Simplification automatique de textes techniques et spécialisés

Rémi Cardon

Université de Lille
École doctorale Sciences Humaines et Sociales
U.M.R. 8163 – Savoirs, Textes, Langage

Numéro attribué par la bibliothèque

Thèse

Pour obtenir le grade de
Docteur en Sciences du Langage
Discipline : Traitement Automatique des Langues

Préparée et soutenue publiquement
par

Rémi Cardon

Le 19 avril 2021

Simplification Automatique de Textes Techniques et Spécialisés

Directrices de thèse :

Natalia Grabar, Chargée de recherches, CNRS
Anne Carlier, Professeure des universités, Sorbonne Université

Composition du jury :

Cécile Fabre	Professeure des universités	Université de Toulouse Jean Jaurès (rapporteuse)
Thomas François	Chargé de cours	Université Catholique de Louvain (rapporteur)
Emmanuelle Canut	Professeure des universités	Université de Lille (examinatrice)
Pascal Denis	Chargé de recherches	Inria Nord Lille-Europe (examinateur)
Thierry Hamon	Maître de conférences	Université Sorbonne Paris Nord (examinateur)
Horacio Saggion	Associate professor	Universitat Pompeu Fabra (examinateur)

Remerciements

Je tiens tout d'abord à remercier Natalia Grabar. Mener une thèse sous la direction de Natalia signifie bénéficier non seulement de tout l'accompagnement nécessaire à la réalisation des différentes étapes pour mener une thèse à son terme, mais aussi de toute la liberté indispensable à la formation d'un chercheur autonome. Ces trois années de travail en relation de confiance mutuelle resteront comme un moment marquant de ma vie professionnelle.

Je remercie également Anne Carlier, avant tout pour avoir rendu cette collaboration possible, mais aussi pour sa bienveillance et pour sa disponibilité malgré ses nombreuses activités.

Merci à tous les membres du jury d'avoir immédiatement accepté d'en faire partie, et pour les échanges provoqués lors de la soutenance. Bien que la soutenance ait eu lieu en distanciel, je suis heureux d'avoir pu rencontrer chacun d'entre eux en personne à différentes occasions. J'espère que de telles occasions se renouvelleront.

Merci à toutes les personnes avec qui j'ai pu interagir durant cette thèse, trop nombreuses pour être nommées, et parfois anonymes. Je pense à tous les collègues doctorantes et doctorants, chercheuses et chercheurs confirmé(e)s, croisé(e)s lors des divers événements que sont les conférences, séminaires, colloques ou encore écoles d'été. Que ce soit en personne lors des événements scientifiques (ou en marge de ces derniers) ou indirectement à l'occasion de la réception des tant redoutées relectures qui accompagnent les notifications d'acceptation ou de rejet d'article, toutes ces interactions furent stimulantes et enrichissantes. Merci aussi à tous les auteurs et toutes les autrices dont les noms se trouvent dans la section Bibliographie du présent document.

J'adresse un grand merci à tout mon entourage hors de ma vie professionnelle, ma famille, mes amis. Merci aux personnes qui en majorité ont essayé de s'intéresser à ce que je faisais. Merci aussi à la minorité qui n'a pas essayé de s'y intéresser.

Merci tout particulièrement à Alba. Subir un an de confinement est une épreuve en soi, mais avec un doctorant en dernière année de thèse cela tourne à l'exploit.

Enfin, merci à l'ANR pour le financement de cette thèse, menée dans le cadre du projet CLEAR, n°ANR-17-CE19-0016-01.

Table des matières

Table des figures

Liste des tableaux

1. Introduction

1.1. Motivation et contexte

La disponibilité croissante d'informations médicales et de santé sur l'Internet facilite l'accès à ces informations. Il est en effet possible de consulter des publications scientifiques, des articles d'encyclopédies ou des publications de sociétés savantes tout en restant chez soi en quelques clics. Cependant, il a été observé que cette démocratisation d'informations spécialisées n'améliore pas leur compréhension par le grand public (Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs, 1999). Nous illustrons l'opacité des informations médicales librement accessibles dans l'exemple (1), issu d'un article de Wikipédia [1], une encyclopédie en ligne collaborative, libre et gratuite. Comme nous pouvons le voir, ce texte, qui est pourtant créé à destination du grand public, contient de nombreux termes dont la compréhension par des personnes sans formation médicale n'est pas évidente.

(1) *Le cholestéatome est une forme d'otite chronique avec présence d'épithélium pavimenteux stratifié dans l'oreille moyenne. Cet épithélium desquame et se kératinise (structure histologique de l'épiderme), et peut provoquer l'érosion voire la destruction des structures contenues dans et autour de l'oreille moyenne.*
La forme la plus fréquente est le cholestéatome acquis par évolution terminale d'une otite chronique (poches de rétractions essentiellement). Une perforation tympanique acquise post-traumatique ou post-otitique peut également entrainer un cholestéatome par migration de l'épiderme du conduit par la perforation surtout si elle est au contact du sulcus (perforation dite marginale).

