 contient 14 langues au total : basque, chinois, français, allemand, hébreu, hindi, irlandais, italien, grec moderne (1453-), polonais, portugais, roumain, suédois et turc. PARSEME se concentre surtout sur des langues européennes, mais certaines langues asiatiques sont également présentes, y compris le chinois. Dans notre travail, nous nous concentrons sur les parties française et chinoise.

Le corpus français annoté manuellement dans PARSEME a 4 sources : Sequoia Treebank, corpus français de UD « GSD » (fr_gsd-ud), la partie française du corpus ParTUT UD (fr_partut-ud) et les 500 premières phrases françaises du treebank Parallel Universal Dependencies (PUD). Il contient au total 20 961 phrases, et 525 992 tokens.

2. À noter que les systèmes d'écriture des caractères traditionnels ne sont non plus 100% identiques dans ces régions.

Pour l'étude des EPV en chinois, nous profitons de la partie chinoise de PARSEME, qui vient de Universal Dependencies [3] (GSD, HK et PUD) et CoNLL 2017 Shared Task [4] (Crawl-ooo et Wiki-000). Le dernier sont des articles de pages web ou de Wikipédia. À noter que tout le corpus chinois utilise le système d'écriture traditionnel. Il a au total 39 929 phrases et 649 576 tokens.

Premièrement, nous explorons la répartition des EPV. La table 2.2 compte le nombre d'occurrences des EPV par catégorie de 2 façons différentes : l'une ignore l'influence de répétitions en incluant toutes les occurrences ; l'autre essaie d'exclure cet impact en ne comptant que les occurrences des EPV distinctes. Selon cette table, nous remarquons certains points communs, dont les LVC sont plus fréquents que les VID et que les LVC.full sont plus nombreux que les LVC.cause. Ce décompte révèle également deux différences essentielles :

(1) En chinois, VPC.semi est la catégorie la plus fréquente ; en revanche, VPC n'existe même pas en français ; de plus, la catégorie MVC prend la deuxième place en chinois, alors qu'elle est la dernière pour le français.

(2) LVC.full et VID, les deux catégories les plus fréquentes en français, sont beaucoup moins fréquentes en chinois. Surtout pour VID, nous voyons beaucoup plus de répétitions en français, mais ce n'est pas le cas en chinois. La liste des 10 EPV les plus fréquentes pour chaque catégorie (voir l'annexe A) démontre également cet écart de répétition.

Type	Décompte avec répétitions		Décompte sans répétitions	
	Nombre	Pourcentage (%)	Nombre	Pourcentage (%)
ZH				
VPC.semi	3564	38.89	1388	33.50
MVC	3622	39.52	1337	32.27
LVC.full	1054	11.50	675	16.29
VID	757	8.26	625	15.09
LVC.cause	167	1.82	118	2.85
Total	9164		4143	
FR				
LVC.full	1878	33.22	857	47.53
VID	2156	38.13	577	32.00
IRV	1501	26.55	283	15.70
LVC.cause	97	1.72	80	4.44
MVC	22	0.39	6	0.33
Total	5654		1803	

TABLE 2.2 – Bilan des nombres des EPV

En fait, ces différences vérifient statistiquement un argument précédemment proposé sur l'idiomatisme sémantique, c'est que les EP les plus souvent utilisées sont

3. https://universaldependencies.org/
4. http://universaldependencies.org/conll17/

généralement plus compositionnelles, ou au moins plus prêtes pour la décomposition [Keysar and Bly, 1995]. En général, dans la langue française, les idiomes verbaux (VID) les plus utilisés sont les EP les plus basiques et simples, la catégorie la plus lexicalisée – les verbes réflexifs (IRV) sont également très présents. Les LVC, bien que'elles soient les plus compositionnelles selon l'hypothèse, ont un nombre d'occurrences fortement limité par la spécificité de son sens. En revanche, pour la langue chinoise, les EPV les plus fréquentes sont dans MVC et VPC.semi, qui désignent les sens les plus généraux et larges, et sont souvent considérés comme un seul mot. Même si nous pouvons les décomposer en plusieurs lexèmes dans l'analyse sémantique, ces lexèmes sont presque inséparables dans la pratique pour le chinois moderne. D'ailleurs, tous les VID en chinois se répètent moins de 5 fois dans tout le corpus. Une explication de cette distinction radicale est que beaucoup de VID en chinois ont une source historique (des expressions remarquables dans la littérature classique, et puis la postérité les réutilisent jusqu'à nos jours) ou une histoire/un mythe derrière, de ce fait, leur forme figée les rend moins décomposables que les VID en français, et la signification de la plupart d'entre eux est obscure pour les personnes qui ignorent leur origine.

Deuxièmement, nous poursuivons la comparaison en comptant le nombre de tokens pour chaque occurrence d'EPV (voir la table 2.3). En chinois, 48% des EPV sont déjà segmentées comme un seul token, et les EPV ayant plus de deux tokens sont rarement présentes. Au contraire, le corpus français n'a que 5 occurrences d'expressions contenant un seul token (c'est l'EPV « contre-indiqué »), et 23% des EPV contiennent plus de deux tokens. Plus précisément, en chinois, les EPV de token unique apparaissent plutôt dans VPC.semi et MVC, mais elles représentent une proportion beaucoup plus importante dans VID (88%), qui signifie que la plupart des VID en chinois sont déjà fortement lexicalisés comme une unité figée, et que les VPC.semi et les MVC ont une forte tendance d'inséparabilité dans la pratique.

En fait, la différence sur la répartition des longueurs d'EPV est fortement liée à la manière dont on segmente les phrases et à des considérations linguistiques. Les annotateurs du corpus chinois annotent les EPV selon les lexèmes au lieu des tokens afin d'être indépendant du problème de segmentation en chinois [5], et l'unité la plus petite d'un lexème est un seul caractère, c'est pour cette raison qu'il y a des EPV ayant un seul token (mais certainement plusieurs lexèmes). Néanmoins, le corpus français considère chaque token comme un composant dans l'EPV.

Finalement, à l'égard de la discontinuité des EPV, nous voyons une différence notable entre ces deux langues selon nos constatations précédentes. La table 2.4 montre la proportion des EPV contiguës dans chaque catégorie et au total. 87% des EPV chinois sont contiguës, contre 58% en français. À part les MVC, dont les composants sont presque toujours collés ensemble dans les deux langues, nous constatons aussi cette forte tendance d'adjacence dans les VID et VPC.semi en chinois. En fait, ce voisinage fort est préalable pour la présence fréquente des EPV ayant un seul token, comme ce qui est mentionné précédemment. Mais dans les deux langues, les VID ont beaucoup plus de chances d'être contiguës que les LVC.

5. Par exemple, dans l'annexe A, nous remarquons que le VPC.semi (zh) 看到 'voir' est quelquefois segmenté comme un seul token 看到 et quelquefois comme deux tokens 看 到, mais le nombre de lexèmes est toujours 2.

Catégories	**ZH**		**FR**	
	Pourcentage d'EPV ayant un seul token		**Pourcentage d'EPV ayant un seul token**	
LVC.full	0		0	
LVC.cause	0		0	
VID	88%		0.20%	
VPC.semi	63%		_	
IRV	_		0	
MVC	41%		0	
Total	48%		0.09%	
	Nombre de tokens dans un EPV	**Nombre d'occurrences**	**Nombre de tokens dans un EPV**	**Nombre d'occurrences**
Total	1 token	4416	1 token	5
	2 tokens	4614	2 tokens	4351
	plus de 2 tokens	134	plus de 2 tokens	1298

TABLE 2.3 – Bilan des nombres de tokens par EPV

	ZH	FR
Nombre de EPV contiguës	MVC : 3480 VPC.semi : 3293 VID : 725 LVC.full : 413 LVC.cause : 31	VID : 1660 IRV : 1283 LVC.full : 347 MVC : 21 LVC.cause : 5
Proportion des EPV contiguës	MVC : 96.1% VID : 95.8% VPC.semi : 92% LVC.full : 39% LVC.cause : 19%	MVC : 95% IRV : 85% VID : 76% LVC.full : 18% LVC.cause : 5%
Proportion globale des EPV contiguës	87%	58%

TABLE 2.4 – Bilan de la continuité des EPV

2.3 Influences de ces spécificités sur l'identification des EP en chinois

Selon les caractéristiques présentées ci-dessus, nous voyons clairement que la variation des formes n'est plus un obstacle majeur pour l'identification des EP en chinois, la tâche peut même être simplifiée jusqu'à une simple recherche des lexiques EP dans le corpus. Si nous disposons de ces lexiques, nous pouvons facilement obtenir un résultat satisfaisant. La difficulté se concentre maintenant sur la découverte des EP moins fréquentes et des néologismes, ainsi que la création automatique de lexiques fiables. De plus, sauf les LVC, la majorité des EP en chinois est contiguë, donc l'identification des EP pourrait se transformer à une question de segmentation en mots : il nous faut trouver les frontières entre les caractères qui regroupe ces multiples lexèmes en un seul mot, ou corriger les segmentations fausses afin que les composants d'une EP soient concaténés ensemble. Si la phrase est segmentée correctement, l'essentiel du problème serait déjà résolu sauf pour les cas discontinus, même s'ils reste marginaux. Puisque le chinois est beaucoup moins structuré grammaticalement, l'étiquetage POS et l'analyse syntaxique sont moins fiables, il vaut mieux se pencher sur les méthodes statistiques ou sémantiques pour résoudre ce problème.

En conclusion, nous pointons d'abord le fait que le chinois ne présente pas de difficulté morphologique, qui peut d'un côté diminuer les variations des EPV, d'un autre côté introduire plus d'erreurs pendant l'étiquetage de POS. Par conséquent, l'identification des EPV ne doit pas trop dépendre des informations POS. Deuxièmement, l'absence d'espace entre les mots demande un prétraitement de segmentation en mots avant la tokenisation, qui peut engendrer de nombreuses erreurs, étant une difficulté majeure. Si les phrases sont correctement segmentées en mots, la détection des EPV se réduit à une recherche directe des EPV. Nous remarquons finalement que les EPV en chinois ont plus tendance d'être contiguës par rapport aux EPV en français.

CHAPITRE 3

Hypothèse

Selon notre présentation précédente, les EP se caractérisent principalement par la non-compositionnalité sémantique, qui signifie que le sens global d'une expression n'est pas régulièrement déductible à partir des sens de ses composants. Puisque les composants contribuent peu au sens de l'expression entière, ils ne devraient pas être accessibles par les entités externes de l'expression d'un point de vue sémantique. Nous étudions cet aspect en le croisant avec un autre phénomène linguistique dont l'analyse nécessite des expressions compositionnelles : la coréférence.

D'après [Désoyer et al., 2015], le concept de coréférence relève de la notion d'anaphore en linguistique, qui désigne une procédure référentielle de renvoi d'une expression anaphorique du discours immédiat à un antécédent. L'interprétation de l'expression anaphorique s'appuie sur son antécédent, et leur relation est asymétrique. Quant à la coréférence, la notion est généralisée à une « relation symétrique entre plusieurs expressions référant à une même entité », la distinction entre l'antécédent et l'anaphore n'existe plus. Autrement dit, c'est le phénomène où plusieurs éléments d'un discours réfèrent à une même entité du monde du discours. Dans l'exemple (1) suivant, « sa veste » et « l' » forme une chaîne de coréférence, ils sont tous les deux des « mentions ».

(1) Il a retourné sa veste et l'a suspendue dans l'armoire.

(2) Le président a **retourné sa veste** au dernier moment : il soutient finalement le maire.

Revenons à notre expression polylexicale non compositionnelle, dans l'exemple (2), qui utilise l'interprétation figurée de l'expression « retourner sa veste », donc une vraie EP, le composant « sa veste » ne signifie pas une entité « veste » dans le monde du discours donc il est impossible de le lier avec une autre entité mentionnée dans le texte. En fait, [Laporte, 2018] propose de considérer les restrictions sur les chaînes référentielles comme critères définitoires des expressions polylexicales. Par exemple, l'expression compositionnelle dans la phrase (3) accepte la coréférence, tandis que l'expression non-compositionnelle dans phrase (4) l'interdit.

(3) Kathy avait une posture fière. Cette posture a été commentée.

(4) *Kathy était **en mauvaise posture**. Cette posture aurait pu être évitée.

En un mot, notre hypothèse de départ est que « **les composants individuels d'une expression polylexicale sont rarement susceptibles d'appartenir à des chaînes de coréférence** ». Notre expérience suivante essaie de vérifier quantitativement cette hypothèse en corpus.

Deuxième partie

Expérimentations

CHAPITRE 4

Présentation des outils

Sommaire

Dans ce chapitre, nous présentons les 2 outils qui permettent d'étudier notre hypothèse : Seen2Seen et OFCORS, en soulignant plus le côté Seen2Seen. Plus précisément, nous clarifions d'abord le problème à résoudre en TAL – l'identification des EPV et la résolution de coréférence, et résumons ensuite les approches principales de ces deux outils. étant donné notre intérêt particulier sur les EPV, nous présentons aussi les méthodes principalement utilisées pour l'identification des EPV.

4.1 Identification des EPV

4.1.1 Définition de tâche

[Constant et al., 2017] clarifient le traitement des EP comme deux tâches distinctes : la découverte (*MWE discovery* en anglais) des EP et l'identification des EP (*MWE identification*). Le premier se concentre sur la recherche des EP inconnues dans le corpus afin de créer automatiquement un lexique d'EP, mais avant d'être utilisable, ce lexique exige souvent une correction humaine. Quant à la tâche à laquelle nous nous intéressons, l'identification des EP, il s'agit d'annoter automatiquement les EP dans les textes courants, puis de lier les séquences détectées avec des EP connues. Son entrée est souvent un corpus, avec parfois des lexiques d'EP et des règles systématiques à suivre pour la détection. Sa sortie est souvent une liste d'annotations, autrement dit le corpus sera enrichi d'une couche d'annotation des EP. La nature de cette tâche incite à utiliser principalement des apprentissages supervisés. Son évaluation compare automatiquement les étiquettes ajoutées avec les annotations de référence.

Quant à l'utilité de la tâche, elle peut jouer un rôle de prétraitement pour les systèmes de parsing et de traduction automatique. Ce prétraitement permet au parseur de segmenter correctement les phrases, ou à la machine de réaliser des traductions non-littérales pour les EP. Nous pouvons en profiter aussi dans les traitements sémantiques : les parseurs sémantiques peuvent considérer les spécificités de LVC et d'idiomes pour construire

les structures prédicat-argument. Prenons l'exemple de la LVC « **rendre** une **visite** », le nom « visite » est le vrai prédicat sémantique, tandis que le verbe ne fonctionne que comme la liaison entre le sujet et le prédicat nominal. De ce fait, au lieu de considérer le nom « visite » comme un argument syntaxique du verbe « rendre » comme ce qui est fait par les parseurs syntaxiques, un parseur sémantique préfère probablement identifier le nom « visite » comme le prédicat en le mettant comme l'un de ses arguments sémantiques. D'ailleurs, les systèmes de recherche d'information peuvent l'utiliser pour indexer les expressions importantes, et les systèmes de désambiguïsation sémantique de mots peuvent en bénéficier afin d'éviter l'étiquetage erroné sur les composants individuels des EP.

4.1.2 Méthodes principales

Pour identifier les EP dans un texte, [Constant et al., 2017] donnent trois pistes principales suivies par les chercheur : les méthodes basées sur les règles et les approches d'apprentissage automatique, qui sont inspirées soit par la désambiguïsation de sens, soit par l'étiquetage en séquences.

Premièrement, les méthodes basées sur les règles. Les méthodes de ce genre peuvent être aussi simples qu'une recherche directe(*direct matching*) de formes identiques ou devenir plus compliquées en définissant des contraintes. Les premières approches utilisent des transducteurs à états finis (*finite-state transducers*), un choix courant pour la détection des EP contiguës comme les mots composés. Pour obtenir les formes à chercher, les transducteurs génèrent un dictionnaire des EP fléchies à partir d'une liste de leurs formes canoniques. La recherche est extrêmement efficace du point de vue de la complexité en temps, mais la génération manuelle des dictionnaires est vraiment coûteuse. La deuxième approche inverse les deux phases : elle réalise d'abord l'analyse morphologique des mots simples, et puis identifie les EP par une composition des règles qui traite des variations morphologiques et syntaxiques sous restriction. Elle est plus efficace pour traiter la variabilité et la discontinuité à courte distance. Les approches plus récentes simplifient les précédents en ne traitant que les flexions de mots (étiquetage POS + lemmatisation). Mais elles ignorent l'accord éventuel dans l'expression. Finalement, on peut utiliser la recherche de patrons (*pattern matching*) sur les textes prétraités (par exemple, l'étiquetage POS) comme l'outil mwetoolkit [Ramisch, 2014]. Cette approche fonctionne comme les expressions régulières mais au niveau des tokens, cela permet de spécifier ou nier certaines POS, lemmes ou formes des tokens dans la séquence, et de traiter également la répétition ou l'inconnu en leur sein. Mais cet outil nous demande de définir un patron pour chaque EP, et la flexibilité des ordres n'est pas prise en compte.

Deuxièmement, les approches d'apprentissage automatique. D'un côté, inspirées par la désambiguïsation qui met l'accent sur les informations contextuelles, certaines méthodes considèrent l'identification des EP comme une tâche de classification spécialisée en contexte. Le classifieur détermine si le candidat donné est une vraie EP ou seulement une cooccurrence régulière. Les traits contextuels souvent utilisés sont les mots qui les entourent, leurs étiquettes POS, lemmes, caractéristiques syntaxiques comme la dépendance syntaxique et information distributionnelle. Il est à noter que ces méthodes ne prennent pas en compte la détection des candidats, puisqu'elle est considérée comme un prétraitement hors sujet lors des compagnes d'évaluation, mais cette étape est indispensable en situation réelle. Les méthodes comportent des classifieurs supervisés comme machines à vecteurs de support (SVM) et des classifieurs non-supervisés comme le clustering. Certaines méthodes utilisent aussi des plongements de mots, et comparent le vecteur du candidat détecté avec celui de l'usage idiomatique ou de la forme canonique pour décider

s'il est vraiment une EP. Parmi les travaux les plus récents réalisés dans le cadre de PARSEME, les méthodes neuronales sont utilisées pour l'identification des EPV. TRAPACCs [Stodden et al., 2018] combine les couches convolutionnelless avec un SVM ; et SHOMA [Taslimipoor and Rohanian, 2018] intègre d'abord le plongement de mots en utilisant les ressources externes et utilise ensuite un réseau contenant des couches de convolution et des couches récurrentes accompagné d'une couche CRF optionnelle. Une version améliorée SHOMA2019 [Rohanian et al., 2019] est proposée après afin de mieux traiter la discontinuité. L'amélioration se base sur les réseaux convolutifs sur graphe (*graph convolutional network*) et le self-attention multi-têtes (*multi-head self-attention* en anglais).

D'un autre côté, certaines méthodes s'inspirent de l'étiquetage en séquences. Ce type de question peut être considéré comme une tâche de classification spécifique qui demande de donner à chaque token dans la phrase une étiquette, souvent influencée par les étiquettes précédemment annotées, afin de localiser des segments conformes à nos demandes. L'étiquetage en POS et l'annotation des entités nommées sont deux tâches typiques de ce genre. S'agissant de notre tâche précise, les étiqueteurs d'EP utilisent l'approche d'apprentissage supervisé structuré, donc ils sont entraînés à partir du corpus annoté en schéma BIO. Beaucoup parmi eux utilisent champ aléatoire conditionnel (*Conditional Random Field* en anglais, CRF), mais il ne convient que pour les EP contiguës. D'autres choix possibles sont les SVM ou perceptrons structurés [1]. Les traits donnés pour la prédiction sont souvent le contexte local (les mots à droite et à gauche), les traits au niveau des tokens (comme lemme et POS) et les ressources externes qui visent à établir les données annotées ou construire le modèle statistique.

4.1.3 Notre outil de départ : Seen2Seen

Notre outil de départ pour l'identification des EPV, Seen2Seen [Pasquer et al., 2020], a été développé dans le cadre de PARSEME 2018 (édition 1.1), dont la tâche se concentre sur l'identification des EPV connues [2]. Selon la définition du projet, une EPV est connue (*seen VMWE*) si une EPV ayant le même *multiset* [3] de lemmes est annotée au moins une fois dans l'ensemble d'entraînement ou l'ensemble de développement. La tâche est aussi divisée en deux pistes : la piste fermée (sans ressources externes) et la piste ouverte (avec ressources externes).