Par ailleurs, des études ont montré qu'une meilleure compréhension des informations de santé par les patients et leurs familles conduit à une meilleure adhésion au traitement et à un processus de soins plus réussi (Berkman *et al.*, 2011). En effet, cette étude indique qu'une faible compréhension des informations de santé mène vers des comportements inadaptés consistant en un recours moindre aux services de soin, en une mauvaise aptitude à suivre des traitements ou des recommandations de santé publique. De plus, l'incompréhension des informations de santé peut également détériorer la qualité de la communication entre le médecin et le patient et diminuer la confiance mutuelle. La compréhension d'informations médicales par le grand public est donc un enjeu sociétal important.

1. `https://fr.wikipedia.org/wiki/Cholest%C3%A9atome`

1.2. Deux aspects liés à la compréhension de documents de santé

Pour assurer la bonne compréhension de documents, deux aspects entrent en jeu. Chacun des ces aspects est géré par un domaine de recherche spécifique et complémentaire :

1. D'un côté, nous avons la personne qui lit le document. Il s'agit typiquement du patient ou de sa famille, qui n'ont en général pas de connaissances spécifiques dans le domaine médical. Selon son expérience personnelle, le patient présente ainsi une certaine connaissance et alphabétisation médicale, et donc une certaine capacité à comprendre les informations médicales. Le domaine de recherche qui s'occupe de cet aspect s'appelle éducation thérapeutique du patient (ETP). Ce domaine a pour objectif de rendre les patients plus aptes à recevoir l'information médicale et à la traiter [2]. C'est également ce domaine qui est en capacité de définir les besoins informationnels du patient.
2. De l'autre côté, nous avons le document lu par le patient. Le document a ses propres caractéristiques et, entre autre, son niveau de lisibilité. Pour améliorer la lisibilité du document, il est nécessaire de diagnostiquer les difficultés de compréhension et de les simplifier. Ce processus peut être effectué manuellement ou automatiquement. Lorsque ce processus est effectué automatiquement, il relève du domaine de traitement automatique de langues (TAL).

Pour assurer une compréhension optimale d'un document, les deux aspects doivent être appariés : le document doit présenter un niveau de lisibilité satisfaisant pour le niveau d'alphabétisation du lecteur. Ainsi, les informations présentées doivent comporter le niveau optimal de détails par rapport aux attentes informationnelles de son lecteur.

Notre travail s'intéresse à la simplification automatique de documents de santé : il se place donc du côté de cette deuxième question de recherche. Nous allons donc proposer et mettre en oeuvre des méthodes de TAL.

À ce jour, il existe plusieurs initiatives qui poursuivent l'objectif de faciliter l'accès aux informations à destination de différents types de population. Nous présentons deux de ces initiatives :

— FALC *facile à lire et à comprendre* (Audiau, 2009) est une initiative active au niveau européen. Il s'agit d'un ensemble de recommandations définies pour permettre une présentation d'informations accessible au plus grand nombre d'utilisateurs. Ces recommandations sont accessibles en ligne sur le site d'UNAPEI [3]. Ces recommandations portent sur différents aspects des documents, comme par exemple leur mise en page mais surtout la présentation d'informations. Ainsi, une des recommandations indique qu'il faut placer le texte avec une phrase par ligne, faire des phrases courtes, ne pas utiliser de négations, ne

2. https://www.has-sante.fr/jcms/c_1241714/fr/education-therapeutique-du-patient-etp

3. https://www.unapei.org/wp-content/uploads/2018/11/L%E2%80%99information-pour-tous-Re%CC%80gles-europe%CC%81ennes-pour-une-information-facile-a%CC%80-lire-et-a%CC%80-comprendre.pdf

pas utiliser plusieurs polices de caractères dans un même texte, ne pas utiliser de notions abstraites, etc.
— Dans le cadre de l'ETP, la Haute Autorité de Santé a également publié un cadre méthodologique pour la conception de documents d'informations écrits [4]. Ces recommandations sont spécifiquement dédiées aux documents de santé.

1.3. Simplification automatique

La simplification automatique de textes est un domaine du traitement automatique des langues, qui a pour objectif d'appliquer des transformations sur les phrases d'un texte afin de les rendre plus lisibles, tout en conservant leur sens intact. La tâche a une importance aussi bien à destination des humains (Carroll *et al.*, 1999) que pour faciliter d'autres applications du TAL (Chandrasekar *et al.*, 1996).

Concernant la simplification effectuée à destination des machines, différentes applications concrètes du TAL sont donc concernées. Nous présentons ici quelques exemples :

— La première application de simplification cherchait à simplifier les structures de phrases pour qu'elles soient plus faciles à traiter par les analyseurs syntaxiques (Chandrasekar *et al.*, 1996).
— Par ailleurs, la simplification a été utilisée pour adapter certains types de textes à des outils, qui n'ont pas été entraînés pour les traiter spécifiquement, comme par exemple l'analyse d'un texte biomédical effectuée avec des outils entraînés sur des textes journalistiques (Jonnalagadda *et al.*, 2009).

Concernant la simplification effectuée à destination des humains, ces méthodes sont explorées à destination de différents publics :

— les personnes mal ou non alphabétisées (Williams & Reiter, 2005) ;
— les personnes sourdes qui ont également des difficultés de lecture et d'écriture (Inui *et al.*, 2003) ;
— les lecteurs dyslexiques (Rello *et al.*, 2013) :
— les personnes atteintes d'autismes (Barbu *et al.*, 2013).

Dans le domaine médical – dans lequel nous nous plaçons ici – la simplification peut également servir à faciliter l'éducation thérapeutique des patients (Brin-Henry, 2014) ou l'accès à l'information par les enfants (De Belder & Moens, 2010).

1.4. Objectifs

L'objectif de notre travail consiste à contribuer au domaine de la simplification automatique de textes de spécialité, en ciblant la méthode sur le traitement et la simplification de textes médicaux. Si la majorité de travaux de simplification traitent les données en langue anglaise, nous travaillons avec des données en français. Cela introduit donc une double difficulté :

— absence de données pour la création d'algorithmes pour la simplification de textes en français,

4. https://www.has-sante.fr/jcms/c_430286/fr/elaboration-d-un-document-ecrit-d-information-a-l-intention-des-patients-et-des-usagers-du-systeme-de-sante

— absence de données pour la création d'algorithmes pour la simplification de textes médicaux.

Au commencement de notre travail, aucune ressource dédiée à la simplification de documents médicaux en français n'était donc disponible. Ainsi, notre travail se décompose en plusieurs tâches :

— Nous commençons par constituer un corpus comparable dont les documents sont différenciés par leur degré de spécialisation (chapitre 2).
— Nous effectuons un alignement manuel de phrases à partir de ce corpus comparable afin de faire une analyse des procédés mis en place lors de la simplification (chapitre 2).
— Nous décrivons ensuite notre méthode automatique proposée pour la constitution d'un corpus parallèle et aligné à partir du corpus comparable collecté (chapitre 3).
— Enfin, nous présentons des expériences en simplification automatique effectuées avec une méthode neuronale issue de la traduction automatique, ainsi que les résultats que nous obtenons (chapitre 4).

Le schéma en figure 1.1 montre les étapes successives de notre travail.

FIGURE 1.1. – Les étapes du travail présenté.

Nous terminons la présentation de notre travail par un bilan général de nos contributions (chapitre 5).

2. Création du corpus comparable

2.1. Introduction

Dans ce chapitre, nous nous concentrons sur la ressource fondamentale à partir de laquelle nous effectuons nos travaux de simplification automatique de textes : le corpus. Les méthodes de simplification, qu'elles soient à base de règles ou d'apprentissage automatique, nécessitent des corpus parallèles. De tels corpus mettent en regard des textes ou des phrases qui se différencient par leur degré de technicité ou de lisibilité. Les corpus parallèles alignés peuvent être obtenus en extrayant des couples de phrases parallèles à partir de corpus parallèles ou comparables. Ces corpus contiennent donc des couples de phrases où les deux phrases expriment le même sens. Notons que nous utilisons deux termes (*paire de phrases* et *couple de phrases*) selon les étapes de la méthode :

— le terme *paire de phrases* est utilisé lorsque deux phrases n'ont pas de relation identifiée entre elles. Par exemple, nous parlons de paires de phrases tant que nous n'avons pas la certitude qu'il s'agisse bien de phrases parallèles et alignables (ou alignées) ;

— le terme *couple de phrases* est utilisé lorsque les deux phrases candidates à l'alignement ont pu être alignées et lorsque cet alignement est validé. La relation entre ces deux phrases est alors explicitement catégorisée. De même, les données de référence avec des phrases alignées manuellement contiennent des couples de phrases.

Aucun corpus de phrases parallèles et alignées n'était disponible pour le français médical quand nous avons commencé notre travail. Notre première tâche a donc consisté en création d'un corpus parallèle du domaine médical pour le français. Dans ce chapitre, nous présentons d'abord un panorama des corpus comparables et parallèles existants dans le domaine de la simplification automatique de textes (section 2.2). Dans un second temps, nous décrivons les données collectées pour constituer un corpus comparable du domaine biomédical en français (section 2.3). Ensuite, nous décrivons l'étape d'alignement manuel de données à partir d'un échantillon du corpus comparable (section 2.4). Entre autres, cette étape permet d'avoir une idée de la taille du corpus parallèle qu'il serait possible d'obtenir à partir du corpus comparable. Nous menons enfin une analyse détaillée des processus de simplification observés dans les phrases alignées manuellement (section 2.5). Cette analyse montre les spécificités de notre corpus spécialisé par rapport à des typologies élaborées pour la langue générale. Enfin, nous concluons le chapitre en faisant le bilan du travail effectué : il s'agit entre autres de la construction d'un corpus comparable et d'un ensemble de couples de phrases alignées (section 2.6). Nous proposons également quelques perspectives pour poursuivre ce travail.