Étant donné 1) la difficulté majeure de l'identification des EPV inconnues, 2) l'espace d'amélioration encore existant pour les occurrences morphosyntaxiquement variées des EPV connues, et 3) la proportion dominante des EPV connues dans le corpus, Seen2Seen vise à améliorer le résultat global par l'amélioration du résultat des EPV connues.

Fondée sur des règles fortement interprétables, sa méthodologie se divise en deux étapes.

La première étape extrait des candidats possibles à partir des *multisets* de lemmes des EPV déjà vues dans le corpus d'entraînement. Cette étape d'entraînement garantit un rappel élevé au prix de précision.

La deuxième étape constitue une composition des filtres, qui essaie de sélectionner les EPV correctes. Ces filtres se fondent sur quatre hypothèses a priori :

— **h1** : il faut chercher dans le corpus d'entraînement les cooccurrences des lexèmes précises (avec le POS) qui sont annotées comme EPV ;

1. Modèle structurel entraîné sur beaucoup de traits.
2. La dernière édition de PARSEME, édition 1.2, se concentre sur l'identification des EPV inconnue, qui n'est pas traité dans cet article. Mais notre corpus utilisé vient de l'édition 1.2.
3. Un ensemble désordonné qui permet à l'apparition plusieurs d'un même élément.

— **h2** : il faut permettre seulement les variantes morphosyntaxiques déjà vues dans l'entraînement ;
— **h3** : il ne faut pas compter trop sur l'annotation automatique des étiquettes POS et des relations de dépendance, puisqu'elle peut être bruitée ;
— **h4** : il faut considérer la cohérence syntaxique afin d'éliminer les cooccurrences par hasard.

Huit filtres binaires et indépendants sont proposés :

— **f1** - désambiguïsation par POS : Ce filtre se fonde sur h1 et h3. Il demande aux candidats et leurs EPV comparées (annotées dans le corpus d'entraînement) d'avoir les mêmes *multisets* des étiquettes POS.
— **f2** – précision de l'ordre : basé sur h2, ce filtre vérifie si l'ordre des POS des composants lexicalisés détectés existe déjà dans les EPV observées en ignorant la contrainte de discontinuité.
— **f3** – précision de l'ordre avec la prise en compte de discontinuité : idem que le filtre précédent mais ce troisième filtre considère aussi la possibilité des insertions externes dans les composants lexicalisés.
— **f4** – limite de la distance : inspiré par h4 et h3, ce filtre élimine les candidats qui ont une distance de discontinuité plus longue que la distance maximale observée dans cette catégorie pendant l'entraînement.
— **f5** – préférence des candidats : ce filtre renforce le précédent en préférant les candidats qui ont une discontinuité plus courte dans la phrase.
— **f6** – contrainte syntaxique : basé sur h4, ce filtre exige que les composants soient syntaxiquement connectés. Plus précisément, pour une EPV contenant plus de 2 composants lexicalisés, ces composants doivent former un sous-arbre de dépendance ; pour celles ayant 2 composants, ces deux doivent être liés dans une chaîne syntaxique dans laquelle au plus une insertion externe est tolérée entre eux.
— **f7** – contrainte sur le composant nominal : basé sur h2, ce filtre limite les flexions nominales aux cas déjà vus dans l'entraînement pour les EPV contenant un seul nom, du fait que l'inflexibilité des flexions nominales existe fortement dans les idiomes. De plus, ce filtre vérifie la déclinaison aux cas pour les langues casuelles (comme l'allemand) et la déclinaison aux nombres pour d'autres langues, telles que le français.
— **f8** – contrainte sur l'EPV imbriquée : pour l'imbrication des EPV, la conservation des EPV imbriqués est décidée par l'EPV qui les englobe : ce filtre ne permet que les imbrications déjà constatées lors de l'entraînement.

Afin de trouver la combinaison optimale des filtres activés parmi ces 8 choix, toutes les possibilités, soit 256 combinaisons, sont testées sur le corpus de développement. L'outil garde finalement celle qui obtient la F-mesure la plus élevée pour l'identification des EPV connues. Pour la langue française, la combinaison optimale sur PARSEME 1.2 est f1,f2,f5,f6,f7 et f8 dans l'ordre, qui est également la combinaison utilisée par nous pour l'annotation des EPV dans notre tâche. La F1-mesure des EPV connues est 89.75 pour le français, avec une meilleure performance en rappel (93.86) qu'en précision (85.99) [4], contre 80.47 comme score global de toutes les 14 langues [5], et la F1-mesure stricte (correspondance par EPV, pas de correspondances partielles des tokens) de toutes les EPV

4. Ce chiffre est l'évaluation du résultat obtenu dans notre expérience, la définition de « seen » utilise la nouvelle règle : EPV apparue dans train ou dev. Le résultat est un peu différent de celui sur le site. Dans notre expérience, le résultat global de seen+unseen : MWE-based F1 – 78.68 ; token-based F1 : 79.85

5. Ce score et les scores suivants viennent de la page officielle du PARSEME : http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_02_MWE-LEX_2020___lb__COLING__rb__&subpage=CONF_50_Shared_task_results#definition-change

(connues et inconnues) est de 78.63. La F1-mesure stricte pour toutes les EPV en chinois est seulement 49.28, ce qui est probablement dû au fait que moins de travail est fait sur cette langue lors la création du Seen2Seen.

4.2 Résolution de coréférence

4.2.1 Aperçu de la tâche

La résolution de coréférence constitue un objet d'étude important en TAL daté des années 70. Comme ce qui est présenté précédemment (chapitre 3), chaque pronom ou groupe nominal (y compris des noms propres) dans le discours, appelés spécifiquement comme des « mentions » pour cette question, désigne une entité dans le monde réel, ou au moins dans le monde du discours s'il est fictif. À noter que plusieurs mentions peuvent pointer vers une même entité. Ces mentions qui réfèrent les unes aux autres construisent une chaîne de coréférence, et cette chaîne peut évidemment dépasser les frontières des phrases. La résolution de coréférence cherche à lier les mentions à leur entité désignée, afin d'aider la machine à la compréhension des textes. Concrètement, il faut d'abord identifier les mentions dans le texte et puis relier celles qui réfèrent à la même entité. Il existe deux difficultés majeures dans cette tâche : d'un côté, l'identification des mentions est gênée par la variation riche des formes de mentions ; de l'autre côté, qui est aussi le plus difficile, l'enchaînement correct des anaphores infidèles exige souvent des inférences fondées sur la compréhension sémantique des mentions, voire des connaissances du monde (ressources externes) quelquefois. Par conséquent, la reconnaissance automatique des chaînes de coréférence est toujours une tâche très difficile et généralement assez partiellement résolue, avec de faibles performances.

4.2.2 Notre outil de départ : OFCORS

Afin d'obtenir les chaînes de coréférence dans un texte, nous utilisons l'outil end-to-end OFCORS. Il implémente la partie de détection des mentions d'un autre outil DeCOFre [Grobol, 2019] et fait la clustering à partir de ces mentions détectées. DeCOFre applique une méthode totalement neuronale, tandis qu'OFCORS choisit une approche plus traditionnelle d'apprentissage automatique supervisé.

La logique principale d'OFCORS suit celle du système antérieur CROC [Désoyer et al., 2015]. Plus précisément, dans un premier temps, l'outil annote le texte en entrée avec Spacy. Disposant des mentions trouvées par DeCOFre, la construction d'une chaîne de coréférence suit une stratégie *pairwise* : il transforme la tâche en une classification binaire – le classifieur doit décider si une paire de mentions est coréférente ou non. L'annotation automatique permet au classifieur d'extraire un ensemble de traits linguistiques pour les mentions paires, tels que leur distance (en nombre de mentions/mots/caractères), l'identité de leur genre et nombre, leur catégorie morpho-syntaxique (nom. pronom, etc.), l'identité de leurs formes, etc. À partir de ces traits, différents modèles d'apprentissage automatique sont entraînés pour réaliser la classification. Pour former les paires candidats, l'algorithme cherche soit toutes les paires (stratégie « combinations ou permutations »), soit les paires dans une fenêtre (stratégie fenêtrage, par exemple, on choisit les combinaisons de la mention et ses k précédentes mentions à gauche pour une fenêtre de taille k). La reconstruction de la chaîne de coréférence à partir des paires a aussi deux stratégies différentes à choisir : « *closest-first* » (la mention concernée et sa mention voisin la plus proche à gauche) et « *best-first* » (la mention concernée et celle ayant le plus haut score de « probabilité coréférentielle »

parmi toutes les mentions précédentes). Ayant des paires coréférentes, l'algorithme finit par regrouper les paires qui possèdent les mentions en commun dans de mêmes clusters, soit les chaînes de coréférence sortantes.

Dans notre tâche, d'après une observation qualitative des résultats obtenus sur plusieurs essais, nous utilisons finalement la stratégie « fenêtrage » avec fenêtre de taille 8 pour la génération des paires ; quant au classifieur, nous choisissons le modèle Random Forest, qui est le plus performant selon le développeur de l'outil. Une dernière remarque est que l'outil est entraîné sur un corpus français de l'oral transcrit, donc l'appliquer sur un corpus écrit conduit à une forte baisse de performance.

En conclusion, les deux outils utilisés pratiquent des stratégies différentes : Seen2Seen se base sur l'analyse linguistique des EPV et utilise une méthode fondée sur les règles, tandis qu'OFCORS applique des méthodes neuronales et statistiques.

Chapitre 5

Corpus et méthode

Sommaire

Dans ce chapitre, nous présentons d'abord nos critères pour le choix de corpus à examiner, et nous décrivons chaque corpus : Sequoia, Est Républicain et ANCOR. Dans la deuxième section, nous présentons d'abord le pipeline pour fusionner les résultats de deux systèmes de manière générale, et mettons en avant les étapes clés dans notre travail : la démarche pour obtenir le corpus sous le format d'entrée demandé, la présentation du format de sortie et l'alignement des tokens pour la fusion des résultats.

5.1 Corpus utilisé

5.1.1 Choix de corpus

La vérification de l'hypothèse proposée exige que les annotations des EPV et des chaînes de coréférence soient suffisamment fiables, donc il est nécessaire d'examiner manuellement les annotations automatiques avant de passer à la suite. Afin d'alléger ce travail manuel, nous partons des corpus possédant déjà des annotations manuelles pour l'un de ces deux aspects, c'est-à-dire les corpus de référence de Seen2Seen et d'OFCORS, soit PARSEME et ANCOR.

Il est à noter que la coréférence ne peut être traitée qu'à l'intérieur d'une même unité discursive, par conséquent, le corpus doit pouvoir être séparé au niveau des textes. Néanmoins, l'information sur les délimitations des textes est perdue dans PARSEME et une grande partie de ce corpus utilise même les phrases mélangées, sauf une partie du sous-corpus Sequoia. Ayant extrait cette partie, nous l'annotons avec OFCORS, et notre contrôle de validation ne s'effectue que sur les chaînes de coréférence.

Dans un deuxième temps, nous prenons le corpus oral ANCOR et nous l'annotons avec Seen2Seen. Cette fois-ci, notre vérification se concentre sur les EPV détectées.

Finalement, pour élargir notre validation, nous effectuons notre expérience sur un corpus brut, Est Républicain, qui est un corpus écrit regroupant des sujets variés. Ce corpus ne contient aucune annotation donc il faut faire attention à tous les deux aspects pendant la vérification manuelle.

5.1.2 Sequoia de PARSEME

Parmi les 4 sources du corpus français dans PARSEME, soit GSD-UD, ParTUT-UD, PUD et Sequoia, afin de vérifier notre hypothèse, nous utilisons seulement une partie de Sequoia, puisqu'elle est la seule partie où les phrases des mêmes articles sont bien ordonnées. Plus précisément, ce sont : 1) emea : contient 2 rapports de l'agence européenne du médicament, 2) frwiki : est constitué de 19 articles Wikipédia parlant des histoires sur des affaires sociales ou politiques, 3) annodis.ER : comporte 36 articles courts du journal Est Républicain avec des thèmes variés. La table 5.1 montre l'état des parties que nous avons utilisées :

Corpus	**Nombre de textes**	**Nombre total de phrases**	**Taille moyenne de textes (nombre de mots)**	**Nombre de mots par texte**
emea	2	Environ 1000 phrases	8267	Plus de 8 000 mots
frwiki	19	Environ 1000 phrases	988	13 articles ont moins de 1000 mots et 6 ont plus de 1000 mots
annodis.ER	36	Environ 500 phrases	264	Entre 120 et 600 mots, la majorité est moins de 400 mots
Total	57	Environ 2500 phrases	786	

TABLE 5.1 – Sous-corpus utilisés du corpus Sequoia

5.1.3 ANCOR et Est Républicain

ANCOR est un corpus oral transcrit qui est constitué de 4 parties : ESLO_ANCOR (environ 25 000 phrases au total), ESLO_CO2 (environ 2 500 phrases au total), OTG (environ 2 800 phrases au total) et UBS (environ 700 phrases au total). ESLO_ANCOR contient des entretiens divisés en sous-dialogues thématiques cohérents, ESLO_CO2 comporte 3 entretiens complets, OTG regroupe de dialogues interactifs entre des individus et le personnel d'accueil de l'office de tourisme de Grenoble et UBS est formé par des dialogues interactifs par téléphones recueillis auprès du standard téléphonique d'une université (voir l'annexe C.1 pour des exemples d'extrait). Il est à noter que l'ANCOR est annoté sous format TEI, de ce fait, il faut extraire ses textes bruts avant d'utiliser Seen2Seen et transformer ses annotations de coréférence au format sorti par OFCORS. De plus, étant un corpus oral, il ne dispose pas de ponctuations sauf les points d'interrogation. Dans notre traitement, nous considérons les frontières d'un tour de parole comme les délimiteurs des

phrases, mais ce manque de frontières des phrases pourrait impacter la performance du parseur de dépendance et celle de Seen2Seen. Finalement, par rapport au corpus écrit, sa caractéristique orale conduit également à un nombre beaucoup plus important de disfluences et de répétitions ou réutilisations des expressions, qui rend l'analyse des phrases demandée par Seen2Seen plus difficile. De plus, ANCOR annote toutes sortes d'anaphores y compris celles qui viennent de ces phénomènes, par exemple :

— …
— d' accord oui oui oui avec quoi est -ce que vous écrivez euh vous m' avez déjà dit un en général un stylo à bille mais vous essayez de réécrire avec un s- un stylo à plume un stylo à encre oui quand vous **avez le temps** oui d' accord oui oui oui oui d' accord quel type de papier est -ce que vous utilisez ?
— ça quand j' **ai le temps** hein parce que
— …

Dans ce cas-là, la deuxième personne reprend l'EPV utilisée par la parole précédente, même si la deuxième mention « le temps » est anaphorique, cette reprise ne demande pas une décomposition du sens de l'EPV. Par conséquent, il est fortement possible que les croisements des EPV avec ce genre de coréférence ne nous intéressent pas.

Finalement, le corpus Est Républicain est un recueil d'articles parus en 1999, 2002 et les deux premiers mois de 2003 dans le journal régional Est Républicain. Ce sont des articles courts portant sur divers thèmes. Considérant la taille considérable du corpus dans sa totalité et notre contrainte à la fois sur le calcul pour l'annotation et sur la force humaine pour la vérification, nous limitons notre expérience aux 100 premiers articles ayant plus de 300 mots de l'année 2003, soit environ 3 000 phrases en tout (ci-après nommé ER).

La table 5.2 récapitule les informations sur ces deux corpus.

Corpus	**Nombre de textes**	**Nombre total de phrases**	**Taille moyenne de textes (nombre de mots)**
ANCOR : ESLO_ANCOR	99	Environ 25 000 phrases	3837
ANCOR : ESLO_CO2	3	Environ 2500 phrases	12 246
ANCOR : OTG	313	Environ 2800 phrases	83
ANCOR : UBS	40	Environ 700 phrases	176
Est Républicain	100	2923 phrases	501
Total	555	environ 33 923 phrases	900

Table 5.2 – Corpus ANCOR et Est Républicain

5.2 Méthode

5.2.1 Aperçu de la chaîne de traitement

La figure 5.1 illustre notre chaîne de traitement d'une manière générale.

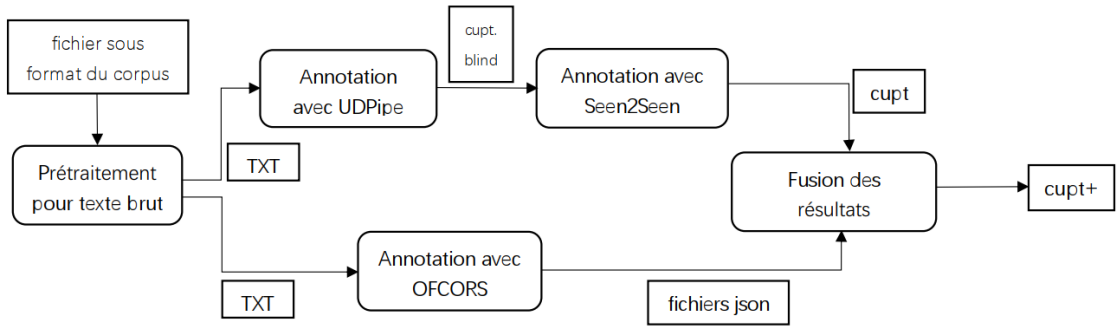


FIGURE 5.1 – Chaîne de traitements pour l'annotation et la fusion

Afin d'annoter le corpus avec des outils, il faut d'abord uniformiser les fichiers de différents formats définis par leur corpus d'origine au format texte brut. Ensuite, du côté Seen2Seen, l'outil requiert un texte en entrée annoté par UDPipe et sous format conllu+. Il nous fournit comme sortie une extension de ce format (nommée cupt) en ajoutant les annotations des EPV. Du côté OFCORS, l'outil nous fournit 3 fichiers json : la liste des tokens du texte, une liste des mentions et une liste des clusters. Nous fusionnons ces deux sorties en étendant encore le format cupt.

Les sections suivantes s'articulent autour des étapes clés dans notre traitement, soit l'extraction et la segmentation en articles des textes, la transformation des formats de fichier et l'alignement des résultats des outils pour la fusion.

5.2.2 Extraction des textes

La méthode pour extraire les textes dans chaque corpus est différente.

Premièrement, pour les sous corpus de Sequoia, il nous faut les fichiers TXT servant à l'entrée d'OFCORS et les formats de sortie de Seen2Seen – cupt (abréviation de « *CoNLL-U Plus Format* »).

Dans un premier temps, pour EMEA et FRWIKI dans Sequoia, afin de faciliter la lecture et l'annotation, nous commençons par extraire l'indice (sent_id) et le contenu (text) des phrases de chaque sous-partie en respectant l'ordre de `sent_id`. Ensuite, la délimitation des articles se réalise manuellement en ajoutant une marque (« `## DEBUT DOC` » et « `## FIN DOC` ») avant et après chaque article. Finalement, à partir de ces frontières annotées et les indices de phrases, nous arrivons à créer 2 fichiers pour chaque article : l'un est un fichier TXT, contenant seulement le contenu, dans lequel chaque ligne correspond à une phrase ; l'autre est un fichier cupt, qui regroupe pour chaque article les blocs d'annotation de ses phrases dans le fichier d'origine.

Dans un deuxième temps, nous trouvons que la sous-partie annodis.ER du Sequoia utilise le texte du corpus ANNODIS [1], qui est en fait une version annotée du corpus Est Républicain, donc elle a aussi une structure textuelle et un ordre parmi les phrases. Nous profitons de l'annotation du titre et du nom de chaque fichier d'article (les fichiers

1. Corpus ANNODIS : http://redac.univ-tlse2.fr/corpus/annodis/ANNODIS_rr.zip

`*.seg`[2]) pour obtenir un dictionnaire des dates et titres (`ER_info.json`). Après avoir corrigé manuellement certaines incohérences, nous utilisons ces titres pour localiser le début de chaque article dans le fichier cupt global de la sous-partie annodis.ER, afin d'extraire et créer le fichier du texte brut et le fichier cupt pour chaque article.