Les travaux présentés dans ce chapitre sont essentiellement basés sur les publica-

tions suivantes :

— constitution d'un corpus pour la simplification en français médical (Grabar & Cardon, 2018a),
— analyse et typologie de procédés de simplification (Koptient *et al.*, 2019).

2.2. Corpus existants

Dans le domaine de simplification automatique, les corpus comparables contiennent des textes complexes et simples traitant des mêmes sujets. Ils nécessitent des méthodes spécifiques, ou un pré-traitement, avant d'être transformés en corpus parallèles alignés et pouvoir être exploités pour les travaux sur la simplification (Brunato *et al.*, 2014). Si les corpus comparables ont l'avantage d'être plus facilement disponibles que les corpus parallèles, les traitements requis pour leur transformation en corpus parallèles alignés sont assez conséquents.

Il existe quelques corpus parallèles alignés, obtenus à partir de corpus comparables, disponibles pour la recherche. Ils sont principalement le résultat de simplifications manuelles. Le corpus le plus fréquemment utilisé est le corpus SEW-EW (*Simple English Wikipédia – English Wikipédia*). SEW-EW propose des couples de documents issus d'articles de Wikipédia en anglais [1] et de leur version en *Simple English* [2]. C'est un corpus disponible librement, qui a été souvent utilisé dans les travaux de recherche (Zhu *et al.*, 2010a; Biran *et al.*, 2011; Coster & Kauchak, 2011a). Actuellement, il existe deux autres corpus, qui sont également utilisés dans les travaux sur la simplification automatique en anglais :

— Newsela [3] (Xu *et al.*, 2015) est un corpus d'articles de journaux réécrits selon quatre niveaux de simplification. Il est à noter que les consignes données aux rédacteurs ne sont pas publiques. Newsela n'est pas libre d'utilisation. Il est notamment interdit de publier les modèles entraînés sur ces données. De plus, le corpus n'est pas distribué avec une division en ensembles d'entraînement, de test et de validation, ce qui rend difficile la comparaison entre différents travaux qui exploitent ce corpus ;
— WikiLarge (Zhang & Lapata, 2017) est la compilation de trois corpus publiés précédemment (Zhu *et al.*, 2010a; Woodsend & Lapata, 2011; Kauchak, 2013), tous issus de Wikipédia. WikiLarge est disponible librement sans restriction pour la recherche, avec une division en ensembles d'entraînement, de test et de validation.

Il est également possible de construire un corpus parallèle à partir de la version francophone de Wikipédia [4] et de Vikidia [5], une encyclopédie en ligne créée sur le modèle de Wikipédia à destination des enfants de 8 à 13 ans. Cette source a servi pour un travail sur la simplification syntaxique en français (Brouwers *et al.*, 2014) mais n'est pas rendue disponible pour d'autres chercheurs.

1. https://en.wikipedia.org/wiki/Main_Page
2. https://simple.wikipedia.org/wiki/Main_Page
3. https://newsela.com/
4. https://fr.wikipedia.org
5. https://fr.vikidia.org

Notons aussi que des corpus parallèles alignés ont été créés pour les travaux de simplification dans d'autres langues :

— Espagnol : Simplext (Saggion *et al.*, 2015)
— Portugais du Brésil : PorSimples (Caseli *et al.*, 2009)
— Danois : D-Sim (Klerke & Søgaard, 2012)
— Italien : TERENCE et TEACHER [6] (Brunato *et al.*, 2015), PaCCSS-IT [7] (Brunato *et al.*, 2016)
— Français : ALECTOR (Gala *et al.*, 2020)

Certains de ces corpus indiquent explicitement ce qui a été simplifié et comment (suppression, segmentation...). C'est le cas du corpus italien PaCCSS-IT et du corpus français ALECTOR. Notamment, un schéma d'annotation a été proposé pour la simplification avec plusieurs classes de modifications : découpage, regroupement, réorganisation, insertion (verbes, sujets ou autres), suppression (verbes, sujets ou autres), transformation (substitution lexicale, élucidation des anaphores, changement de voix, changement des traits verbaux...) (Brunato *et al.*, 2014). Ce schéma d'annotation couvre la simplification lexicale et la simplification syntaxique.

2.3. CLEAR : un corpus médical comparable pour la simplification

Pour pouvoir mener nos travaux de recherche sur la simplification de textes biomédicaux en français, un corpus parallèle relevant du domaine biomédical est donc nécessaire. Ce corpus doit consister en une compilation de couples de phrases qui expriment le même sens et où l'une est formulée de façon plus complexe que l'autre. Comme nous l'avons déjà mentionné, aucun corpus de simplification pour le domaine biomédical en français n'existait au moment où nous avons commencé notre travail.

Afin de construire ce corpus, nous avons exploité trois types de ressources francophones liées au domaine médical : des articles d'encyclopédie en ligne (section 2.3.1), des notices d'information sur les médicaments et leur utilisation (section 2.3.2), et des résumés de revues systématiques produits par la fondation Cochrane (section 2.3.3). Ces trois sources sont libres d'utilisation pour la recherche, avec des modifications non autorisées pour les résumés Cochrane. Les textes couvrent différents sujets : des sujets divers liés à la médecine pour les articles encyclopédiques, les médicaments pour les notices, et des questions liées au diagnostic et au traitement des maladies pour les résumés Cochrane. Une partie de ces données a été alignée manuellement au niveau des phrases afin de permettre la création automatique d'un corpus parallèle par la suite. Nous décrivons ce processus d'extraction automatique de phrases parallèles en section 2.4.

6. http://www.italianlp.it/resources/terence-and-teacher/
7. http://www.italianlp.it/resources/paccss-it-parallel-corpus-of-complex-simple-sentences-for-italian/

2.3.1. Articles encyclopédiques

La partie *Encyclopédie* du corpus est composée d'articles de deux encyclopédies collaboratives en français disponibles en ligne : Wikipédia et Vikidia. Wikipédia en français s'adresse au grand public francophone, alors que Vikidia a été créée pour donner accès à des informations similaires à un public d'enfants de 8 à 13 ans. Ces deux encyclopédies fournissent des articles sur une grande diversité de sujets : politique, économie, médecine, culture, histoire, géographie, etc. Wikipédia a une couverture plus large que Vikidia au regard du nombre de sujets couverts et du volume d'informations donné. En effet, Wikipédia est un projet plus ancien et plus renommé. La rédaction d'articles dans ces encyclopédies doit se plier à des règles précises : les articles doivent être clairs et compréhensibles, formels, avec un usage aussi limité que possible du jargon des domaines spécialisés. De plus, comme Vikidia s'adresse à des enfants, les articles doivent aussi contenir les éléments suivants : une introduction et des définitions simples, un développement clair, des exemples, des sources ainsi que des liens externes et, dans la mesure du possible, des images, des schémas, des ressources audio et vidéo. La participation des enfants au processus de rédaction des articles est encouragée [8]. Bien que les articles de ces deux encyclopédies traitent des mêmes sujets et que les articles de Wikipédia inspirent parfois ceux de Vikidia, les articles de ces deux encyclopédies sont rédigés indépendamment.

Les articles encyclopédiques ont été recueillis dans les *dumps* du 17 septembre 2017 pour Wikipédia et d'août 2017 pour Vikidia. Globalement, Wikipédia comprend environ deux millions d'articles et Vikidia autour de cinquante mille articles. De l'ensemble des articles de Wikipédia, nous conservons seulement les 20 972 articles faisant partie du portail médical [9]. Parmi ceux-ci, 575 articles partagent leur titre avec un article de Vikidia. Ces 575 sujets et couples d'articles sont retenus pour intégrer le corpus. Ces articles de Wikipédia contiennent environ trois millions d'occurrences de mots, contre un peu moins de deux cent mille pour ceux de Vikidia.

Un effort complémentaire est fait pour couvrir la langue générale, en plus de la langue médicale. Ainsi, un travail similaire est effectué en janvier 2019 sur l'ensemble de Wikipédia et Vikidia. Cela a permis de répertorier 19 879 articles communs entre les deux encyclopédies, avec 59 022 858 occurrences de mots pour la partie Wikipédia, et 6 144 521 occurrences de mots pour Vikidia.

Ce corpus sera dorénavant mentionné sous le nom de « corpus *Encyclopédie* ».

2.3.2. Notices de médicaments

Par obligation légale, tout médicament mis sur le marché en France doit être accompagné d'une notice qui informe les citoyens sur sa composition, les indications de prescription qui y sont liées, les effets indésirables connus et affiche divers avertissements. Ces informations sont créées sous deux formes différentes. Une version est créée à destination des professionnels de santé. Elle donne des informations techniques et complètes sur un médicament donné. De plus, cette version utilise une structure spécifique et une terminologie médicale très riche. Une autre version est

8. `https://fr.vikidia.org/wiki/Aide:Comment_cr%C3%A9er_un_article`
9. `https://fr.wikipedia.org/wiki/Portail:M%C3%A9decine`

destinée aux patients, et contient des informations essentielles mais accessibles. Le texte s'adresse au lecteur. Des expressions récurrentes sont utilisées, comme *votre santé*, *votre médecin* ou *vous pouvez*. L'information est structurée sous forme de questions-réponses : *Qu'est-ce que [nom du médicament] ?*, *Quels sont les effets indésirables éventuels ?* Ces versions simplifiées sont systématiquement créées pour chaque médicament qui entre sur le marché. Elles sont ensuite imprimées et insérées sous forme de notice dans les boîtes des médicaments. Ce corpus est collecté à partir des documents disponibles dans la *base de données publique des médicaments*[10] gérée par le Ministère de la Santé en France. Ces documents ont été téléchargés au mois de juin 2017. Le corpus contient 11 800 médicaments, avec les informations techniques et les informations simplifiées pour chacun d'entre eux. La partie technique contient plus de cinquante millions d'occurrences de mots, la partie simplifiée en contient plus de trente millions.