Ensuite, pour le corpus ANCOR, le problème principal est différent : les conversations différentes sont déjà séparées dans différents fichiers, mais sous format TEI, et nous voulons 1) extraire les fichiers TXT pour l'annotation de Seen2Seen et 2) transformer les annotations au format sortant d'OFCORS, soit 3 fichiers json : `*_tokens.json`, `*_mentions_output.json` et `*_resulting_chains.json`. Le fichier `*_tokens.json` est un dictionnaire des tokens dans l'article ayant son indice comme clé et le token comme valeur ; `*_mentions_output.json` est le dictionnaire de mentions, pour chaque mention numérotée dedans, il contient l'information de ses tokens, les contextes avant et après, les indices des tokens initial et final de la mention[3] ; finalement, `*_resulting_chains.json` contient les chaînes de référence annotées, appelées « clusters », et il inclut seulement les indices de clusters et celles des mentions. Des extraits de ces fichiers sont présentés dans l'annexe D. Il est à noter que généralement les ponctuations marquant les frontières des phrases n'existent pas dans ce corpus, de ce fait, nous considérons le changement de locuteurs comme frontières de phrases et marquons cette délimitation par un saut de ligne entre deux tours de parole, afin que l'outil UDPipe les traite vraiment comme deux phrases différentes.

Finalement, pour le corpus Est Républicain, nous n'avons qu'à extraire le texte brut servant de fichier d'entrée pour les deux outils. Les différentes fragments du corpus d'origine sont séparés par leurs dates et stockés sous format XML simple. Chaque fichier regroupe tous les articles publiés ce jour-là, séparés par les lignes vides et par les titres signalés par la balise `<head>`. À partir de ces fichiers XML, nous arrivons à créer un dictionnaire des articles contenant la date, le titre, le contenu et le nombre de mots de chaque article, ce qui nous permet de sélectionner 100 articles contenant chacun plus de 300 mots pour notre expérience.

5.2.3 Transformation des formats

Disposant au moins des fichiers TXT après les prétraitements, nous arrivons à l'étape de l'annotation automatique et de la fusion des résultats. Nous présentons ici une chaîne de traitement si seul le texte brut est disponible, il est à noter que certaines annotations automatiques ne sont plus nécessaires en fonction du corpus utilisé.

Seen2Seen a besoin d'un prétraitement supplémentaire sur les fichiers TXT puisqu'il prend en entrée un fichier au format cupt.blind, qui est le format CoNLL-U avec une colonne vide en plus, qui seront remplies par les annotations des EPV dans les fichiers de sortie. Afin d'obtenir les fichiers CoNLL-U, nous utilisons l'annotateur UDPipe et son modèle français `french-gsd-ud-2.5-191206.udpipe`[4], qui est la version utilisée pour le corpus PARSEME.

Le format cupt mentionné ici est le format spécifiquement utilisé par le corpus PARSEME[5]. Ce format sert à résoudre le problème d'étiquetage provoqué par les superpo-

2. Nous utilisons l'annotation des experts donc les fichiers dans `annotations_expert/texte/A/`

3. Le fichier de mentions transformé diffère des vrais fichiers de sortie sur 2 détails : 1) les indices des mentions commencent à 0 au lieu de 1 dans le fichier sortant, 2) le contexte avant et après la mention parlée est remplacé par son id dans ANCOR. Mais ces différences n'ont aucune influence sur notre résultat.

4. https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3131

5. http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_04_LAW-MWE-CxG_2018___lb__COLING__rb__&subpage=CONF_45_Format_specification

sitions des expressions polylexicales, telles que le cas de coordination ou l'imbrication d'une EP dans l'autre. Le schéma BIO, un schéma d'annotation populaire pour les entités nommées, conduit à des confusions dans ce cas. La figure 5.2 illustre un exemple fourni par [Constant et al., 2017] :

	Now	that	I	looked	the	dirty	word	up	,	I	understand	.
No. expression	1	1		2		3	3	2				
BIO	B	I	O	B	O	B	I	I	O	O	O	O
BIO (Schneider)	B	I	O	B	O	b	i	I	O	O	O	O
PARSEME	1:cat1	1	*	2 :VPC	*	3 :cat2	3	2	*	*	*	*
Signification	Maintenant que j'ai cherché (le sens de) ce gros mot, je comprends.											

FIGURE 5.2 – Schéma d'annotation pour les EP

Ici, le schéma BIO ne permet pas de distinguer les deux expressions « look up » (chercher) et « dirty word » (gros mot). Pour résoudre ce problème d'imbrication, [Schneider et al., 2014] propose d'utiliser les minuscules pour les EP imbriquées. Mais cela ne fonctionne pas avec les chevauchements. De ce fait, dans la colonne de l'annotation des EP [6], le format cupt numérote chaque expression en précisant sa catégorie après ce numéro sur la ligne de son premier composant lexical et en les séparant par deux points. Les étoiles dans cupt jouent le rôle de « O » dans BIO. Si un token appartient à plusieurs EP, toutes ces appartenances sont annotées et séparées par un point-virgule. Il est à noter que le numéro des EP recommence par 1 pour chaque phrase. La figure 5.3 est un exemple d'extraits du format cupt :

```
37  get         get         VERB  VB   VerbForm=Inf                    38  aux:pass     38:aux:pass  _          3:VID;4:IAV
38  rid         rid         ADJ   JJ   Degree=Pos                      35  advcl        35:advcl     _          3;4
39  of          of          ADP   IN   _                               41  case         41:case      _          4
40  distracting distracting ADJ   JJ   Degree=Pos                      41  amod         41:amod      _          *
41  elements    element     NOUN  NNS  Number=Plur                     38  obl          38:obl       SpaceAfter=No  *
42  .           .           PUNCT .    _                               3   punct        3:punct      _          *

# source_sent_id = http://hdl.handle.net/11234/1-2515 UD_English/en-ud-train.conllu email-enronsent11_01-0057
# text = Shucks, guess none of us three paid much attention to that.
1   Shucks  shucks  INTJ  UH   _                                       3   discourse    3:discourse  SpaceAfter=No  *
2   ,       ,       PUNCT ,    _                                       3   punct        3:punct      _          *
3   guess   guess   VERB  VBP  Mood=Ind|Tense=Pres|VerbForm=Fin        0   root         0:root       _          *
4   none    none    NOUN  NN   Number=Sing                            8   nsubj        8:nsubj      _          *
5   of      of      ADP   IN   _                                       6   case         6:case       _          *
6   us      we      PRON  PRP  Case=Acc|Number=Plur|Person=1|PronType=Prs  4   nmod     4:nmod       _          *
7   three   three   NUM   CD   NumType=Card                           6   nummod       6:nummod     _          *
8   paid    pay     VERB  VBD  Mood=Ind|Tense=Past|VerbForm=Fin        3   ccomp        3:ccomp      _          1:LVC.full
9   much    much    ADJ   JJ   Degree=Pos                             10  amod         10:amod      _          *
10  attention attention NOUN  NN   Number=Sing                         8   obj          8:obj        _          1
11  to      to      ADP   IN   _                                       12  case         12:case      _          *
12  that    that    PRON  DT   Number=Sing|PronType=Dem                8   obl          8:obl        SpaceAfter=No  *
13  .       .       PUNCT .    _                                       3   punct        3:punct      _          *
```

FIGURE 5.3 – Extrait d'un fichier cupt

Ayant l'annotation des EP, il faut ensuite envoyer le texte brut à OFCORS afin d'obtenir les mentions et les chaînes détectées. Puisque la sortie d'OFCORS n'est pas assez

6. Puisque PARSEME ne travaille que sur les EPV, il n'a pas défini des acronymes pour les mots composés nominaux et fonctionnels. Donc nous utilisons dans le tableau "cat1" et "cat2" pour signifier une catégorie spécifique.

intuitive, et le format CoNLL-U est non seulement extensible, mais aussi capable de visualiser plusieurs couches d'annotation en même temps, nous utilisons finalement le format cupt comme la base pour la fusion des annotations.

Nous suivons principalement la même logique pour concevoir notre format de fusion cupt+ [7]. Une colonne de mention (la 12e colonne) et une colonne de chaîne de coréférence (la 13e colonne, nommée `COREF`) sont ajoutées après la colonne de EP. La colonne de mention annote le numéro de mention si le token en fait partie, et s'il appartient à plusieurs mentions, toutes ces indices seront présents ici et séparés par points-virgules. Quant à la colonne de `COREF`, seulement les mentions dans une chaînes sont annotées ici, et l'indice de la chaîne le précède. Ces deux indices sont séparés par deux points, tandis que plusieurs appartenances aux mentions différentes sont divisées par des points-virgules. La figure 5.4 illustre l'extrait du fichier pour la phrase suivante, qui contient la 7e (en bleu) et 8e (soulignée) mention du texte :

« …L'affaire des caporaux de Souain, fusillés pour l'exemple, est un des cas parmi les plus flagrants et les plus médiatisés de l'injustice militaire durant la Première Guerre mondiale … »