Ce corpus sera dorénavant mentionné sous le nom de « corpus *Médicaments* ».

2.3.3. Résumés Cochrane

Le but de la fondation Cochrane est de fournir l'information médicale basée sur des preuves de haute qualité (Sackett *et al.*, 1996). Depuis plusieurs années, la fondation Cochrane demande à des chercheurs du domaine médical de travailler sur la création de revues systématiques. Les revues systématiques traitent de différentes questions de recherche, souvent en relation avec le diagnostic et le traitement de maladies. La démarche de production des revues systématiques consiste à collecter et faire lire les travaux existants autour d'une question précise à des experts, qui en font une synthèse, avec une validité scientifique et méthodologique plus importante que celle des travaux individuels. Cela fournit de l'information de haute qualité aux professionnels de santé. Pour chaque revue de ce type, un court résumé est créé. En plus de ces résumés techniques à destination des experts, des résumés simplifiés sont créés pour le grand public. Cette démarche de simplification des résumés à destination du grand public est une évolution récente des activités de la fondation Cochrane.

Ce corpus est construit avec les documents disponibles sur le site de la Bibliothèque Cochrane[11]. Les documents ont été téléchargés au mois de novembre 2017. Cela représente 8 789 résumés systématiques. Parmi celles-ci, 3 815 résumés proposent une version technique et une version simplifiée. La partie technique contient un peu moins de trois millions d'occurrences de mots et la partie simplifiée en contient un million et demi.

Ce corpus sera dorénavant mentionné sous le nom de « corpus *Cochrane* ».

2.3.4. Bilan

Le tableau 2.1 récapitule la taille des trois parties du corpus. Pour chaque corpus, nous indiquons le nombre de couples de documents, le nombre d'occurrences de mots dans la partie technique et le nombre d'occurrences de mots dans la partie simplifiée.

10. `http://base-donnees-publique.medicaments.gouv.fr/`
11. `http://www.cochranelibrary.com/`

Corpus	# doc.	# tokens (tech.)	# tokens (simple)	# lemmes (tech.)	# lemmes (simple)
Médicaments	11 800×2	52 313 126	33 682 889	59 420	44 131
Cochrane	3 815×2	2 840 003	1 515 051	24 252	16 663
Encyclopédie	575×2	2 293 078	197 672	84 245	10 407
Total	16 190×2	57 446 207	35 395 612	127 371	51 686

TABLEAU 2.1. – Taille du corpus comparable, en couples de documents, en nombre d'occurrences de mots et en nombre de lemmes uniques, pour la partie technique et la partie simple.

Les trois corpus présentent des disparités. Le corpus *Médicaments* est le plus fourni avec 11 800 couples de documents, suivi par le corpus *Cochrane* qui en compte 3 815. Ces deux corpus présentent approximativement deux fois moins d'occurrences de mots dans leur partie simple par rapport à leur partie technique. Le corpus *Encyclopédie* montre une différence plus grande entre les deux parties : la partie simple contient est environ dix fois moins volumineuse que la partie technique. C'est aussi le sous-corpus le moins fourni, avec 575 couples de documents. Le nombre unique de lemmes est moins élevé dans la partie simple des trois sous-corpus que dans la partie technique. Comme pour les occurrences de mots, le corpus *Encyclopédie* présente la différence la plus importante entre les deux parties. Cela peut être expliqué non seulement par la rédaction plus simple mais aussi par la différence de taille entre les deux parties du corpus. Les deux autres corpus présentent un écart similaire : le nombre de lemmes dans la partie simple représente environ deux tiers du nombre de lemmes dans la partie technique.

2.4. Création des données de référence : couples de phrases alignées manuellement

Le corpus comparable collecté et décrit dans la section précédente (section 2.3) est la source de données à partir de laquelle nous créons un corpus de phrases parallèles alignées. En effet, le corpus de phrases parallèles alignées est une ressource nécessaire pour l'extraction de règles de transformation indispensables pour effectuer la simplification automatique de textes. Afin d'entamer la création d'un corpus parallèle, nous avons aligné manuellement un sous-ensemble du corpus comparable au niveau de la phrase. Cet alignement manuel sert à créer des données de référence annotées pour permettre l'utilisation de méthodes d'alignement basées sur l'apprentissage automatique.