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC PARSEME:MWE MENTION COREF
...
15  un  un  PRON  _  Gender=Masc|Number=Sing 0  root  _  _  *  7;8 50:8
16-17  des  _  _  _  _  _  _  _  *  *  *
16  de  de  ADP  _  _  18  case  _  _  *  7;8 50:8
17  les le  DET  _  Definite=Def|Number=Plur|PronType=Art  18  det  _  _  *  7;8 50:8
18  cas cas NOUN  _  Gender=Masc|Number=Plur 15  nmod  _  _  *  7;8 50:8
19  parmi  parmi  ADP  _  _  22  mark  _  _  *  8  50:8
20  les le  DET  _  Definite=Def|Number=Plur|PronType=Art  22  det  _  _  *  8  50:8
21  plus  plus  ADV  _  _  22  advmod  _  _  *  8  50:8
22  flagrants  flagrant  ADJ  _  Gender=Masc|Number=Plur 18  acl  _  _  *  8  50:8
23  et  et  CCONJ  _  _  26  cc  _  _  *  8  50:8
24  les le  DET  _  Definite=Def|Number=Plur|PronType=Art  26  det  _  _  *  8  50:8
25  plus  plus  ADV  _  _  26  advmod  _  _  *  8  50:8
26  médiatisés  médiatisé  ADJ  _  Gender=Masc|Number=Plur 22  conj  _  _  *  8  50:8
27  de  de  ADP  _  _  29  case  _  _  *  8  50:8
28  l'  le  DET  _  Definite=Def|Number=Sing|PronType=Art  29  det  _  SpaceAfter=No  *  8;9 50:8;40:9
29  injustice  injustice  NOUN  _  Gender=Fem|Number=Sing  18  nmod  _  _  *  8;9 50:8;40:9
30  militaire  militaire  ADJ  _  Number=Sing 29  amod  _  _  *  8;9 50:8;40:9
31  durant  durant  ADP  _  _  34  case  _  _  *  *  *
```

Figure 5.4 – Extrait d'un fichier cupt+

Les tokens « un des cas » constituent la 7e mention, et ces 3 tokens ne sont qu'une partie de la 8e mention « un des cas parmi les plus flagrants et les plus médiatisés de l'injustice militaire ». Dans ce cas-là, sur la colonne de mention, il y a « 7 ;8 » pour les trois premiers tokens et un seul « 8 » pour les tokens suivants. De plus, seulement la 8e mention est considérée dans une chaîne, donc dans la dernière colonne on ne voit que « 8 » et l'indice de la chaîne « 50 ».

La fusion des annotations prend en compte la cohérence des annotations sur les articles contractés entre les 2 systèmes, ce qui est illustré dans la 7e mention. L'annotation UDPipe décompose les articles contractés en les affichant sur 3 lignes : une pour la forme contractée et les autres pour ses composants. Mais l'annotateur utilisé par OFCORS, Spacy, ne fait jamais cette décomposition. Puisque notre format fusionné se base sur le format cupt et PARSEME n'annote jamais la forme contractée dans les EPV, nous choisissons d'ajouter aussi l'annotation sur les composants des articles contractés. Par conséquent, pour « un des cas », les tokens annotés sont finalement « un de les cas ».

Ayant ce format qui inclut toutes les annotations de deux côtés, nous pouvons finalement passer à l'analyse des croisements de ces deux phénomènes linguistiques.

7. Dans la pratique, les fichiers utilisent toujours l'extension cupt. Nous utilisons ici l'appellation cupt+ simplement pour faire la distinction entre les deux.

5.2.4 Alignement des sorties

Un problème principal pendant notre fusion des annotations est la tokenisation différente de 2 systèmes. Puisque Seen2Seen utilise UDPipe et OFCORS utilise Spacy dès le début, ce qui est impossible d'être unifié, il nous faut trouver un moyen permettant de s'affranchir de cette incohérence pendant l'ajout des colonnes `MENTION` et `COREF`.

L'incohérence provient des différents choix de tokenisation de chaque outil tant de manière générale que sur les tokens de mots multiples (ci-après MWT, l'acronyme de *multiword token* en anglais) spécifiques. Plus précisément, nous les remarquons principalement dans les aspects suivants :

(1) UDPipe ne considère aucun genre d'espaces (espaces, retour à la ligne, etc.) comme tokens mais Spacy prend les retours à la ligne `\n` en compte de temps en temps. De plus, dans UDPipe, les indices de tokens sont ceux dans la phrase au lieu des indices globaux dans un texte.

(2) Certaines erreurs de tokenisation, par exemple, la fraction « 1/10 000 » est correctement traitée comme un seul token dans Spacy, mais elle est segmentée en une séquence `['1', '/', '10', '000']` dans UDPipe. De plus, la tokenisation des symboles spéciales n'est pas non plus stable, prenons l'exemple de degré celsius « ℃ », cette fois-ci, tandis qu'UDPipe le traite correctement comme un seul token, Spacy le sépare en « ° » et « C », et dans le cas de « 2 ℃- 8 °C », Spacy colle faussement « C » et « - » comme un seul token.

(3) À part le traitement systématiquement différent sur les articles contractés présenté précédemment, les mots composés tels que « week-end » peuvent être traités comme un seul token ou une séquence de token (`['week', '-', 'end']`).

Face à tout cela, nous proposons la démarche suivante pour aligner les tokens des deux côtés par leurs indices :

(1) Du côté UDPipe, indexer tous les tokens de manière globale. Pour la décomposition des MWT, comme les articles contractés, seulement la forme contractée est numérotée, mais les composants sont notés dans le dictionnaire de ce token ;

(2) Du côté OFCORS, supprimer les tokens qui n'ont que le retour chariot ;

(3) Supprimer tous les espaces entre les chiffres avant la comparaison des formes de tokens ;

(4) En effet, la tokenisation différente ne conduit qu'à deux résultats : soit le token dans cupt est plus long que celui d'OFCORS, soit il est plus court que le dernier. De plus, les deux doivent retrouver une cohérence après avoir traité les tokens problématiques. De ce fait, nous concaténons d'abord le token actuel et ses tokens suivants du côté où le token est plus court jusqu'à ce qu'il ait une longueur égale ou supérieure à l'autre, pour ensuite reprendre la comparaison. De cette manière, nous trouvons les correspondances des indices pour les cas de « un à plusieurs ». Néanmoins, il existe aussi les cas de « plusieurs à plusieurs » tels que `['°C', '-']` et `['°', 'C-']`, qui sont non seulement insolubles avec un algorithme simple, mais aussi problématiques pour le choix des indices à relier. Puisque ce cas est rare, nous choisissons de laisser tomber les annotations suivantes d'OFCORS pour la phrase actuelle si les tokens sont toujours différents après la concaténation, et recommencer l'alignement à partir de la phrase suivante.

De cette manière, nous obtenons finalement les fichiers cupt+, dont chacun correspond à un texte. La table 5.3 récapitule les volumes finaux des corpus.

Corpus	Sous-corpus	Volume des corpus produits (cupt+)
ANCOR	ESLO_ANCOR	29.4 MB
	ESLO_CO2	2.89 MB
	OTG	2.10 MB
	UBS	564 KB
ER	2003_len300_1-100	4.31 MB
Sequoia	annodis.ER	891 KB
	emea	1.52 MB
	frwiki	1.74 MB
TOTAL		43.38 MB

Table 5.3 – Les volumes finaux des corpus

En conclusion, nous choisissons une partie de Sequoia dans PARSEME, 100 articles de l'Est Républicain et l'ensemble du corpus ANCOR comme notre corpus de vérification. Nous choisissons d'étendre encore le format cupt du PARSEME pour inclure les annotations sur les chaînes de coréférence. Face à la tokenisation différente, l'alignement se réalise par comparer au fur et à mesure les tokens venus de 2 systèmes, dans le cas où ils sont inégaux, nous essayons de trouver une correspondance entre 1 token d'un côté et la concaténation de plusieurs tokens de l'autre côté.

CHAPITRE 6

Résultats

Sommaire

Dans ce chapitre, nous présentons d'abord notre méthode pour classifier les croisements détectés en 4 cas différents, ensuite notre définition sur la validité des croisements et les sources d'erreurs s'il y en a. Nous montrons également les erreurs principales de Seen2Seen selon nos expériences d'annotation. Finalement, nous produisons une analyse statistique de nos annotations, dans le but de prouver quantitativement notre hypothèse.

6.1 Catégorisation des croisements

Face au nombre important de croisements détectés, nous cherchons d'abord un moyen qui peut accélérer et faciliter notre travail. En espérant pouvoir éliminer certains exemples avant la validation manuelle, nous proposons de classifier les croisements en 4 catégories différentes : EPV incluse dans une mention, segment identique, mention incluse dans une EPV et chevauchement.

— Cas 1 : EPV incluse dans une mention
Dans ce cas-là, tous les tokens de l'EPV sont inclus dans la mention et au moins un token de la mention n'est pas inclus dans l'EPV. Autrement dit, l'EPV est un sous segment propre de la mention. Voir l'exemple dans la table 6.1 [1].

— Cas 2 : segment identique
Dans ce cas-là, les tokens de l'EPV correspondent exactement aux tokens de la mention. Autrement-dit les deux segments sont égaux. Voir l'exemple dans la table 6.2.

— Cas 3 : mention incluse dans une EPV
Ici, la mention correspond seulement à une partie de l'EPV, elle est donc incluse

1. x est l'indice de l'EPV et y est l'indice de mention, ignorons ici l'indice de chaîne et la catégorie de l'EPV puisqu'ils n'influencent pas le résultat, les colonnes dans cupt+ sont transposées en lignes pour faciliter l'affichage. Idem pour les tableaux suivants.

tokens	...	la	question	posée	...
annotation EPV	...	*	x	x	...
annotation COREF	...	y	y	y	...

TABLE 6.1 – Cas 1 : EPV incluse dans mention

tokens	...	la	page	tournée	...
annotation EPV	...	x	x	x	...
annotation COREF	...	y	y	y	...

TABLE 6.2 – Cas 2 : segment identique

dans l'EPV. Autrement-dit, cette fois-ci c'est la mention qui est un sous segment de l'EPV. Voir l'exemple dans la table 6.3.

tokens	...	tourner	la	page	...
annotation EPV	...	x	x	x	...
annotation COREF	...	*	y	y	...

TABLE 6.3 – Cas 3 : mention incluse dans EPV

— Cas 4 : chevauchement
Finalement, il nous reste le cas où la mention et l'EPV ont une partie en commun et une partie propre à chacune. C'est-à-dire l'intersection de l'ensemble des tokens de la mention et celui de l'EPV est non vide, et ces deux ensembles n'ont pas une relation d'inclusion. Les tables 6.4 et 6.5 illustrent deux exemples différents.

tokens	...	faire	l'	objet	d'	autorisation	...
annotation EPV	...	x	x	x	*	*	...
annotation COREF	...	*	y	y	y	y	...

TABLE 6.4 – Cas 4a : chevauchement

Nous remarquons qu'il existe deux situations différentes : l'un est le cas du tableau 6.4, où le substantif dans l'EPV est modifié par un complément non lexicalisé dans l'expression, qui se trouve souvent après le substantif ; l'autre est montré par le tableau 6.5, où le substantif dans l'EP est flexible en nombre donc son déterminant n'est pas annoté dans l'expression, mais les déterminants sont toujours inclus dans la mention.

Nous constatons que certains substantifs dans l'expression peuvent être détectés comme plusieurs mentions dans différentes chaînes. Par exemple dans « …a fait tout d'abord une rétrospective des travaux qui ont été accomplis … », deux mentions « une rétrospective des travaux » et « des travaux » sont découvertes par OFCORS et elles se trouvent dans différentes chaînes. Et l'annotation des cas de croisement est réalisée sur chacune de ces deux mentions. Nous pouvons donc disposer de plusieurs cas différents pour une seule EPV identifiée dans le texte. Dans ce cas là, nous préférons les mentions qui sont plus longues, concrètement, l'ordre de choix est cas 1 > cas 2> cas 4> cas 3.

Puisque notre hypothèse porte sur l'accessibilité des substantifs à l'intérieur de l'EP à la coréférence, les croisements du premier et du deuxième cas sont évidemment à exclure de la discussion. De plus, étant donné que le quatrième cas arrive souvent pour les EP plus compositionnelles ou même discutables d'être EP, le troisième cas permettra de mieux

tokens	...	se	faire	des	soucis	...
annotation EPV	...	x	x	*	x	...
annotation COREF	...	*	*	y	y	...

Table 6.5 – Cas 4b : chevauchement

tester notre hypothèse, mais nous utilisons tous les exemples dans la suite pour réaliser les annotations humaines.

6.2 Validité des croisements

Afin d'analyser le résultat, nous effectuons une annotation sur la validation et les sources d'erreur des exemples. Il est à noter que les exemples examinés n'incluent que les croisements détectés, et que l'annotation est réalisée principalement par 2 annotateurs et la validité des exemples est décidée après leur discussion.

6.2.1 Annotation de la validation des croisements

Parmi ces croisements détectés automatiquement, il y a des exemples qui ne nous intéressent pas, et des croisements formés suite à des erreurs des systèmes. De ce fait, il nous faut vérifier manuellement la validité de chaque exemple et l'annoter en choisissant parmi les 5 valeurs suivantes : « vrai », « faux », « non concerné », « discutable » et « répétitions ».

Si la validation est annotée « vrai », l'exemple annoté est utilisable pour valider notre hypothèse, plus précisément l'expression trouvée est vraie, et la mention croisée est dans une vraie chaîne de coréférence selon l'interprétation humaine [2]. Il est à noter que même si de petites erreurs sont commises dans certains exemples, nous annotons toujours « vrai » à condition que ces erreurs n'influencent pas l'utilité de l'exemple selon un jugement humain. Par exemple, dans l'exemple valable « … une rétrospective des **travaux** qui ont été **accomplis** dans la commune. », le système détecte « une rétrospective des travaux » comme la vraie mention, tandis que la mention vraiment coréférente avec « qui » est le nom « travaux ».

Pour les exemples restants, nous effectuons une analyse simple pendant l'annotation de la validation. Si l'exemple est annoté « faux », le croisement détecté provient des erreurs des outils, donc ce n'est pas un vrai croisement. S'il est annoté « non concerné », c'est peut-être un vrai croisement mais il n'aide pas à valider notre hypothèse, cela arrive souvent pour le cas 1, soit l'EP incluse dans la mention. Dans le corpus oral ANCOR, nous constatons beaucoup de chaînes de coréférence formées à cause des disfluences et des réutilisations des expressions complètes. Par exemple, les deux « un détour » sont annotés coréférentes dans la phrase :

« je ferais un détour pour aller à Chambord je ferais un détour vous voyez ça c' est c' est c' est c' est sans doute dans mon tempérament vous comprenez ça c' est c' est c' est personnel ça »

Le 2e « un détour » est anaphorique indiscutablement, mais au lieu d'être utilisé dans un autre contexte, il est toujours utilisé dans la même expression « faire un détour ». Par conséquent, c'est plutôt une répétition de l'expression qui ne réfute pas notre hypothèse. Nous annotons « répétitions » pour la validation de ce type. Finalement, la validation est

2. À noter que du côté de mention, nous annotons « vrai » quand la mention dans la chaîne de coréférence a au moins une mention vraiment coréférente dans cette chaîne, donc le « vrai » ne signifie pas que la chaîne soit 100% correcte.

annotée « discutable » pour les exemples qui peuvent être vrais ou faux selon l'interprétation humaine. Prenons la phrase suivante :

« - Créé par la Fédération nationale qui perpétue le souvenir de l'homme d'Etat 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. »

Nous pouvons dire qu'il existe une coréférence entre « son nom » et « André-Maginot », mais cette coréférence est également mise en question puisqu'on peut considérer que le dernier est le nom du prix mais pas le nom de la personne.

6.2.2 Annotation des sources d'erreur

Les erreurs du système proviennent de deux modules : l'expression polylexicale ou la chaîne de coréférence. Les sources d'erreur sont annotées en même temps s'il y en a. Elles sont classifiées en les cinq types suivants :

- `MWE incorrecte` : malgré l'identité des lemmes, l'expression détectée n'est pas une vraie expression polylexicale.
- `MWE littérale` : c'est un sous-cas de « MWE incorrecte », dans ce cas-là, l'expression détectée demande une lecture littérale dans le contexte.
- `MWE type incorrect` : l'expression détectée est correcte mais elle est classée dans une catégorie incorrecte.
- `chaîne incorrecte` : Aucune des mentions de la chaîne n'est coréférente avec la mention dans l'expression.
- `mention incorrecte` : La mention utilisée dans la chaîne est incorrecte, mais la chaîne serait correcte si la mention contenait plus ou moins de mots. Cette annotation arrive souvent lorsque l'erreur n'influence pas la validation de l'exemple, comme l'exemple précédent « une rétrospective des travaux » et « qui ».

6.2.3 Analyse des erreurs commises par Seen2Seen

L'annotation des EPV dans le corpus ANCOR est réalisée par Seen2Seen. Nous constatons certaines erreurs courantes au cours de notre validation des exemples.

Premièrement, les erreurs résultent de la difficulté à traiter la discontinuité. À cause de l'absence de ponctuation, les paroles extrêmement longues ne sont pas découpées en plusieurs phrases. De plus, le filtrage optimisé de Seen2Seen a désactivé la limite de distance. En conséquence, nous remarquons de nombreuses EP faussement détectées par Seen2Seen dont les composants se trouvent dans des syntagmes vraiment éloignés, même dans des phrases différentes selon une segmentation humaine.

Deuxièmement, les erreurs proviennent d'une restriction insuffisante des pronoms. Pour les EPV (souvent les idiomes verbaux (VID)) contenant un pronom impersonnel, le pronom « il » doit être inchangeable. Mais Seen2Seen identifie aussi l'expression avec d'autres pronoms qui possèdent le même lemme. Par exemple, nous remarquons l'expression « elle y a » est annotée à cause du VID « il y a ».

Puisque PARSEME n'annote que les composants lexicalisés des EPV, sauf les pronoms dans les IRV, les autres pronoms annotés doivent a priori être figés. Nous vérifions cette supposition dans le corpus français du PARSEME (voir la table 6.6). Parmi toutes les EPV distinctes [3], seulement 388 EPV disposent d'au moins un pronom. Si nous nous limitons aux EPV qui ont un nombre d'occurrences supérieur à un afin d'effectuer une comparaison des formes de pronoms utilisées, les EPV conformes ne seront que les VID

3. Deux EPV sont considérées comme distinctes si elles possèdent différents multisets de lemmes.

et les IRV, et le nombre total diminue à 205, dont 45 VID et 160 IRV. Parmi ces VID, 30 expressions utilisent toujours les mêmes formes de pronoms, et les différences apparues dans les 15 restants sont dues à l'élision, l'adhérence du tiret au pronom « il » dans l'ordre inversé (« il » vs « -il »), l'omission incorrecte des diacritiques (« ça » vs « ca ») et les pronoms réfléchis compris dans le VID. De plus, « il » est toujours utilisé comme pronom impersonnel, il est donc inchangeable. La table 6.7 montre les ratios des VID de différents degré d'inflexibilité du pronom par rapport à ces 45 VID. [4].

Type	VID	IRV	LVC.full	Total
Pas de contrainte sur le nombre d'occurrence	282	103	3	388
Nombre d'occurrence supérieur à 1	45	160	0	205

Table 6.6 – Bilan des nombres des EPV ayant au moins 1 pronom

Degré	Nombre de VID	Ratio
1	30	67%
>=0.9	32	71%
<1 & >=0.7	9	20%
<=0.5	6	13%

Table 6.7 – Ratio des VID ayant différents degrés d'inflexibilité du pronom

Finalement, une partie des erreurs provient de contraintes insuffisantes de l'ordre des composants. Dans la réalité, nous trouvons que la variabilité des ordres arrive souvent dans les LVC, et certaines VID sont trop figés si bien qu'un simple changement de l'ordre des composants exclut son statut de VID. Par exemple, la combinaison « fait ça » est trouvée à cause du VID « ça fait », mais ce n'est pas du tout une EPV.

Logiquement, l'ordre des composants dans un VID est moins flexible que celui des LVC, et nous vérifions aussi cette différence de flexibilité de manière statistique. Parmi toutes les EPV distinctes, seulement 696 EPV ont un nombre d'occurrences supérieur à 1, et 83.6% d'entre eux sont toujours dans le même ordre. La table 6.8 montre le taux des EPV ayant un ordre fixe :

Ce tableau relève clairement que sauf les LVC.full, toutes les autres catégories demandent une inflexibilité beaucoup plus élevée sur l'ordre des composants, mais de nouveau, à cause de manque de données, le haut ratio des LVC.cause n'est pas fiable.

6.3 Analyse des résultats

Après tous ces traitements, nous pouvons effectuer une analyse sur les nombres de croisements et leur distribution selon les 4 cas ou les 5 catégories, ainsi qu'une synthèse des sources d'erreurs.

4. La manière pour calculer le degré de l'inflexibilité des pronoms : pour chaque EPV, le nombre d'occurrences de la forme la plus fréquente d'une EPV / nombre de toutes les occurrences de cette EPV

Type	Nombre d'EPV dont son nombre d'occurrences est supérieur à 1	Nombre d'EPV suivant le même ordre	Ratio
LVC.full	314	217	69.10%
VID	210	196	93.30%
IRV	160	157	98.10%
LVC.cause	8	8	100%
MVC	4	4	100%
Total	696	582	83.62%

TABLE 6.8 – Taux de l'ordre fixe des composants dans EPV par catégorie

6.3.1 Analyse des croisements vrais par rapport aux catégories

Premièrement, nous effectuons une comparaison entre le nombre des EPV annotées/-détectées, le nombre des croisements détectés et celui des croisements validés.

La figure 6.1 illustre un bilan de l'état global par catégorie. La proportion des croisements vrais et faux est montrée sous forme de barres. Les chiffres sur les barres gris foncé indiquent les nombres de croisements détectés. De plus, étant donné que nous voulons analyser la proportion de vrais croisements dans les EPV, il faut que ce nombre de EPV soit correcte. Nous choisissons d'éliminer les croisements faux des EPV fausses lors de notre décompte des nombres de croisements détectés, donc, pour ce qui reste, la seule raison pour laquelle on a une validation fausse c'est que le croisement avec une chaîne de coréférence est plutôt impossible pour l'EPV examinée. [5].

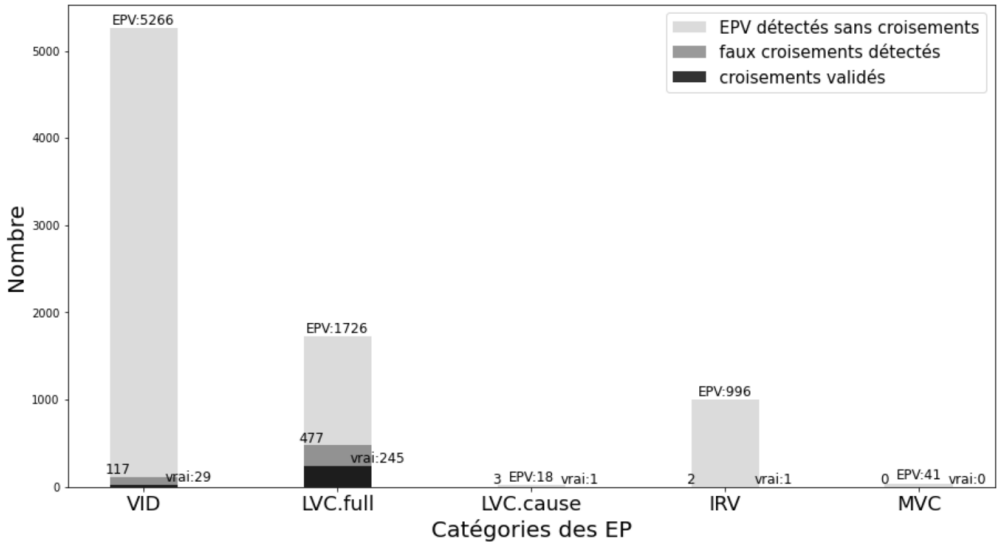


FIGURE 6.1 – Comparaisons des nombres d'EP détectées, de croisements détectés et de vrais croisements

Dans la figure, nous remarquons immédiatement que les vrais croisements des EPV et les mentions dans des chaînes sont vraiment rares. Il est à noter que les MVC ne contiennent

5. ci-après, si nous n'expliquons pas spécifiquement, le nombre de croisements détectés signifiera par défaut ce nombre sans EPV erronées

pas de substantif puisqu'ils sont composés des verbes, et les pronoms réfléchis ne sont pas annotés dans le corpus ANCOR, donc OFCORS n'est pas capable de les trouver à priori. Les deux pronoms détectés ici le sont à cause de leur forme figée comme le cas de « nous ». En conséquence, les croisements détectés pour MVC et IRV sont nuls ou presque et nous pouvons nous en passer dans l'analyse suivante.

Parmi les 3 catégories qui restent, LVC.cause n'a pas non plus un nombre élevé de croisements. Néanmoins, il n'y a que 18 LVC.cause au total, donc les conclusions pour cette catégorie peuvent difficilement être considérées fiables.

Nous concentrons notre analyse sur les VID et LVC.full. Il est évident que LVC.full a un nombre absolu de croisements validés beaucoup plus élevé que celui de VID, même si le nombre total des VID est extrêmement large, soit deux fois plus élevé que celui de LVC.full. Néanmoins, puisque Seen2Seen privilégie le rappel, il y a beaucoup de bruits dans les EPV détectés. De ce fait, afin d'avoir des nombres fiables, nous profitons du corpus Sequoia, dans lequel les EPV sont manuellement annotées.

Nous constatons dans la figure 6.2 que les LVC.full sont plus nombreux que les VID. De plus, sans les EPV faussement détectées, le nombre de croisements détectés dans les figures (le chiffre de « croisements détectés » sur barres gris foncé) est en fait le nombre des EPV correctes parmi toutes les EPV examinées par nous. Et nous voyons également que le nombre des vrais LVC.full dans les croisements est plus élevé que celui des vrais VID dans tous les corpus. Finalement, s'agissant des vrais croisements, ceux de LVC.full sont toujours plus nombreux que ceux de VID.

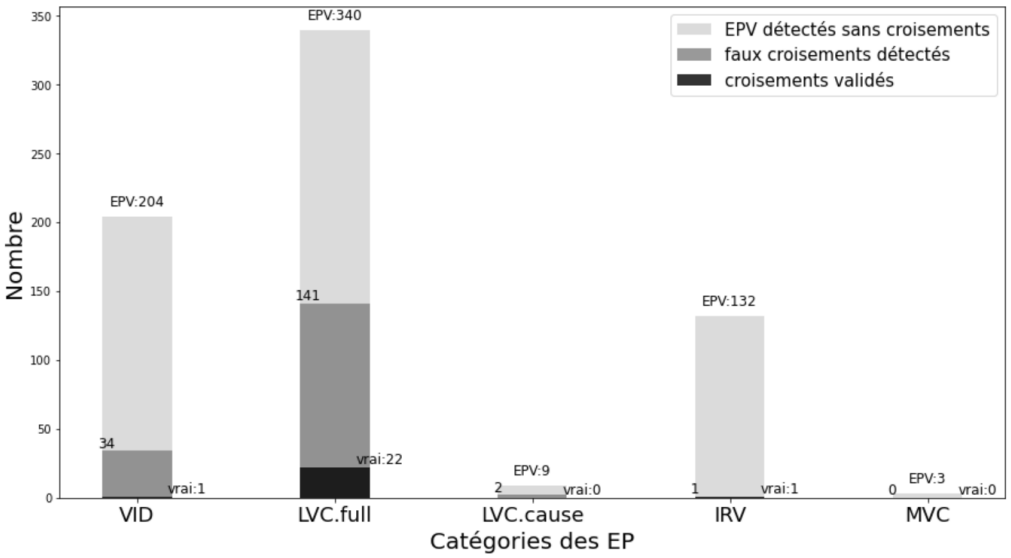


Figure 6.2 – Comparaisons des nombres d'EP détectées, de croisements détectés et de vrais croisements (Sequoia)

Par ailleurs, bien que les LVC.full soient plus nombreux que les VID à la fois pour le nombre des EPV et pour le nombre de vrais croisements, nous remarquons que la proportion de vrais croisements parmi les LVC.full est plus élevée. Ce fait est relevé par la figure 6.5 [6]. Nous utilisons deux critères différents pour calculer le ratio des croisements vrais par

6. Les EPV détectées pour le corpus total est le nombre total des EPV identifiées par le système, quelle que soit l'existence d'un croisement détecté ; les EPV annotées pour le corpus Sequoia est le vrai nombre total d'EPV dans ce corpus ; les EPV vérifiées sont les EPV correctes qui se trouvent dans les croisements détectés.

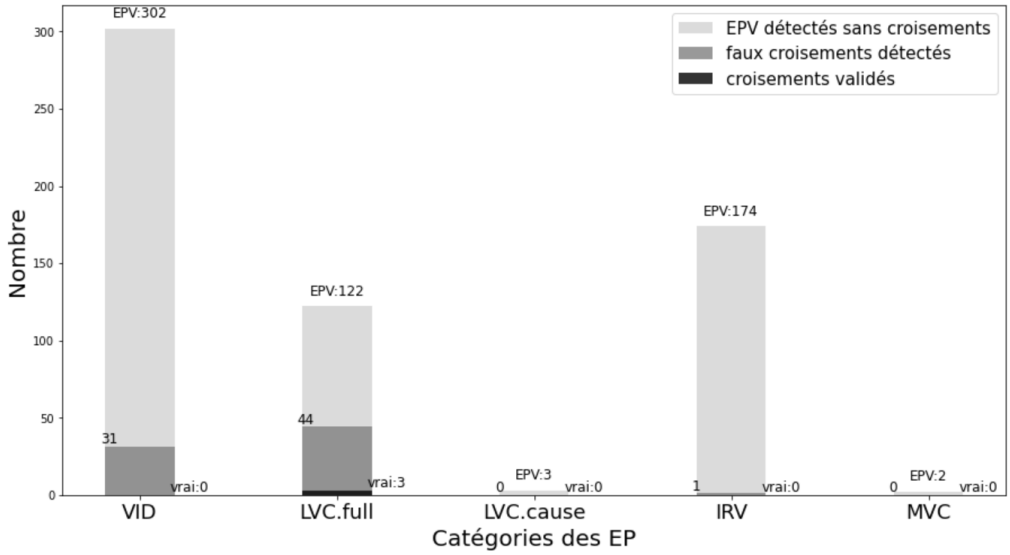


FIGURE 6.3 – Comparaisons des nombres d'EP détectées, de croisements détectés et de vrais croisements (Est Républicain)

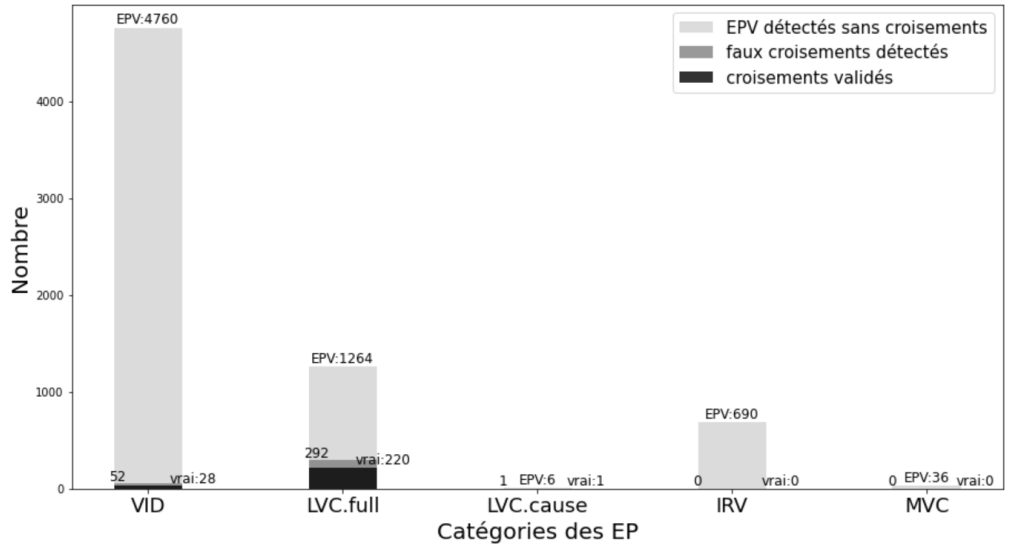


FIGURE 6.4 – Comparaisons des nombres d'EP détectées, de croisements détectés et de vrais croisements (ANCOR)

rapport aux différents corpus : 1) le nombre des EPV vérifiées [7] pour chaque sous-corpus et le corpus global (ANCOR + ER + Sequoia), 2) le nombre des EPV annotés/détectés pour le corpus Séquoia et le corpus global. Il est à noter que le dernier sur le corpus global est probablement biaisé par le bruit de l'annotation automatique des EPV. Nous constatons que le ratio des croisements vrais est toujours élevé dans la catégorie LVC, quel que soit le corpus et la méthode de calcul.

Ces chiffres sont conformes à la différence de ces deux catégories sur le degré de compositionnalité : les LVC.full sont plus compositionnels que les VID. Selon une hypothèse

7. Comme ce qui est expliqué avant, c'est en fait le nombre des croisements détectés par le systèmes en supprimant ceux avec une EPV erronée (« MWE incorrecte », « MWE littérale » et « MWE type incorrect »)

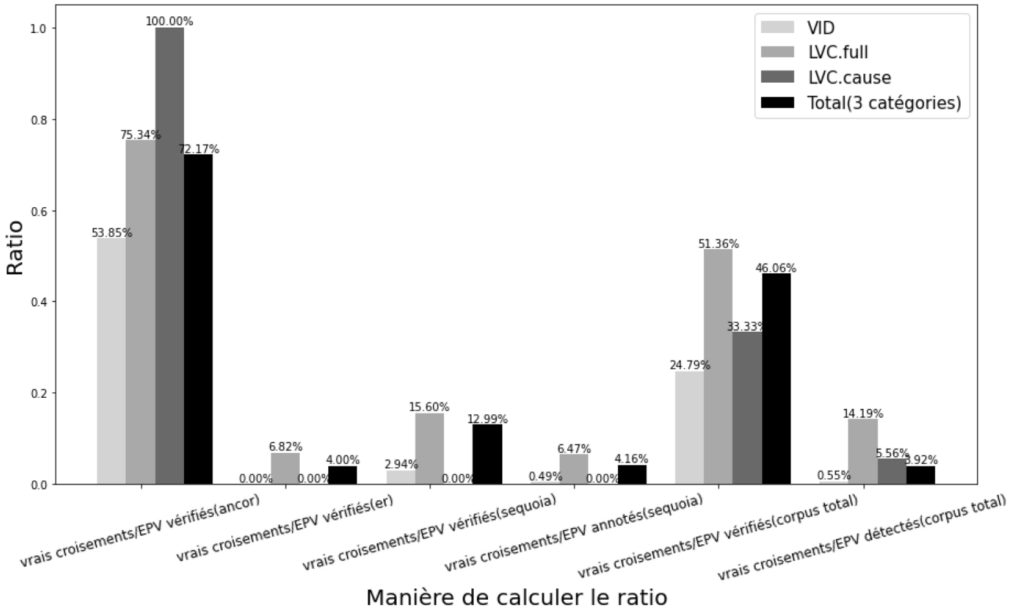


Figure 6.5 – Ratio de croisements vrais

précédemment mentionnée, les EP plus compositionnelles ont tendance à être plus présentes dans les corpus, et les LVC.full doivent être plus nombreux. En plus, notre hypothèse à vérifier se base sur la caractéristique de non-compositionnalité des EP, donc la catégorie VID a statistiquement moins de chances d'avoir de vrais croisements.

6.3.2 Analyse de l'influence du type de corpus

Deuxièmement, il faut aussi faire attention à l'influence du type de corpus sur le résultat. Les différences entre les diagrammes sur les sous-corpus (Figure 6.2, Figure 6.3 et Figure 6.4) peuvent servir d'indicateurs des biais potentiels. Étant donné que le corpus ANCOR est beaucoup plus grand que les deux corpus écrits, l'analyse doit être réalisées sur des ratios ou des comparaisons relatives, au lieu de nombres absolus.

Dans un premier temps, nous remarquons qu'au contraire de l'annotation manuelle dans Sequoia, le nombre des VID sont beaucoup plus élevés que celui des LVC.full selon l'annotation de Seen2Seen, et cet écart est encore plus grand dans le corpus ANCOR [8]. Cela peut s'expliquer par les deux raisons suivantes :

1) Seen2Seen privilégie plus le rappel pour les VID que pour les LVC.full, donc la précision de VID est moins bonne que les LVC.full, et le score de précision est encore pire dans les corpus oraux.

2) Nous utilisons plus de VID à l'oral qu'à l'écrit.

En fait, il y a des arguments à l'appui de ces deux suppositions.

D'abord, selon l'évaluation du système Seen2Seen, la précision de VID est un petit peu moins élevée que LVC.full, soit 0.8503 contre 0.8675, mais son rappel est beaucoup plus élevé que le rappel de LVC.full, soit 0.7762 contre 0.5988. De plus, la table 6.9 montre la précision du système sur les EPV examinées [9]. Dans les deux corpus, la précision des VID est beaucoup moins bonne que celle des LVC.full, et celle de VID dans ANCOR est

8. nombre de VID détectés / nombre de LVC.full détectés est 2.5 pour Est Républicain et 3.8 pour ANCOR

9. Étant donné qu'il n'existe pas de croisement détecté pour la catégorie MVC, nous l'ignorons ici.

extrêmement faible, même inférieure à 0.1. Notre supposition sur la précision est bien démontrée avec cet échantillon. Par conséquent, nous imaginons qu'il y a beaucoup de bruit parmi les VID détectés dans ANCOR.

corpus	**VID**	**LVC.full**	**LVC.cause**	**IRV**	**Total**	**macro-moyenne**
ER	0.63	0.98	0	1	0.79	0.65
ANCOR	0.09	0.64	1	0	0.33	0.43
ER+ANCOR	0.13	0.67	0.5	0.33	0.37	0.41

TABLE 6.9 – Précisions du Seen2Seen sur les EPV des croisements détectés

Ensuite, en examinant de plus près la liste des VID et des LVC.full apparus dans les vrais croisements, nous remarquons que malgré un nombre de croisements vrais relativement élevé, 28 dans ANCOR contre 1 dans Sequoia et ER, seulement 9 VID différents sont en fait concernés, et ce sont des expressions fréquemment utilisées dans la vie réelle, telles que « **avoir le temps** », « **poser problème** » (voir dans l'annexe B la liste complète). Nous pouvons imaginer que certaines expressions extrêmement courantes à l'oral sont catégorisées dans les VID, et leur forte présence augmente le nombre total de VID à l'oral. En revanche, les VID présents à l'écrit ont tendance à être plus figurés et donc moins compositionnels : parmi les croisements détectés selon le système, 65 VID sont vérifiés et sont corrects, mais un seul est un vrai croisement (c'est l'expression « **porter** le **nom** »). Ces mêmes proportions se retrouvent également dans d'autres catégories, ce qui est évident dans la Figure 6.5 : les barres du corpus ANCOR sont beaucoup plus hautes que les barres de ER et de Sequoia (le troisième groupe de barres). Nous supposons que la situation est identique pour les LVC.

6.3.3 Analyse sur les 4 cas

Finalement, nous présentons l'analyse des cas de croisements détectés selon notre définition, par rapport aux types d'EPV et à la validation des croisements détectés. Pour rappel, le cas 1 correspond aux EPV incluses dans les mentions ; le cas 2 lorsqu'elles ont un segment identique, le cas 3 lorsque la mention est incluse dans l'EPV et le cas 4 lorsqu'il y a chevauchement.

La Figure 6.6 montre le nombre d'occurrences pour chaque cas par type d'EP. Le cas 2 n'est jamais détecté parmi les croisements et le cas 1 est également rare. Pour les VID, le cas 3 est plus fréquent que le cas 4, mais c'est l'inverse pour les LVC et le cas 4 est remarquablement plus élevé dans le LVC.full. Nous constatons la même tendance pour les croisements vrais (voir la Figure 6.7). Cela pourrait résulter de deux faits : 1) les LVC.full sont plus compositionnels, donc leurs substantifs ont une plus grande probabilité d'avoir des compléments non lexicalisés dans l'EPV, mais OFCORS traite le nom et ses compléments comme une seule unité ; 2) les différents choix pour traiter les déterminants augmente aussi beaucoup le nombre de cas 4.

Nous examinons finalement la distribution des valeurs de validation pour les cas 1, 3 et 4 (voir la figure 6.8) [10]. Pour le cas 1, les 3 « vrai » sont en fait des mentions longues contenant un EPV dans son complément comme un modifieur et ce sont tous LVC.full. Par exemple, dans « patients **présentant** des **symptômes** après l'injection », le substantif « **symptômes** » peut être utilisé à l'extérieur de l'expression. De ce fait, au contraire de notre supposition au départ, le cas 1 pourrait nous intéresser dans certains

10. Le cas 2 est ignoré ici puisqu'il n'y a pas de données.

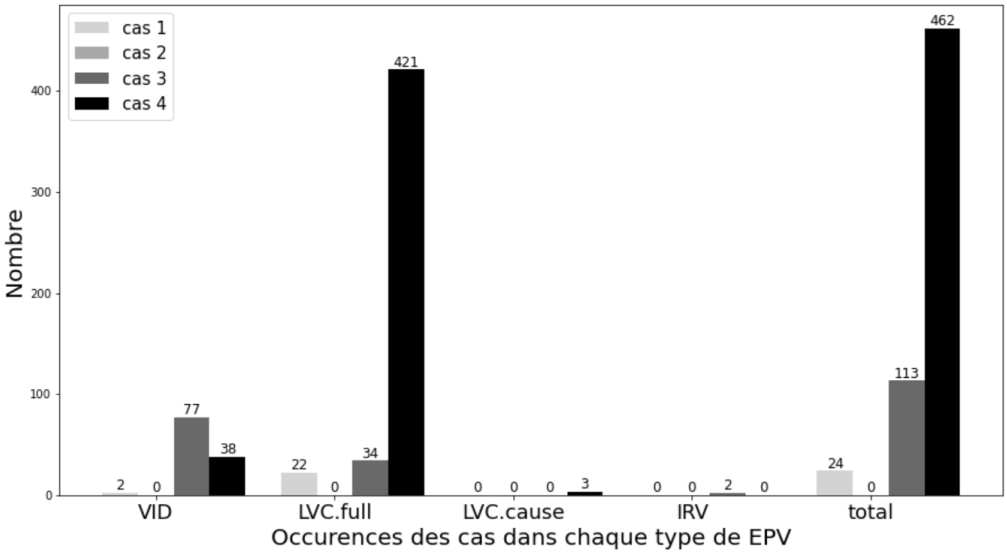


Figure 6.6 – Nombre de cas par type EPV (croisements détectés)

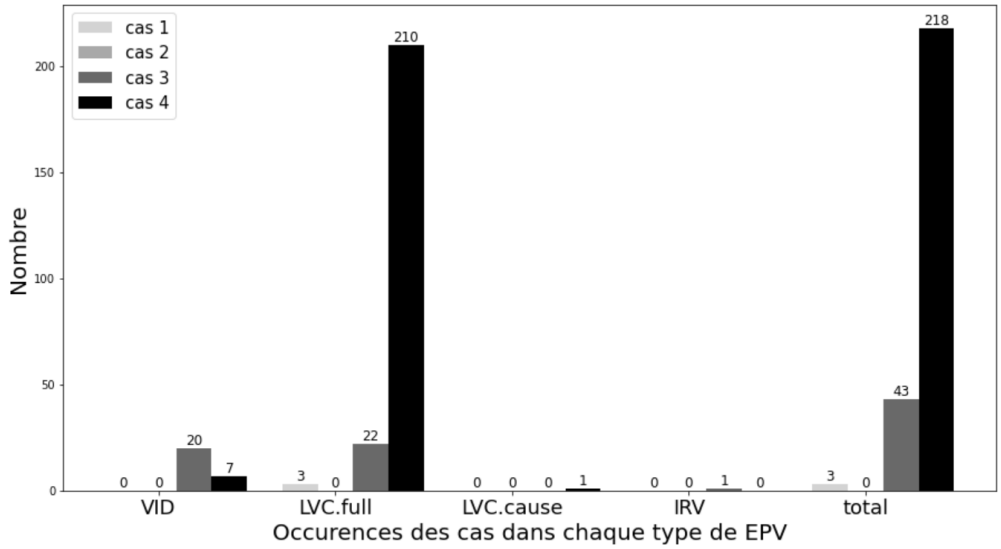


Figure 6.7 – Nombre de cas par type EPV (croisements vrais)

contextes. Nous remarquons beaucoup de répétitions pour les cas 3 et 4, qui viennent plutôt du corpus ANCOR. Un dernier point intéressant : la proportion des « vrai » dans le cas 4 constitue presque la moitié des croisements, ce qui est beaucoup plus fréquent que son pourcentage pour d'autres cas. Puisque les croisements de cas 4 sont essentiellement des LVC.full, qui sont plutôt compositionnels, nous supposons que les expressions plus compositionnelles ont une plus grande probabilité d'être composées des substantifs qui sont plus accessibles à l'extérieur de l'expression et aussi plus susceptibles d'avoir des compléments flexibles.

En conclusion, les croisements peuvent être classifiés en 4 cas différents selon l'appartenance des tokens : EPV incluse dans mention, identique, mention incluse dans EPV et chevauchement des deux. Selon la validité que nous avons définie, les croisements peuvent aussi être divisés en 4 catégorie : « vrai », « faux », « discutable » et « répétitions ». Nous

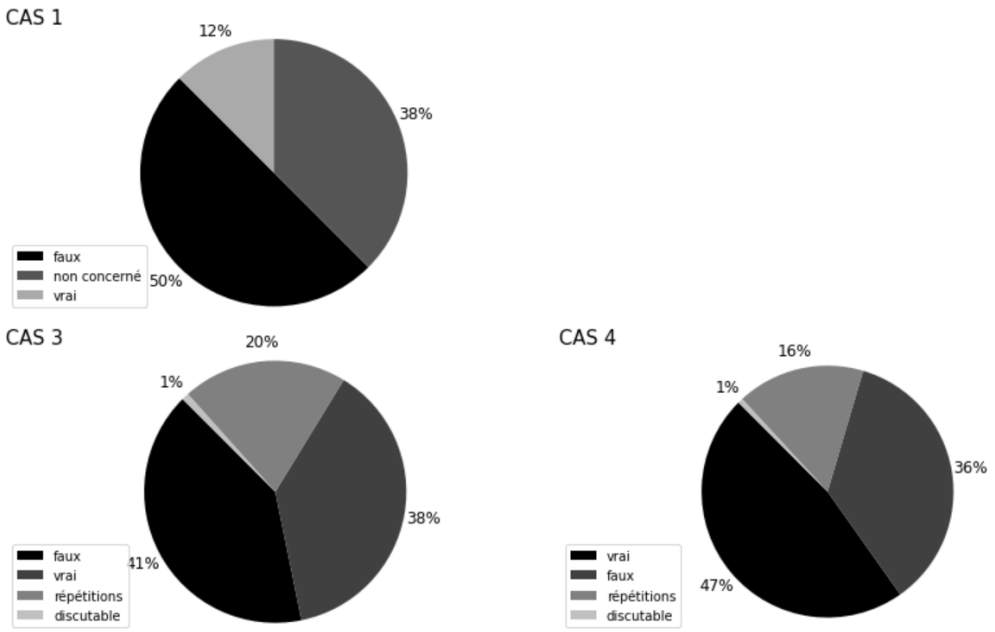


FIGURE 6.8 – Les validations dans chaque cas

trouvons que le bruit de Seen2Seen viennent essentiellement du contrôle insuffisant sur la forme du pronom et sur la discontinuité de l'EPV. Finalement, selon les statistiques obtenues, nous remarquons que les composants dans un LVC.full ont beaucoup plus de chances d'être repris dans une chaîne de coréférence par rapport à ceux dans un VID, et cela est conforme à leur différent niveau de compositionnalité. Du point de vue des différents cas, le 4e cas représente une proportion majoritaire parmi les croisements vrais.

CHAPITRE 7

Analyse de non-compositionnalité d'EPV

Sommaire

Dans les chapitres ci-dessus, en supposant que les LVC soient plus compositionnels que les VID, notre hypothèse est quasi prouvée de manière générale, mais nous n'avons pas testé le croisement avec le degré de compositionnalité de chaque EPV. Dans ce chapitre, nous présentons notre proposition pour définir une méthode pour mesurer le degré de compositionnalité d'un EPV, le résultat de l'annotation et les problèmes rencontrés pendant l'annotation, ainsi que la comparaison de cette annotation manuelle avec les statistiques obtenues auparavant.

7.1 Annotation de compositionnalité

L'étude de compositionnalité est une problématique qui intéresse beaucoup de chercheurs, mais il est difficile de trouver une manière pour la mesurer sans ambiguïté.

[Caillies, 2009] a effectué une étude pour décrire les expressions idiomatiques en français sur différentes dimensions, dont un élément est la « décomposabilité ». Cette caractéristique « réfère au degré avec lequel la signification des mots de l'expression contribue à la signification figurée globale », ce qui est donc similaire à notre appréhension de la compositionnalité. Pour chaque expression, ce degré a été jugé par 20 adultes, qui sont tous locuteurs natifs de la langue française et possèdent au moins le baccalauréat. On demande aux annotateurs de déterminer ce degré de « décomposabilité » sémantique de l'expression sur une grille de 1 à 6, 1 pour les expressions non décomposables comme « voir midi à sa porte » et 6 pour les idiomes très décomposables tels que « cacher son jeu ».

Dans les études plus récentes, afin d'évaluer leur cadre de prédiction sur la compositionnalité des mots composés nominaux, [Cordeiro et al., 2019] ont préparé une référence des scores humains pour ces expressions à prédire, qui sont sélectionnées manuellement dans les dictionnaires et ont l'une des structures suivantes : « nom + nom », « nom + adjectif » et « adjectif + nom ». Le degré de compositionnalité est défini par trois niveaux : 1) totalement compositionnel (les 2 composants dans l'EP sont toutes littérales), 2) partiellement compositionnel (un composant est utilisé avec son sens littéral mais l'autre non), 3) idiomatique (la signification globale n'a rien à voir avec les significations des composants). L'annotation de ce niveau est réalisée par des natifs non spécialisés en linguistique. On leur demande de donner un score entre 0 (idiomatique) et 5 (littéral) pour le nom principal, le modifieur et l'expression globale respectivement pour chaque mot composé. Pour

chaque expression, la moyenne arithmétique des scores sur l'expression globale est utilisée finalement comme score de référence.

Nous essayons d'élaborer un guide d'annotation pour annoter le degré de compositionnalité des EPV en nous basant sur le questionnaire de [Cordeiro et al., 2019]. La plus grande différence dans notre cas, qui est également la difficulté majeure, est le fait que la structure d'une EPV est beaucoup plus flexible que celle d'un mot composé nominal. Afin d'adapter ce questionnaire aux EPV, nous demandons aux annotateurs de donner un score entre 0 et 5 pour 6 questions, qui contiennent non seulement l'évaluation de contribution de chaque composant (verbe, ses dépendants, modifieurs éventuels de dépendants) au sens global de l'expression, mais aussi des tests structurels et grammaticaux, soit un test sur la flexibilité de la structure de l'expression, et un test de l'apport de la tête verbale sur le rôle sémantique du sujet et sur des informations morphologiques (voir le guide complet dans l'annexe F). Nous posons successivement quatre questions générales : les trois premières se concentrent sur les composants dans l'expression et la quatrième demande une idée générale sur l'expression évaluée. La question sur la tête verbale et celle sur l'expression globale ont deux sous-questions respectivement, et différents poids sont donnés à ces sous parties. Le score final de chaque expression est calculé de deux manières différentes : 1) nous moyennons d'abord les scores de différents annotateurs pour chaque question. Par la suite, nous calculons la moyenne des questions sur les composants pour obtenir un score des composants. Le score final de degré de compositionnalité pour une expression est la moyenne entre ce score de composants et le score de la quatrième question, ce score est utilisé pour décider le « degré moyenné », qui est aussi le degré associé à chaque croisement vrai concret ; 2) nous utilisons directement la moyenne des scores des annotateurs sur la 4e question pour décider le degré de compositionnalité, le degré obtenu est appelé « degré Q4 » ci-après. Pour chacun de ces deux scores, nous pensons que le degré de compositionnalité est fort si l'EPV ayant un score supérieur à 4, et faible si c'est inférieur à 2.

7.2 Résultat des annotations manuelles

L'annotation est réalisée par deux étudiants et deux chercheurs qui connaissent déjà assez bien les EPV. De ce fait, afin de diminuer le biais des informations externes sur la décision des annotateurs, nous avons dissimulé l'information sur l'origine de l'EPV – le corpus et le type de l'EPV dans la fiche d'annotation. De plus, afin de pouvoir effectuer une comparaison par la suite, nous évaluons finalement 119 expressions différentes, dont 93 ont au moins un croisement vrai et les 26 autres qui n'en ont pas[1]. Ces EPV sans croisements vrais sont aléatoirement sélectionnées du corpus Sequoia et ce sont des EPV du type VID, LVC.full ou LVC.cause. Chaque expression venue de croisements vrais est annotée par quatre annotateurs et les autres par trois annotateurs. Voici le résultat final de cette phase d'annotation.

Nous remarquons que la plupart des EPV examinées ont un degré de compositionnalité « moyen », ce qui est logique puisque la plupart parmi elles sont des LVC à cause de leur forte présence dans les croisements vrais. Mais nous constatons aussi que les EPV présentes dans les croisements vrais ont tendance à avoir un degré de compositionnalité plus élevé :

1. C'est le nombre après correction selon l'annotation, le nombre initial des EPV automatiquement sélectionées est 95 pour les EPV ayant croisements vrai et 25 pour les EPV sans croisement vrai. Nous effectuons aussi un traitement supplémentaire : si selon l'annotateur l'expression trouvée n'est pas vraiment une EPV, nous mettons pour toutes les questions un score de 5, puisque les non EPV devraient être totalement compositionnelles.

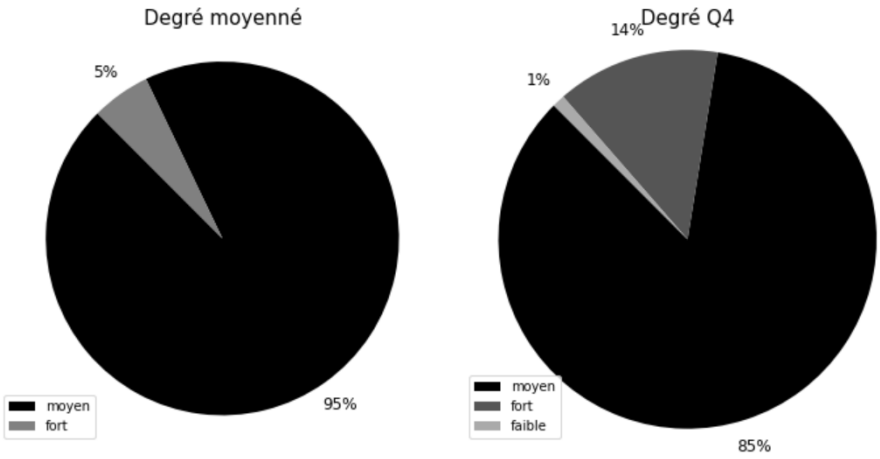


Figure 7.1 – Degré de compositionnalité pour les expressions ayant un croisement vrai

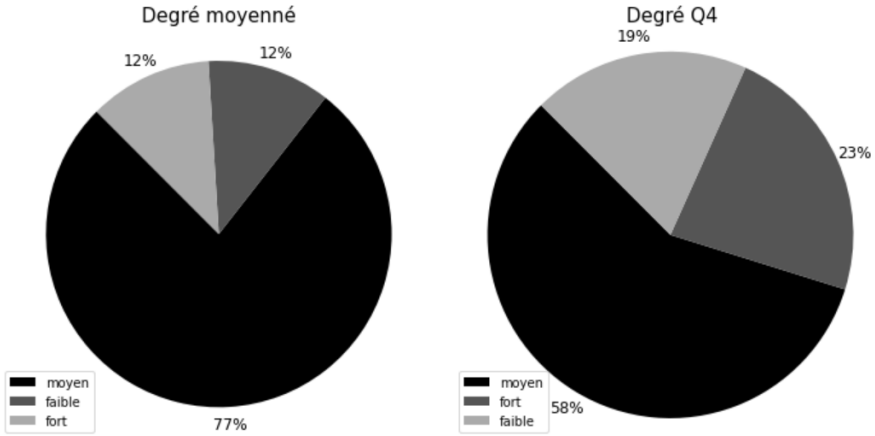


Figure 7.2 – Degré de compositionnalité pour les expressions sans croisement vrai

il n'y a pas de niveau faible selon « degré moyenné » et une proportion importante (14%) parmi elles a même un niveau fort selon « degré Q4 » (voir la Figure 7.1). En revanche, nous remarquons un pourcentage beaucoup plus important d'EPV du niveau « faible » pour les expressions sans croisement vrai (voir la Figure 7.2). Étant donné que ces expressions sont choisies au hasard parmi les expressions annotées dans tout le corpus Sequoia, qui ne contient pas seulement les 3 sous-corpus (frwiki, emea et annodis.ER) utilisés dans notre vérification de l'hypothèse, il est possible que les EPV annotées ici n'apparaissent pas dans notre corpus de travail, donc il est raisonnable que le niveau fort soit assez représenté parmi les EPV sans croisement.

7.3 Fiabilité de l'annotation

Même si nous essayons de préciser le plus possible les règles à suivre pour déterminer le score de chaque question, nous remarquons toujours certains problèmes au cours de l'annotation.

Premièrement, la légitimité de certaines EPV est remise en question. Ces expressions sont peut-être correctes avec leurs voisins dans le corpus PARSEME, mais quand nous les examinons de manière générale en fournissant d'autres contextes, ce n'est pas toujours le

cas. En fait, le jugement sur le statut EPV d'une combinaison de mots varie beaucoup en fonction de personnes. Cette incertitude existe aussi pendant la validation des croisements détectés dans le corpus ANCOR et ER.

Deuxièmement, nous constatons que les réponses aux questions données par différents annotateurs sont très hétérogènes, notamment pour la contribution du sens des composants à la signification globale. Ces réponses sont fortement influencées par les connaissances linguistiques ou même la langue maternelle de l'annotateur. Par exemple, un annotateur non natif de la langue française pourrait être influencé par la manière de penser dans sa langue maternelle, puisqu'il a tendance de suivre une démarche de décision en traduisant mot à mot l'expression en sa langue maternelle et vérifiant ensuite si la traduction de chaque composant contribue au sens global de l'expression. Pour la même expression « avoir influence », l'annotateur natif pense que le verbe « avoir » ne porte aucun sens, mais pour un annotateur dont la langue maternelle est le chinois, puisque le verbe dans la traduction littérale de l'expression contribue au sens global en chinois, il pense que ce verbe contribue quand même un peu au sens global aussi en français.

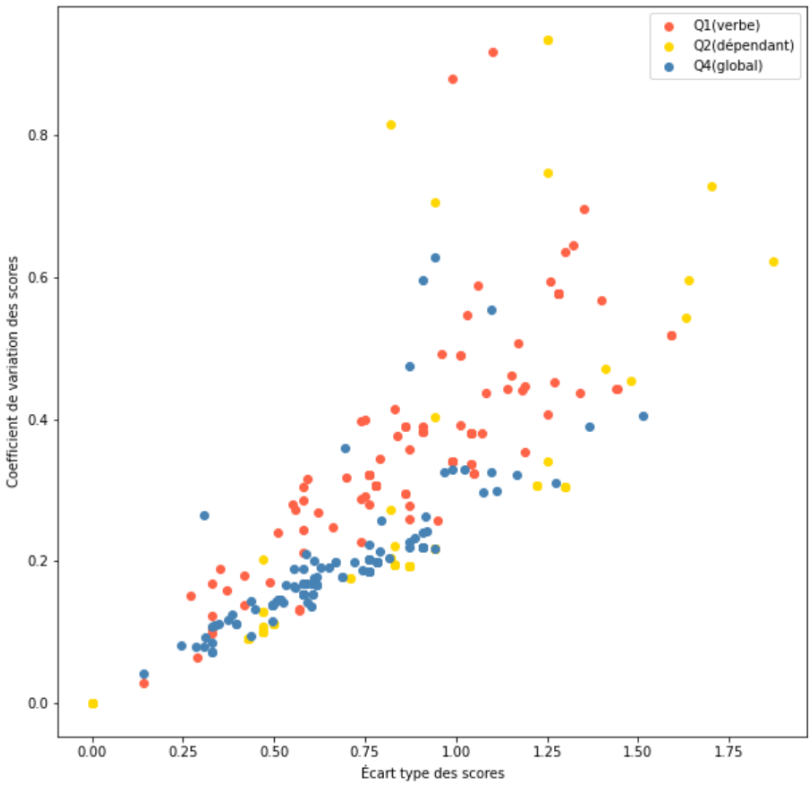


FIGURE 7.3 – L'écart type et le coefficient de variation entre les scores de différents annotateurs

La Figure 7.3 illustre les écarts-types et les coefficients de variation (écart type / moyenne entre annotateurs) de toutes les expressions annotées. Nous constatons que le désaccord entre les annotateurs est assez large pour certaines expressions. De plus, contrairement à notre idée de départ, les annotateurs sont plus d'accord sur les questions portant sur l'expression globale. Le désaccord est généralement plus élevé pour les verbes, mais les différences les plus extrêmes se trouvent parmi les décisions sur les dépendants dans l'expression.

Finalement, comme illustré par la figure 7.1 et la figure 7.2, le « degré moyenné »

a tendance à atténuer l'influence des scores extrêmes de la 4e question donc plus d'expressions sont mises dans « degré moyen » avec cette méthode de calcul.

En conclusion, inspiré par les travaux précédents des autres chercheurs, nous définissons un test de compositionnalité pour déterminer le degré de compositionnalité de chaque expression. Ensuite, en comparant l'annotation de compositionnalité avec l'existence de croisement vrai pour des EPV, nous montrons qu'il existe une corrélation entre ces deux aspects : les EPV ayant un croisement vrai sont généralement plus compositionnelles. Donc le taux de reprise coréférentielle des composants pourrait être un bon indicateur de la compositionnalité de l'EPV. Néanmoins, la méthode d'annotation sur cette question reste à améliorer pour que le résultat soit plus fiable.

Troisième partie

Perspectives

CHAPITRE 8

Discussion

Sommaire

Dans ce chapitre, à partir de notre constatation sur les EPV et les chaînes de coréférence concrètes, ainsi que les statistiques finalement obtenues, nous mettons en avant d'abord certains points ambigus à clarifier avant de poursuivre l'étude de l'hypothèse, et aussi quelques étapes à améliorer pour mieux prouver ou réfuter l'hypothèse. Nous présentons ensuite une autre hypothèse inspirée par cette expérience, dont la vérification est faisable à partir de notre travail actuel. Nous discutons des résultats possibles pour la vérification de cette hypothèse en chinois à la fin de la première section.

Dans la section 2, selon les erreurs commises par le système (surtout pour Seen2Seen) et les spécificités de la langue chinoise, nous proposons certaines pistes possibles pour améliorer la performance de Seen2Seen sur la langue française et aussi la possibilité de l'adapter à la langue chinoise.

8.1 Discussion sur l'étude de l'hypothèse

8.1.1 Affinage de l'hypothèse

Après avoir étudié les différents types des EPV et les exemples concrets, nous trouvons que l'hypothèse doit se limiter à la catégorie VID pour être plus exacte. Étant donné que les MVC et les VPC ne contiennent pas de substantifs, il n'y a aucune chance que leurs composants soient pris dans une chaîne de coréférence, donc nous pouvons les éliminer dès le début. De plus, pour les IRV, à cause de la difficulté pour les annoter, nous n'avons pas pu tester quantitativement l'hypothèse sur cette catégorie.

D'ailleurs, les LVC, qui représentent une majorité de croisements vrais dans notre expérience, ont l'air de contredire notre hypothèse. Mais en considérant la caractéristiques des LVC, dont le sens global de l'expression est essentiellement contribué par le substantif et le verbe joue un rôle essentiellement grammatical, nous pensons que le substantif dans une LVC, surtout un LVC.full, peut représenter le sens global de l'expression dans un

discours au niveau sémantique, la reprise de ces substantifs est donc logique. De l'autre côté, si le verbe apporte du sens, sa raison principale d'être une EPV est plutôt lié au fait que le verbe et le substantif s'utilisent toujours ensemble (idiomatisme statistique). Par conséquent, nous supposons que l'hypothèse peut aussi exclure la catégorie LVC de ces EPV discutées.

De plus, quant à la catégorie VID, nous remarquons que les VID concrets dans le corpus ne sont pas comme ce que nous imaginions au début. Les VID réellement rencontrés peuvent être non seulement les EPV ayant des substantifs plutôt figurés comme dans « couper l'herbe sous le pied », mais aussi celles contenant un nom un peu abstrait dont l'interprétation est mitigée dans différents contextes, telles que « prendre sa place » ou « il est question » dans nos croisements vrais. Ces noms peuvent comporter un sens plutôt général (même un sens latent) donc leur reprise dans une chaîne de coréférence est impossible ; mais, ils peuvent aussi porter un idée plutôt précise qui peut désigner une entité concrète donc peut être prise dans une chaîne de coréférence. La détermination du statut EPV de ces VID diverge souvent en fonction de l'interprétation personnelle de ces noms.

Finalement, l'hypothèse doit également préciser le type de coréférence concerné. Pour ce qui est constaté dans le corpus ANCOR, il existe des cas où la reprise d'un composant dans l'EPV est due à la disfluence et la reformulation de phrases par l'interlocuteur, ou même à la réutilisation de la même EPV dans la parole suivante. Par exemple :

(1) …
- vous regrettez que la langue française se dégrade ou bien que ça **a** pas beaucoup d'**importance** ?
- oh si moi je trouve que ça a de l'importance ah oui
- importance oui ? oui
…

(2) …j'ai toujours du temps je **prends** toujours **le temps** …

Ces deux exemples sont tous typiques dans un corpus oral. L'exemple (1) a répété 2 fois le mot clé « importance », mais la première est due à réutilisation de l'expression, et la 2e est simplement une vérification par l'autre interlocuteur. Nous avons éliminé ce genre de croisements en mettant « répétitions » comme la valeur de sa validation. Mais nous n'avons pas examiné particulièrement les cas de l'exemple (2). L'interlocuteur a reformulé sa phrase afin de trouver une manière de parler plus appropriée. Le substantif de l'EPV a été réutilisé mais il se trouve dans un contexte différent. Nous pensons que l'intérêt de ce genre d'exemple dépend de la distance de ces deux mentions et la similarité de leur contexte. Nous pouvons extraire ce genre d'exemples pour les étudier plus profondément.

8.1.2 Amélioration possible sur la vérification de l'hypothèse

Premièrement, comme mentionné précédemment, nous pouvons introduire plus de catégories pour annoter la validation des exemples, afin de mieux qualifier les croisements détectés. Par exemple, introduire une classe spécifique pour regrouper les exemples venant de reformulations.

Deuxièmement, la taille des corpus écrits et oraux dans notre corpus de validation n'est pas assez équilibrée, et celle des corpus écrits est un peu limitée. Pour que les statistiques obtenues soient plus convaincantes, il vaudrait mieux élargir le corpus de validation du type écrit.

Troisièmement, s'agissant de l'annotation du degré de compositionnalité, nous trouvons que le nombre d'annotateurs est un peu limité. Puisque les réponses des questions se basent essentiellement sur l'annotateur, leur détermination peut être fortement biaisée par ses expériences personnelles. Dans les études similaires, les chercheurs ont choisi d'avoir au minimum 20 annotateurs pour chaque expression afin de diminuer les biais individuels. Nous pouvons pratiquer la même stratégie pour avoir un score final plus juste. D'ailleurs, pour la sélection des EPV sans croisement, nous pouvons limiter notre choix aux corpus emea, frwiki et annodis.ER, afin d'assurer que la raison pour laquelle l'EPV sélectionnée n'a pas de croisement est plutôt à cause de sa nature, mais pas parce qu'elle n'apparaît pas dans le corpus quand nous examinons son croisement avec les chaînes de coréférence.

8.1.3 Hypothèse inspirée par ce travail

Dans les figures 6.6 et 6.7, nous constatons une différence intéressante entre VID et LVC.full : c'est que les VID ont plus de croisements du cas 3 (mention totalement incluse dans l'EPV) que du cas 4 (chevauchement entre les deux), tandis que les LVC.full (même les LVC.cause, mais il manque de données ici) ont beaucoup plus de croisements du type cas 4. Cela pourrait provenir du fait que les LVC.full sont plus compositionnels et les substantifs y contenus sont plus susceptibles d'avoir un modifieur ou un déterminant flexible. Par exemple, dans la phrase suivante :

« …Aclasta ne doit être utilisé, chez les patients **souffrant de** la **maladie** osseuse de Paget, que par un médecin expérimenté dans le traitement de cette maladie.… ».