Nous avons sélectionné aléatoirement 14 couples d'articles d'encyclopédie, 12 couples de documents d'information sur les médicaments et 13 couples de résumés Cochrane. L'alignement a été fait par deux annotateurs travaillant dans le domaine du traitement automatique des langues appliqué aux textes biomédicaux. L'alignement a été mené indépendamment par les deux annotateurs, avec pour seule consigne d'aligner deux phrases lorsqu'elles expriment le même sens. L'accord est comptabilisé quand les annotateurs proposent le même alignement de phrases et le désaccord quand une paire de phrases n'est proposée à l'alignement que par un des deux anno-

tateurs. Dans notre expérience, l'accord inter-annotateurs – le κ de Cohen (Cohen, 1960) – est de 0,76. Il a été calculé avec l'ensemble de phrases proposées pour alignement par les deux annotateurs. Un tel score est qualifié d'accord substantiel selon l'échelle d'interprétation habituellement utilisée (Landis & Koch, 1977) et indique une bonne fiabilité des données obtenues. Dans un deuxième temps, les désaccords sont discutés afin d'arriver à un consensus. À la suite de ces discussions, nous avons également défini plusieurs critères pour l'alignement ou le non-alignement de deux phrases, technique et simplifiée. Dans les exemples qui suivent, les phrases sont tirées des résumés Cochrane. Pour chaque exemple, la phrase technique précède la phrase simplifiée.

1. Les phrases identiques ou qui varient seulement par la ponctuation ou des mots grammaticaux ne sont pas alignées. Même si de telles paires de phrases ont un contenu sémantique très proche ou identique, nous considérons qu'elles ne contiennent pas de transformations utiles à la création ou l'apprentissage de règles de simplification de textes ;
2. Un verbe conjugué doit être présent dans les deux phrases ;
3. Les phrases à aligner doivent exprimer le même sens ou presque (équivalence sémantique) et doivent au moins faire apparaître des adaptations lexicales et/ou syntaxiques. Nous illustrons cela à l'exemple (1)
4. Une des deux phrases exprime toutes les informations présentes dans l'autre phrase, mais également des informations absentes de l'autre phrase. Ceci est le cas de l'inclusion sémantique. Dans l'exemple (2), le contenu de la phrase simplifiée est inclus dans la phrase technique.
5. Des informations sont partagées entre les deux phrases mais chacune d'entre elles apporte de l'information supplémentaire qui lui est propre. Ce cas est appelé *intersection sémantique*, que nous illustrons avec l'exemple (3) :

(1) *Les enfants prématurés présentent un risque d'hémorragie préventriculaire (HPV).*
Les bébés nés très tôt (avant 34 semaines) présent un risque de saignement au cerveau (hémorragie périventriculaire).

(2) *Nous n'avons trouvé aucune étude qui rendait compte de l'effet des régimes contenant des céréales complètes sur la mortalité cardiovasculaire totale ou le nombre total d'événements cardiovasculaires (infarctus du myocarde, angor instable, pontage aorto-coronarien, angioplastie coronaire transluminale percutanée, avc).*
Nous n'avons trouvé aucune étude rendant compte de l'effet des céréales complètes sur les décès de maladies cardiovasculaires ou les événements cardiovasculaires.

(3) *Des études à plus grande échelle sont nécessaires pour déterminer la possibilité d'événements indésirables lorsque les ultrasons sont utilisés pour confirmer le positionnement des sondes.*
Des études à plus grande échelle sont nécessaires pour déterminer si les ultrasons pourraient remplacer les rayons x pour confirmer la mise en place d'une sonde gastrique, et pour évaluer si les ultrasons pourraient permettre

> *de réduire les complications graves, telles que la pneumonie résultant d'un tube mal placé.*

Ces critères ont permis d'atteindre le consensus. Le résultat est une caractérisation des paires de phrases comme étant alignées ou non. Nous avons choisi de garder les cas d'inclusion car ils permettent de conserver une trace des alignements multiples, en cas de découpage ou de regroupement de phrases. En revanche, notre démarche nous conduisant à travailler au niveau de la phrase, les cas d'intersection sémantique liés à la simplification sont plus complexes et relèvent plutôt d'un travail sur l'organisation de l'information au niveau du document. Pour cette raison, nous excluons les cas d'intersection. Pour appuyer cette remarque, considérons l'exemple suivant des articles *Cauchemar* du corpus *Encyclopédie* :

(4) - *Dans le langage populaire, le cauchemar est un mauvais rêve.*
- *Un cauchemar est un mauvais rêve, créé par le cerveau durant le sommeil.*

Les deux articles définissent le cauchemar comme étant *un mauvais rêve*. Les informations propres à chaque phrase (*dans le langage populaire*, *créé par le cerveau durant le sommeil*) ne relèvent pas d'une réorganisation syntaxique : la mention de langage populaire est absente de l'article de Vikidia [12] et le mot *cerveau* n'apparaît pas dans l'article de Wikipédia [13]. En effet, il faut aller dans l'article *Rêve* [14] pour en trouver la mention. Avec cet exemple, nous illustrons que les cas d'intersection sémantique peuvent impliquer des aspects qui les démarquent nettement des cas de découpage et de regroupement de phrases.