Le nom « maladie » dans le LVC.full « souffrir de maladie » est modifié par un complément fortement lié au contexte du discours – « osseuse de Paget ».

En revanche, dans un VID comme « **retourner sa veste** », il est beaucoup moins vraisemblable que le nom « veste » ait un complément en gardant une signification idiomatique.

Nous pouvons même imaginer que cela existe non seulement dans les EPV dont la partie nominale est reprise dans une chaîne de coréférence, mais aussi dans toutes les EPV qui possèdent un substantif qui pourrait être détecté comme une mention. Autrement dit, nous supposons que les EPV plus compositionnels sont plus susceptibles d'avoir un groupe nominal contenant un modifieur flexible, qui n'est donc pas annoté dans l'EPV, puisque ces modifieurs en fonction de contexte révèlent que l'entité dans l'EPV a un lien plus fort avec le contexte. En fait, il existe déjà certains travaux proches qui examinaient la modification libre d'un composant interne d'une EP. [Sheinfux et al., 2017] soulignent le lien entre la flexibilité d'une EPV et sa transparence ainsi que sa figuration, qui concernent surtout la signification littérale ou idiomatique de l'expression, soit l'aspect de compositionnalité d'une EPV dans notre étude.

Pour le démontrer, nous pouvons réutiliser les mentions détectées par OFCORS et notre distinction des croisements différents, mais cette fois-ci, nous utilisons toutes les mentions trouvées au lieu de seulement celles qui sont reprises dans une chaîne de coréférence. Selon notre constatation précédente, les erreurs principales d'OFCORS viennent principalement du chaînage des mentions donc les mentions trouvées sont plutôt fiables, et n'exigent pas une vérification manuelle. Ensuite, pour les classifier en différents cas, nous avons peut-être besoin de redéfinir la manière de traiter les déterminants, afin que la classification en différents cas ne soit pas influencée par l'absence des annotations des déterminants changeables dans une EPV.

8.1.4 Résultats en chinois

Selon notre étude de spécificités des EPV en chinois, nous imaginons que les VID en chinois auront encore moins de croisements avec les chaînes de coréférence, puisqu'ils sont généralement plus figés et idiomatiques que les VID en français. Néanmoins, pour les LVC, qui ont l'air d'être la catégorie la plus proche de celle en français, ils auront probablement autant de croisements vrais avec les chaînes de coréférence. Voici un exemple possible de ce genre de croisement :

(1) 酒店 提供 很多 服务， 其 中 叫早 服务 是 我 最
jiǔdiàn **tígòng** hěnduō **fúwù**, qí zhōng jiàozǎo fúwù shì wǒ zuì
hotêl offrir beaucoup services, lesquels parmi réveil service est moi le_plus
满意 的 (zh)
mǎnyì de
satisfaire (part.nom)

'L'hotêl **offre** plusieurs **services**, parmi lesquels le service de réveil me satisfait le plus.'

8.2 Proposition des modifications du système

8.2.1 Amélioration possible pour le français

Selon notre analyse des erreurs, nous remarquons certaines erreurs courantes pour les VID, qui proviennent principalement des limites concernant l'ordre des composants et la forme du pronom s'il existe. Prenons l'exemple suivant :

« jamais oh ben c' **est** parce que **nous** n' avons pas le **temps** »

Dans cette phrase, à part l'expression « avoir le temps » (souligné ici), Seen2Seen a également trouvé « est…nous…temps » à cause de l'expression « il est temps ». Cet exemple illustre clairement ce qui peut arriver quand le VID est assez figé mais le filtrage de Seen2Seen n'est pas suffisamment précis. Pour rappel, les filtres qui contrôlent la distance de discontinuité (f4), ainsi que l'insertion et l'ordre en même temps (f3) ne sont pas activés pour une performance maximale au niveau global.

Néanmoins, selon notre étude sur les EPV dans le corpus PARSEME, sauf les LVC.full, toutes les autres catégories suivent presque toujours le même ordre, avec 93.3% pour les VID et 98.1% pour les IRV (Voir la table 6.8)[1]. En plus, la majorité (76%) des VID en français est continue, contre moins d'un cinquième parmi les LVC.full (voir la table 2.4). De ce fait, nous pouvons tirer la conclusion que la flexibilité varie énormément entre différentes catégories d'EPV. D'ailleurs, spécifiquement pour la catégorie VID en français, si un pronom est annoté comme composant lexicalisé dans l'expression, c'est le « il » impersonnel ou un pronom qui ne désigne pas une entité concrète, comme « ça » dans « ça fait » et « en » dans « en avoir ». Donc la forme du pronom doit être inchangeable.

Par conséquent, nous proposons deux pistes possibles pour l'amélioration de la performance Seen2Seen.

Premièrement, au lieu d'utiliser une seule combinaison optimale de filtres pour toutes les EPV, nous pouvons chercher pour chaque catégorie sa propre meilleure combinaison, ensuite utiliser différents filtres activés pour annoter différentes catégories. Puisque le f8 effectue une contrainte sur l'imbrication des EPV, il faut soit organiser prudemment l'ordre

1. Nous ne discutons pas les LVC.cause ici, puisque le nombre des LVC.cause répétitifs est vraiment limité

de ces plusieurs étapes d'annotation et les faire fonctionner en mode cascade, soit séparer ce f8 des autres filtres pour qu'il fasse son travail à la fin.

Deuxièmement, par rapport à l'inflexibilité des pronoms dans VID, nous pouvons ajouter une étape de filtrage. Ce filtre va mémoriser les VID ayant un pronom inflexible pendant l'entraînement, et éliminer par la suite tous les candidats de ces VID qui ont un pronom différent de ce qui est rencontré précédemment. Par ailleurs, étant donné que les pronoms dans VID ne devraient pas désigner une entité concrète, sa présence dans une chaîne de coréférence est très rare. Par conséquent, si nous disposons de chaînes de coréférence suffisamment fiables, nous pouvons en profiter pour éliminer les VID dont le pronom est repris dans une chaîne de coréférence.

8.2.2 Adaptation en chinois

Selon notre étude des EPV chinois, puisqu'elles sont beaucoup moins flexibles que les EPV français, la tâche d'identification des EPV devrait être plus facile, mais la performance de Seen2Seen pour la langue chinoise est manifestement moins bonne que celle pour le français. C'est peut-être pour les raisons suivantes.

Premièrement, le rappel en chinois est beaucoup moins bon, notamment pour VID et LVC, nous supposons que de nombreux EPV ne sont même pas trouvées pendant la première phase de recherche des candidats.

Nous remarquons tout d'abord qu'il y a une différence entre les corpus de ces deux langues. Dans le corpus chinois, le ratio des EPV inconnues est 38% dans l'ensemble de test, mais ce ratio est seulement 22% dans le corpus français. Puisque Seen2Seen n'est capable de trouver que les EPV connues, le plafond de sa performance en chinois est déjà inférieur à celui en français pour ces deux corpus.

Nous imaginons qu'une autre source d'erreur possible vienne de l'incohérence de segmentation en mots. Puisque la segmentation en mots dépend de la compréhension de l'annotateur et sa pensée sur le regroupement des signifiés, même deux annotations manuelles pourraient différer sur certains points, sans parler de la tokenisation automatique, qui est utilisée pour la majorité du corpus chinois. De plus, les erreurs sont seulement corrigées quand elles posent problème pour l'annotation des EPV. Par conséquent, pour une même EPV, une segmentation différente pourrait conduire à des tokens différents. Par exemple, le LVC.full « 造成 »(causer) peut être un seul token ou une combinaison de 2 tokens « 造 (produire) » et « 成 (achevé) ». Dans ce cas-là, si la méthode de tokenisation d'une EPV dans l'ensemble de test n'est jamais rencontrée par Seen2Seen pendant l'entraînement, l'outil n'aura aucune chance de la reconnaître.

Deuxièmement, la précision pour les autres catégories n'est pas non plus bonne, sauf celle de VID, qui a une précision parfaite. Étant donné que les VID chinois sont plutôt des tokens uniques et fixés, cela prouve notre argument selon lequel l'identification des VID en chinois peut être une tâche simple si les phrases sont correctement segmentées en mots.

De ce fait, nous proposons les essais suivants pour adapter Seen2Seen à la langue chinoise.

Tout d'abord, considérant l'inexistence des flexions en chinois et une plus grande possibilité d'avoir des erreurs sur les étiquettes POS et l'analyse syntaxique, nous pouvons supprimer les filtres sur ces aspects et nous concentrer seulement sur la discontinuité et l'ordre des composants dans les EPV. De nouveau, d'après nous, la performance sera meilleure si nous distinguons la combinaison de filtres activés pour différentes catégories.

Ensuite, puisque les informations du parseur ne nous aident pas énormément, nous pouvons pratiquer une stratégie différente de tokenisation : au lieu de segmenter les phrases

en mots, nous pouvons les segmenter en caractères. De cette manière, nous éviterons les difficultés liées à la segmentation. Si nous déplaçons notre travail à ce stade, nous transformons en fait notre question à une question de segmentation. Les frontières des EPV annotés pourraient servir à la désambiguïsation du texte et à la segmentation des mots plus ordinaires par la suite.

Finalement, si nous voulons que le système soit utilisable pour n'importe quel texte brut avec n'importe quel système d'écriture, il faut aussi enrichir l'ensemble d'entraînement par le système d'écriture de caractères simplifiés.

En conclusion, nous présentons dans ce chapitre les aspects à améliorer pour mieux aborder notre hypothèse afin d'être plus précise, en constatant une différence majeure de VID et LVC sur notre hypothèse, ainsi que l'influence du corpus oral, qui produit les types de chaîne de coréférence non conformes à notre hypothèse. Nous clarifions ensuite le travail restant pour améliorer la démarche de vérification. En tant que perspectives, nous proposons un autre indicateur indirect du degré de compositionnalité à vérifier, ainsi que des pistes possibles pour améliorer la performance du Seen2Seen en langues française et chinoise.

Conclusion générale

Dans ce mémoire, nous essayons de valider quantitativement l'hypothèse portant sur la non-compositionnalité des expressions polylexicales, c'est à dire que les composants dans les expressions polylexicales verbales (EPV) sont peu susceptibles d'apparaître dans des chaînes de coréférence.

Notre expérience montre que le degré de compositionnalité d'une expression polylexicale verbale influence effectivement l'accessibilité des chaînes de coréférence à ses composants. Nous constatons que les LVC.full, qui sont les EPV les plus compositionnelles, représentent une forte proportion dans les croisements validés avec les chaînes de coréférence, tandis que les idiomes verbaux (VID), qui sont moins compositionnels, sont beaucoup plus rares. De plus, parmi les VIDs dans les croisements validés par un humain, leur statut d'EPV est quelquefois discutable. Notre supposition sur le degré de compositionnalité des EPV ayant des croisements validés est également confirmée par l'annotation humaine.

Nous pensons que l'hypothèse sera plus exacte si elle se limite à la catégorie VID. De plus, il faut aussi prendre en compte dans l'hypothèse les types d'anaphores apparaissant dans les chaînes de coréférence, surtout quand nous discutons l'hypothèse sur un corpus oral.

Pendant l'annotation, nous constatons également certaines sources de bruit du système d'identification d'EPV (Seen2Seen). Nous proposons que l'outil considère la divergence entre les catégories d'EPV et applique un filtrage plus strict pour les EPV moins flexibles comme les VID.

Finalement, par rapport aux EPV chinoises, nous supposons que le résultat de vérification de l'hypothèse serait similaire : forte présence des composants des LVC.full dans les croisements, contraire à la rareté des composants des VID. De plus, nous pointons que l'identification des EPV est notamment influencée par les spécificités de la langue chinoise, et nous suggérons deux pistes possibles pour que le système améliore sa performance : ignorer les informations morphologiques ou retravailler la segmentation.

Bibliographie

[Baldwin and Kim, 2010] Baldwin, T. and Kim, S. N. (2010). Multiword expressions. *Handbook of natural language processing*, 2:267–292. – Cité page 16.

[Caillies, 2009] Caillies, S. (2009). Descriptions de 300 expressions idiomatiques: familiarité, connaissance de leur signification, plausibilité littérale,«décomposabilité» et «prédictibilité». *LAnnee psychologique*, 109(3):463–508. – Cité page 63.

[Calzolari et al., 2002] Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., and Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. In *LREC*, volume 2, pages 1934–1940. – Cité page 15.

[Carpuat and Diab, 2010] Carpuat, M. and Diab, M. (2010). Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245. – Cité page 16.

[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 17, 20, 35, 36 et 46.

[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 pages 63 et 64.

[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 31 et 39.

[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 page 39.

[Keysar and Bly, 1995] Keysar, B. and Bly, B. (1995). Intuitions of the transparency of idioms: Can one keep a secret by spilling the beans? *Journal of Memory and Language*, 34(1):89–109. – Cité page 27.

[Laporte, 2018] Laporte, É. (2018). Choosing features for classifying multiword expressions. – Cité page 31.

[Melamed and Melamed, 2001] Melamed, D. I. and Melamed, I. D. (2001). *Empirical methods for exploiting parallel texts*. MIT press. – Cité page 21.

[Moon, 1998] Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford University Press. – Cité page 15.

[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, Barcelona, Spain (Online). International Committee on Computational Linguistics. – Cité page 37.

[Ramisch, 2014] Ramisch, C. (2014). *Multiword expressions acquisition: A generic and open framework*. Springer. – Cité page 36.

[Ramisch et al., 2020] Ramisch, C., Guillaume, B., Savary, A., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymme, S., Walsh, A., Xu, H., Palka-Binkiewicz, E., Ehren, R., Stymne, S., Constant, M., Pasquer, C., Parmentier, Y., Antoine, J.-Y., Carlino, C., Caruso, V., Di Buono, M. P., Pascucci, A., Raffone, A., Riccio, A., Sangati, F., Speranza, G., Ramisch, R., Cordeiro, S. R., de Medeiros Caseli, H., Miranda, I., Rademaker, A., Vale, O., Villavicencio, A., Wick Pedro, G., Wilkens, R., Zilio, L., Rizea, M.-M., Ionescu, M., Onofrei, M., Chen, J., Ge, X., Hu, F., Hu, S., Li, M., Liu, S., Qin, Z., Sun, R., Wang, C., Xiao, H., Yan, P., Yih, T., Yu, K., Yu, S., Zeng, S., Zhang, Y., Zhao, Y., Foufi, V., Fotopoulou, A., Markantonatou, S., Papadelli, S., Louizou, S., Aduriz, I., Estarrona, A., Gonzalez, I., Gurrutxaga, A., Uria, L., Urizar, R., Foster, J., Lynn, T., Elyovitch, H., Ha-Cohen Kerner, Y., Malka, R., Jain, K., Puri, V., Ratori, S., Shukla, V., Srivastava, S., Berk, G., Erden, B., and Yirmibeşoğlu, Z. (2020). Annotated corpora and tools of the PARSEME shared task on semi-supervised identification of verbal multiword expressions (edition 1.2). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. – Cité page 25.

[Rohanian et al., 2019] Rohanian, O., Taslimipoor, S., Kouchaki, S., Ha, L. A., and Mitkov, R. (2019). Bridging the gap: Attending to discontinuity in identification of multiword expressions. *arXiv preprint arXiv:1902.10667*. – Cité page 37.

[Sag et al., 2002] Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *International conference on intelligent text processing and computational linguistics*, pages 1–15. Springer. – Cité page 16.

[Schneider et al., 2014] Schneider, N., Danchik, E., Dyer, C., and Smith, N. A. (2014). Discriminative lexical semantic segmentation with gaps: running the mwe gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206. – Cité page 46.

[Sheinfux et al., 2017] Sheinfux, L. H., Greshler, T. A., Melnik, N., and Wintner, S. (2017). Verbal multiword expressions: Idiomaticity and flexibility. *Representation and parsing of multiword expressions: Current trends. Language Science Press, Berlin*, pages 35–68. – Cité page 73.

[Stodden et al., 2018] Stodden, R., QasemiZadeh, B., and Kallmeyer, L. (2018). Trapacc and trapaccs at parseme shared task 2018: Neural transition tagging of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 268–274. – Cité page 37.

[Taslimipoor and Rohanian, 2018] Taslimipoor, S. and Rohanian, O. (2018). Shoma at parseme shared task on automatic identification of vmwes: Neural multiword expression tagging with high generalisation. *arXiv preprint arXiv:1809.03056*. – Cité page 37.

ANNEXE

EPV les plus fréquentes dans PARSEME

TABLE A.1 – EPV fréquentes en français

Type	**Expression**	**Nombre d'occurrences**
LVC.full	avoir droit	50
	jouer rôle	45
	avoir besoin	38
	faire appel	32
	avoir effet	19
	faire apparition	17
	signer contrat	17
	poser question	17
	jouer match	16
	avoir chance	14
LVC.cause	donner impression	4
	donner occasion	4
	entraîner baisse	3
	entraîner mort	3
	entraîner formation	2
	apporter amélioration	2
	donner opportunité	2
	apporter solution	2
	causer extinction	1
	entraîner poursuite	1
VID	il y a	213
	il faut	168
	il s'agit	140
	avoir lieu	104
	faire partie	98
	tenir compte	37
	faire l'objet	35
	il convient	29
	faire face	25
	prendre part	25
MVC	faire remarquer	4
	faire savoir	4
	faire valoir	3
	laisser tomber	3
	entendre parler	1
	faire passer	1
IRV	se trouver	104
	se dérouler	66
	se situer	52
	se rendre	50
	se produire	45
	s'élever	31
	se passer	27
	se retrouver	27
	s'engager	27
	s'étendre	23

Table A.2 – EPV fréquentes en chinois

Type	Expression	Traduction française	Nbre d'occurrences
LVC.full	有 疑問	avoir question	53
	提供 服務	offrir service	32
	負責 管轄	être chargé de gestion	15
	提供 優惠	offrir promotion	8
	舉辦 活動	organiser activité	7
	進行 研究	mener recherches	6
	有 影響	avoir influence	5
	發表 評論	publier commentaire/critique	5
	表示 反對	présenter (son) opposition	4
	出現 變化	apparaître des changements	4
LVC.cause	產生 影響	avoir influence	6
	造成 傷害	causer dommage	5
	引起 注意	attirer attention	4
	造成 破壞	causer destruction	3
	引起 關注	attirer l'attention (du public)	3
	引起 爭議	provoquer polémique	3
	造成 影響	avoir influence	2
	發動 政變	déclencher coup d'état	2
	引起 爆炸	engendrer explosion	2
	引發 爭議	provoquer polémique	2
VID	賓至如歸	faire sentir chez soi	5
	截然不同	totalement différent	4
	不可或缺	indispensable	4
	不可思議	incroyable	4
	意想不到	inattendu	4
	獨一無二	unique	3
	馬不停蹄	sans cesse	3
	美不勝收	extrêmement beau, magnifique	3
	與眾不同	unique en son genre	3
	取而代之	remplacer (une mention précédente)	3
MVC	成 為	devenir	262
	認為	penser, considérer	77
	變成	devenir	76
	登入	login	53
	造成	causer	51
	分 為	diviser en	39
	組成	constituer	37
	回到	revenir, rentrer	37
	來到	venir	34
	改 為	changer en	32
VPC.semi	提出	proposer	83
	得到	obtenir	64
	完成	finir	61
	受到	reecevoir	54
	看到	voir	52
	推出	publier, sortir, lancer	51
	看 到	voir	44
	遇到	rencontrer	34
	發出	émettre	33
	加上	ajouter	31

ANNEXE B

Étude des croisements

B.1 EPV dans les croisements validés

B.2 EPV sans croisements

1	sous-corpus	forme canonique	Moyenne	Degré moyenné	Q4avg	Q4deg
2	ESLO_ANCO	prendre position	2.86	moyen	2.93	moyen
3	ESLO_ANCO	avoir rôle	3.57	moyen	3.95	moyen
4	ESLO_ANCO	avoir idée	3.78	moyen	3.95	moyen
5	ESLO_ANCO	produire résultat	3.96	moyen	4.13	fort
6	ESLO_ANCO	exercer contrôle	3.84	moyen	3.98	moyen
7	ESLO_ANCO	avoir influence	3.64	moyen	3.80	moyen
8	ESLO_ANCO	avoir possibilité	3.43	moyen	3.63	moyen
9	ESLO_ANCO	avoir intérêt	3.34	moyen	3.45	moyen
10	ESLO_ANCO	prendre le temps	2.65	moyen	3.20	moyen
11	emea	avoir fracture	3.74	moyen	3.95	moyen
12	ESLO_ANCO	faire recherche	3.55	moyen	3.70	moyen
13	ESLO_ANCO	exercer activité	3.74	moyen	4.13	fort
14	ESLO_ANCO	faire effort	3.74	moyen	3.95	moyen
15	ESLO_ANCO	avoir contact	3.57	moyen	3.95	moyen
16	emea	se diriger	3.36	moyen	3.50	moyen
17	frwiki	signer ordonnance	3.69	moyen	3.73	moyen
18	ESLO_ANCO	donner enseignement	3.77	moyen	4.58	fort
19	ESLO_ANCO	passer vacance	3.53	moyen	3.53	moyen
20	frwiki	mener combat	3.24	moyen	3.38	moyen
21	ESLO_ANCO	avoir conséquence	3.55	moyen	3.70	moyen
22	ESLO_ANCO	prendre cours	3.43	moyen	3.43	moyen
23	ESLO_ANCO	avoir connaissance	3.67	moyen	3.50	moyen
24	ESLO_ANCO	avoir besoin	3.52	moyen	3.55	moyen
25	ESLO_ANCO	avoir intention	3.71	moyen	3.80	moyen
26	ESLO_ANCO	donner impression	3.48	moyen	3.48	moyen
27	emea	subir traitement	3.78	moyen	3.95	moyen
28	emea	mener étude	3.58	moyen	3.78	moyen
29	ESLO_ANCO	prendre sanction	3.61	moyen	3.78	moyen
30	emea	atteindre insuffisance	3.10	moyen	3.05	moyen
31	emea	présenter saignement	3.30	moyen	3.08	moyen
32	emea	atteindre maladie	3.24	moyen	3.13	moyen

FIGURE B.1 – EPV dans les croisements validés (1)

1	sous-corpus	forme canonique	Moyenne	Degré moyenné	Q4avg	Q4degre
33	ESLO_ANCO	suivre cours	3.61	moyen	3.60	moyen
34	ESLO_ANCO	avoir responsabilité	4.04	fort	3.95	moyen
35	emea	présenter symptôme	3.54	moyen	3.43	moyen
36	emea	souffrir de syndrome	3.71	moyen	3.78	moyen
37	ESLO_ANCO	faire guerre	3.69	moyen	3.78	moyen
38	ESLO_ANCO	donner réponse	3.87	moyen	3.95	moyen
39	ESLO_ANCO	faire service	3.62	moyen	3.63	moyen
40	ESLO_ANCO	avoir importance	3.53	moyen	3.45	moyen
41	ESLO_ANCO	avoir capacité	3.69	moyen	3.78	moyen
42	ESLO_ANCO	faire étude	3.69	moyen	3.78	moyen
43	ESLO_ANCO	faire confiance	3.25	moyen	3.03	moyen
44	ESLO_ANCO	avoir habitude	3.46	moyen	3.30	moyen
45	ESLO_ANCO	recevoir éducation	3.61	moyen	3.70	moyen
46	ESLO_ANCO	faire demande	3.67	moyen	3.78	moyen
47	ESLO_ANCO	prendre son place	3.67	moyen	3.63	moyen
48	ESLO_ANCO	donner conseil	3.91	moyen	4.13	fort
49	ESLO_ANCO	avoir difficulté	3.47	moyen	3.38	moyen
50	ESLO_ANCO	en savoir	3.89	moyen	4.17	fort
51	ESLO_ANCO	avoir activité	3.83	moyen	4.13	fort
52	ESLO_ANCO	prendre photo	3.65	moyen	3.78	moyen
53	ESLO_ANCO	avoir rapport	3.47	moyen	3.45	moyen
54	ESLO_ANCO	avoir le temps	2.88	moyen	3.40	moyen
55	emea	subir pontage	3.65	moyen	3.78	moyen
56	ESLO_CO2	avoir vocation	3.46	moyen	3.55	moyen
57	ESLO_ANCO	comporter risque	3.62	moyen	3.53	moyen
58	ESLO_ANCO	faire travail	3.68	moyen	3.70	moyen
59	ESLO_ANCO	faire essai	3.76	moyen	3.95	moyen
60	ESLO_ANCO	avoir formation	4.04	fort	3.95	moyen
61	0-100	accomplir travail	3.94	moyen	4.13	fort
62	annodisER	entreprendre action	3.68	moyen	3.73	moyen
63	ESLO_ANCO	il est question	2.02	moyen	1.98	faible
64	ESLO_CO2	donner concert	3.69	moyen	3.78	moyen

Figure B.2 – EPV dans les croisements validés (2)

1	sous-corpus	forme canonique	Moyenne	Degré moyenné	Q4avg	Q4degre
65	ESLO_CO2	donner concert	3.69	moyen	3.78	moyen
66	ESLO_ANCO	avoir religion	4.08	fort	4.13	fort
67	0-100	mener action	3.50	moyen	3.60	moyen
68	ESLO_ANCO	avoir opinion	3.52	moyen	3.55	moyen
69	ESLO_ANCO	avoir impression	3.43	moyen	3.38	moyen
70	ESLO_ANCO	prendre décision	3.43	moyen	3.43	moyen
71	ESLO_ANCO	poser problème	3.46	moyen	3.60	moyen
72	ESLO_ANCO	donner ordre	3.48	moyen	3.53	moyen
73	ESLO_ANCO	faire classe	3.26	moyen	3.38	moyen
74	ESLO_ANCO	faire plaisir	3.24	moyen	3.05	moyen
75	ESLO_ANCO	avoir rendement	4.08	fort	4.13	fort
76	emea	atteindre SCA	3.19	moyen	3.05	moyen
77	emea	réaliser étude	3.91	moyen	4.13	fort
78	ESLO_ANCO	avoir projet	3.69	moyen	3.78	moyen
79	frwiki	porter nom	3.51	moyen	3.40	moyen
80	ESLO_ANCO	faire fête	3.43	moyen	3.43	moyen
81	ESLO_ANCO	avoir problème	3.78	moyen	3.95	moyen
82	ESLO_ANCO	avoir question	3.87	moyen	4.13	fort
83	emea	souffrir de maladie	3.76	moyen	3.88	moyen
84	ESLO_CO2	garder souvenir	3.81	moyen	3.88	moyen
85	0-100	commettre crime	4.00	fort	4.30	fort
86	ESLO_ANCO	avoir relation	3.87	moyen	4.13	fort
87	emea	subir angioplastie	3.74	moyen	3.95	moyen
88	ESLO_ANCO	dispenser enseignement	3.78	moyen	3.95	moyen
89	ESLO_ANCO	faire grève	3.47	moyen	3.38	moyen
90	ESLO_CO2	en revenir	3.01	moyen	3.08	moyen
91	ESLO_ANCO	avoir perception	3.50	moyen	3.48	moyen
92	ESLO_ANCO	donner cours	3.74	moyen	3.95	moyen
93	ESLO_ANCO	poser question	3.54	moyen	3.60	moyen
94	ESLO_ANCO	faire course	3.47	moyen	3.38	moyen
	emea	recevoir perfusion	3.70	moyen	3.70	moyen

FIGURE B.3 – EPV dans les croisements validés (3)

1	sous-corpus	forme canonique	Moyenne	Degré moyenné	Q4avg	Q4degre
2	Sequoia	fournir travail	3.83	moyen	4.10	fort
3	Sequoia	remporter succès	3.87	moyen	4.00	fort
4	Sequoia	trouver refuge	4.01	fort	4.33	fort
5	Sequoia	revêtir caractère	2.81	moyen	2.80	moyen
6	Sequoia	avoir air	3.10	moyen	3.20	moyen
7	Sequoia	classer sans suite	3.32	moyen	3.10	moyen
8	Sequoia	entraîner irrégularité	4.08	fort	4.43	fort
9	Sequoia	réaliser enquête	4.28	fort	4.53	fort
10	Sequoia	effectuer liaison	3.97	moyen	4.53	fort
11	Sequoia	couper l'herbe sous le pied	1.24	faible	1.17	faible
12	Sequoia	mettre en scène	2.90	moyen	2.97	moyen
13	Sequoia	prendre à partie	1.83	faible	1.83	faible
14	Sequoia	faire partie	2.92	moyen	3.07	moyen
15	Sequoia	accorder entretien	3.83	moyen	3.97	moyen
16	Sequoia	conduire audition	3.64	moyen	3.83	moyen
17	Sequoia	chasser sur les terres	2.11	moyen	1.93	faible
18	Sequoia	donner permission	3.80	moyen	3.87	moyen
19	Sequoia	recevoir coup de fil	3.14	moyen	3.83	moyen
20	Sequoia	avoir trait	1.98	faible	1.50	faible
21	Sequoia	mettre en pratique	3.25	moyen	3.00	moyen
22	Sequoia	en croire	2.94	moyen	3.07	moyen
23	Sequoia	prendre en otage	3.65	moyen	3.57	moyen
24	Sequoia	rencontrer succès	3.57	moyen	3.63	moyen
25	Sequoia	traiter par le mépris	3.16	moyen	3.00	moyen
26	Sequoia	avoir occasion	3.62	moyen	3.50	moyen
27	ESLO_ANCOR	il est temps	2.06	moyen	1.53	faible

Figure B.4 – EPV sans croisements

ANNEXE C

Extraits du corpus

C.1 ANCOR

```
12
13  est -ce que c' est surtout votre mariage qui vous a ramené à Orléans ou est -ce que c'
    est une autre chose ?
14
15  y a eu de y a eu plusieurs euh plusieurs raisons euh je travaillais chez Simca enfin
    j' ai mon école est une école d' ingénieurs électroniciens
16
17  hm hm
18
19  j' ai travaillé chez Simca dans l' industrie automobile ça n' avait aucun rapport
    parce que je voulais euh me lancer dans l' automobile ça n' a pas marché pour
    différentes raisons et euh sur ce je me suis marié alors comme euh j' ét- je n' avais
    pas une place qui correspondait à à ce que j' aurais pu avoir chez Simca que j' avais
    essayé de chercher une situation à Paris enfin dans différentes branches et puis ça n'
    a pas marché et j' ai décidé de revenir à Orléans pour de- pour me mettre professeur
    de maths chose que j' avais toujours euh voulu faire mais enfin je faisais j' avais
    fait autre chose
20
21  hm hm
```

FIGURE C.1 – ESLO_ANCOR : (extrait d'un fichier)

```
3   vous savez euh enfin tout le monde euh tout le monde Floriot par exemple quand vous
    écoutez ses causeries enfin hein maître Floriot ben quand même on a beau dire c' est
    quand même des gens qui savent s' exprimer
4
5   oui oui ah je crois que tout le monde sait s' exprimer hein
6
7   oui mais enfin je veux dire par là euh euh on ressent ce qu' ils disent mais on n' est
    pas capable de l' exprimer soi-même souvent quoi
8
9   oui oui moi je crois que chacun chacun se débrouille quand même dans la vie
10
11  non mais enfin euh vous savez i- justement ce qui est intéressant c' est la variété
12
13  voilà n' est -ce pas ? oui oui
14
15  de c' est ça
16
```

FIGURE C.2 – ESLO_CO2 : (extrait d'un fichier)

```
madame

je voudrais avoir une documentation sur Grenoble pour des allemands qui vont venir e

je suis très très pauvre hein

qui connaissent pas alors Grenoble et la région

en allemand je suis très pauvre voilà et je voudrais

français ou en allemand

je vous demande quelques secondes ça c' est le musée

voilà

voilà

j' aurais voulu la même chose en français mais pas celui là non voilà c' est ça "
sprache deutsch " et la merci

je vous en prie

merci bien
```

FIGURE C.3 – OTG : (un fichier entier)

```
U B S bonjour

oui bonjour madame monsieur Nom Siemens j' aurais voulu parler à Monsieur Nom2 s' il
était là

oui j' ai pas bien entendu votre nom monsieur

monsieur Nom Siemens

monsieur Nom Société Siemens c' est ça

de oui

d' accord je vais voir

merci

allo

oui j' ai en ligne monsieur Nom de la société Siemens

d' accord
```

FIGURE C.4 – UBS : (un fichier entier)

ANNEXE D

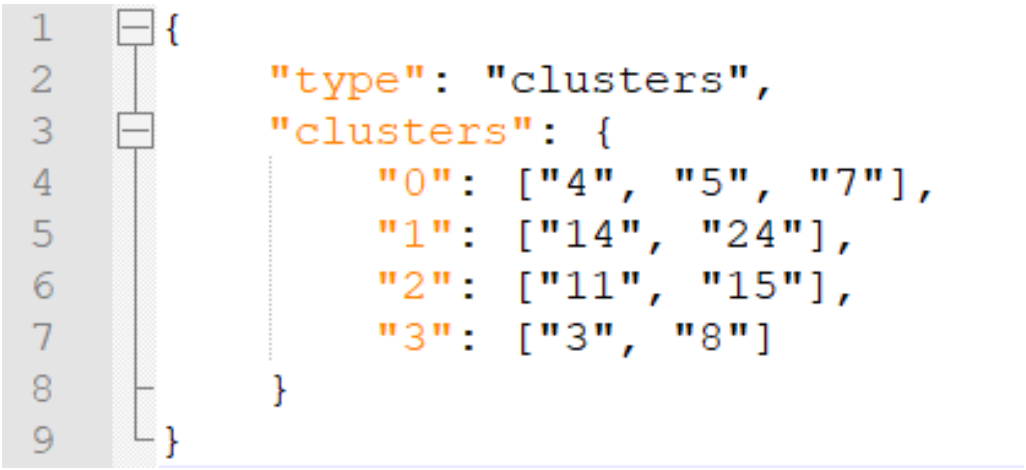
Formats sortants de l'OFCORS


```
{"0": "madame", "1": "je", "2": "voudrais", "3": "avoir", "4": "une", "5":
"documentation", "6": "sur", "7": "Grenoble", "8": "pour", "9": "des", "10":
"allemands", "11": "qui", "12": "vont", "13": "venir", "14": "e", "15": "je",
"16": "suis", "17": "tr\u00e8s", "18": "tr\u00e8s", "19": "pauvre", "20":
"hein", "21": "qui", "22": "connaissent", "23": "pas", "24": "alors", "25":
"Grenoble", "26": "et", "27": "la", "28": "r\u00e9gion", "29": "en", "30":
"allemand", "31": "je", "32": "suis", "33": "tr\u00e8s", "34": "pauvre", "35":
"voil\u00e0", "36": "et", "37": "je", "38": "voudrais", "39": "fran\u00e7ais",
"40": "ou", "41": "en", "42": "allemand", "43": "je", "44": "vous", "45":
"demande", "46": "quelques", "47": "secondes", "48": "\u00e7a", "49": "c'",
"50": "est", "51": "le", "52": "mus\u00e9e", "53": "voil\u00e0", "54":
"voil\u00e0", "55": "j'", "56": "aurais", "57": "voulu", "58": "la", "59":
"m\u00eame", "60": "chose", "61": "en", "62": "fran\u00e7ais", "63": "mais",
"64": "pas", "65": "celui", "66": "l\u00e0", "67": "non", "68": "voil\u00e0",
"69": "c'", "70": "est", "71": "\u00e7a", "72": "\"", "73": "sprache", "74":
"deutsch", "75": "\"", "76": "et", "77": "la", "78": "merci", "79": "je", "80":
"vous", "81": "en", "82": "prie", "83": "merci", "84": "bien"}CRLF
```


FIGURE D.1 – 1AG0141_tokens.json


```
{"0":
    {"CONTENT": ["madame"],
     "MENTION_ID": "u-MENTION-sduchon_1329078552160",
     "START": "0",
     "END": "0",
     "SPAN-ID": "0-0"
     },
 "1":
    {"CONTENT": ["je"],
     "MENTION_ID": "u-MENTION-sduchon_1329078394693",
     "START": "1",
     "END": "1",
     "SPAN-ID": "1-1"
     },
 "2":
    {"CONTENT": ["une", "documentation"],
     "MENTION_ID": "u-MENTION-sduchon_1329078576074",
     "START": "4",
     "END": "5",
     "SPAN-ID": "4-5"
    },
```


FIGURE D.2 – 1AG0141_mentions_output.json (formant transformé du corpus ANCOR)



```
{
    "type": "clusters",
    "clusters": {
        "0": ["4", "5", "7"],
        "1": ["14", "24"],
        "2": ["11", "15"],
        "3": ["3", "8"]
    }
}
```

FIGURE D.3 – 1AG0141_resulting_chains.json


```
{
    "1": {
        "CONTENT": ["Affaire", "des", "caporaux", "de", "Souain"],
        "LEFT_CONTEXT": ["<start>"],
        "RIGHT_CONTEXT": ["<end>"],
        "START": "0",
        "END": "4",
        "SPAN_ID": "0-4"
        },
    "2": {
        "CONTENT": ["des", "caporaux", "de", "Souain"],
        "LEFT_CONTEXT": ["<start>", "Affaire"],
        "RIGHT_CONTEXT": ["<end>"],
        "START": "1",
        "END": "4",
        "SPAN_ID": "1-4"
        },
```


FIGURE D.4 – frwiki_1_mentions_output.json (fromat sortant de l'OFCORS)

ANNEXE E

Notations et abréviations dans les exemples

E.1 Notations

— Dans un exemple de croisement, les tokens annotés comme composants de l'expression polylexicale sont mises en gras.
Exemple : « **poser** une **question** »
— Dans un exemple de croisement, les mentions dans une chaîne de coréférence sont marquées par le soulignement.
Exemple : « Xavier est jeune, mais il n'aime pas faire du sport. »

E.2 Types d'informations dans un exemple d'EP

Voici une liste des types d'informations éventuellement contenus dans un exemple d'EP de ce texte. Une explication plus détaillée se trouve sur https://gitlab.com/parseme/pmwe/-/blob/master/Conventions-for-MWE-examples/PMWE_series_conventions_for_multilingual_examples.pdf.

a un exemple d'usage de l'EP ;

b si ce n'est pas une EP française, le code ISO 639 de la langue parlée est fourni ;

c une translittération de l'exemple si le système d'écriture de la langue n'utilise pas les lettres latines, dans notre cas, les exemples chinois ;

d une glose suivant les règles de glose de Leipzig ;

e une traduction littérale entre guillemets simples, précédée par *lit.* ;

f une traduction idiomatique entre guillemets simples.

E.3 Abréviations

Voici les abréviations des mots grammaticaux utilisées dans les exemples chinois :

* **part.changement** : 了, particule modale, pour exprimer un changement d'état.
* **part.achèvement** : 了, particule placé après un verbe ou un adjectif pour indiquer l'achèvement de l'action ou changement.
* **part.complement** : 得, particule structurelle souvent utilisée pour les compléments.
* **part.nom** :的, particule structurelle utilisée avant les substantifs, précédé d'un mot s'emploie comme qualificatif ou déterminatif.

ANNEXE F

Guide d'annotation sur la compositionnalité des EPV

Attention :L'annotation est à réaliser sur l'expression elle-même, les phrases ne servent que d'exemples de contexte. Elles permettent de montrer le sens de l'expression et de ne pas la confondre avec une autre qui serait composée des mêmes lemmes. Donc, lors des réponses aux questions, il ne faut prendre en compte que les parties annotées comme expressions polylexicales. Par exemple pour l'expression « **présenter saignements** », bien que la phrase d'exemple contient « présentant un saignement actif », il ne faut pas prendre « actif » en compte car il ne fait pas partie de l'expression. On cherche à évaluer la compositionnalité de l'expression dans tous ses emplois.

Les notes de chaque question sont indépendantes.

I **Lire l'expression polylexicale**

II **Lire la phrase contenant cette expression**

III **Transformer l'expression en sa forme canonique (dans le sens du guide PARSEME)**

IV **Donner des synonymes de l'expression**

V **Répondre aux questions suivantes (par rapport à la forme canonique) :**

1 **À quel point la tête verbale (de la forme canonique) de l'expression contribue-telle au sens de l'expression ? [0 à 5]**

— **1A Aspect sémantique (70% du score de Question 1) : [0 à 5]**
- Est-ce que le sens du verbe contribue au sens de l'expression (oui : note haute, non : note basse) ?
Exemple : « **accomplir** travail » vs « **rendre** visite »

- Le sens du verbe est-il idiomatique (note basse) ou littéral (note haute) ?
Exemple : « **avoir** beau » vs. « se **diriger** »

— **1B Aspect grammatical (30% du score de Question 1) : [0 à 5]**
- Le verbe précise-t-il le rôle sémantique du sujet (oui : note haute, non : note basse) ?
Exemple : « **avoir** raison » vs « **donner** raison »

- Est-il possible et courant d'utiliser le verbe avec une forme fléchie ?
Exemple : « traduire c'est trahir », ici tout est figé, note basse

- Dans ce cas, apporte-t-il des informations morphologiques (oui : note haute, non : note basse) ?
Exemple des informations morphologiques : pluriel, futur, personne, conditionnel...

2 **À quel point les dépendants lexicalisés du verbe contribuent-t-il au sens de l'expression ? [0 à 5]**

3 **S'il y a des modifieurs de ce dépendant, à quel point contribuent-ils au sens de l'expression ? [0 à 5]**
Les modifieurs doivent faire partie de l'expression annotée : s'il n'y en a aucun, mettez "/".

4 **À quel point le calcul du sens de l'expression (à partir du sens des composants) est-il cohérent avec sa syntaxe ? [0 à 5]**

— **4A Aspect sémantique (70% du score de Question 4) : [0 à 5]**
- Le sens de l'expression globale est-il idiomatique (note basse) ou littéral (note haute) ?
Exemples :
commettre crime : C'est vraiment un crime commis (note haute)
souffrir de maladie : On ne souffre pas forcément mais on parle bien d'une maladie (note moyenne)
avoir beau : Ce n'est ni avoir ni beau (note basse)

— **4B Aspect structurel (30% du score de Question 4) : [0 à 5]**
- La structure est-elle flexible (oui : note haute, non : note basse) ?
Exemple : « drôle de question » vs « poser problème »
(Aspects possibles à examiner : inflexibilité morphologique (conjugaison de verbes, déclinaison des dépendants etc.) ; syntaxe inchangeable (les composants ont une syntaxe particulière (comme « drôle de question ») / ils sont toujours collés ensemble ou dans un certain ordre (comme « il est temps »), etc.)

ANNEXE G

Recueil des liens vers les corpus, outils et les scripts développés

G.1 Corpus

- **PARSEME :**
  https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3367
- **Sequoia seulement :**
  https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3429
- **ANCOR :**
  https://sharedocs.huma-num.fr/wl/?id=omxQyO3MkOgUgfEVwOlBygZZHYAocTDt
- **Est Républicain :**
  http://redac.univ-tlse2.fr/corpus/estRepublicain.html
- **ANNODIS :**
  http://redac.univ-tlse2.fr/corpus/annodis/ANNODIS_rr.zip

G.2 Outils

- **Seen2Seen :**
  https://gitlab.com/cpasquer/st_2020
- **OFCORS :**
  https://gitlab.com/Stanoy/ofcors
- **Modèle UDPipe 2.5 :**
  https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3131
- **DeCOFre :**
  https://github.com/LoicGrobol/decofre

G.3 Références d'information

- **Guide d'annotation PARSEME :**
  https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/
- **Évaluation du Seen2Seen :**
  http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_02_MWE-LEX_2020___lb__COLING__rb__&subpage=CONF_50_Shared_task_results
- **Explication du format cupt :**

http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_04_LAW-MWE-CxG_2018___lb__COLING__rb__&subpage=CONF_45_Format_specification

G.4 Scripts développés

Veuillez consulter les scripts développés sur la page Git : https://github.com/anaelle-p/MWE_coref