Le tableau 2.2 affiche la taille de l'ensemble à partir duquel la tâche d'annotation a été menée, ainsi que le volume de phrases résultant de l'alignement consensuel. Nous ajoutons le taux d'alignement pour chacun des trois corpus, en rapportant le nombre de phrases alignées au nombre total de phrases des documents d'origine, pour la partie technique et pour la partie simple. Ce taux est calculé pour avoir une vision sur le degré de comparabilité des corpus.

		Technique				*Simplifié*				*Taux*	
		brut		*aligné*		*brut*		*aligné*		*d'alignement*	
Corpus	*doc.*	*ph.*	*occ.*	*ph.*	*occ.*	*ph.*	*occ.*	*ph.*	*occ.*	*tech.*	*simp.*
Médicaments	12×2	4 391	44 684	143	4 227	2 710	27 804	143	8 481	3,25	5,27
Cochrane	13×2	426	8 852	84	2 278	227	4 688	84	2 466	19,71	36,56
Encyclopédie	14×2	2 416	36 703	39	873	235	2 659	39	710	1,61	16,6
Total	39×2	7 233	90 239	266	7 378	3 172	35 131	266	11 657		

Tableau 2.2. – Taille des données de référence avec l'alignement consensuel.

Suite à l'alignement manuel, nous obtenons un total de 266 couples de phrases alignées. Ceci nous fournit les données de référence nécessaires pour entamer le travail d'alignement automatique.

12. `https://fr.vikidia.org/wiki/Cauchemar`
13. `https://fr.wikipedia.org/wiki/Cauchemar`
14. `https://fr.wikipedia.org/wiki/R%C3%AAve`

Un autre point intéressant concerne le parallélisme entre les versions simple et technique des documents. En effet, le degré de parallélisme entre deux corpus peut varier entre corpus presque parallèles, corpus avec beaucoup de phrases parallèles, et corpus assez éloignés (*very-non-parallel corpora*) (Fung & Cheung, 2004). Nous pouvons voir que les phrases alignées sont relativement moins fréquentes dans les corpus *Médicaments* et *Encylopédie* que dans le corpus *Cochrane*. Ceci peut être expliqué par les spécificités des corpus :

— La ligne directrice de rédaction des versions simplifiées des résumés Cochrane affiche explicitement une volonté de simplifier le contenu de ses résumés d'origine pour le grand public. Les rédacteurs prennent donc comme point de départ les résumés techniques et les simplifient ;
— L'objectif de Vikidia est de traiter des sujets présents dans Wikipédia mais pour un public d'enfants. La création d'articles de Vikidia est rarement basée sur les articles de Wikipédia mais résulte le plus souvent d'une écriture indépendante ;
— Quant au corpus *Médicaments*, conformément à la législation, les informations sur les médicaments sont créées à destination des professionnels de santé et des patients. Certaines de ces informations sont propres à la version technique (composition plus détaillée, action sur l'organisme, molécules...), alors que d'autres sont propres à la version simplifiée (précautions d'emploi, mises en garde...). Cela dit, des informations sont communes aux deux versions (posologie, instructions d'utilisation effets indésirables...), ce qui explique la présence de nombreux alignements possibles.

Le taux d'alignement, entre 1,61 et 36,56, illustre ces différences entre les corpus. Il serait intéressant de formaliser la notion de parallélisme entre deux corpus, ce qui pourrait donner lieu à un indicateur du taux de phrases alignées qu'il serait possible d'en extraire.

Les premières observations des phrases parallèles montrent qu'elles fournissent principalement des transformations lexicales et syntaxiques et que les principes de simplification diffèrent en fonction des corpus. Par exemple, les phrases tendent à être découpées dans les notices de médicaments et les articles d'encyclopédie, alors qu'elles sont plutôt regroupées lors du processus de simplification des résumés Cochrane. Nous avons analysé ces phénomènes de façon plus détaillée (Koptient *et al.*, 2019) et faisons état des résultats dans la section suivante.

2.5. Typologie des procédés de simplification

Dans cette section, nous présentons une méthode visant à appréhender un échantillon d'un corpus comparable afin d'en produire une typologie des opérations de transformation de phrases pour la simplification (section 2.5.1). Nous présentons ensuite les résultats de l'application de cette méthode sur un échantillon du corpus CLEAR (section 2.5.2).

Ce travail est basé sur une publication antérieure (Koptient *et al.*, 2019).

2.5.1. Méthode

Notre méthode d'analyse des procédés de simplification repose sur trois tâches principales :

1. le contrôle des relations d'inclusion sémantique (Section 2.5.1.1) ;
2. l'annotation sémantique des couples de phrases pour décrire plus précisément les transformations (Section 2.5.1.2) ;
3. l'étiquetage et l'analyse morpho-syntaxique pour associer les natures sémantique et syntaxique des séquences transformées (Section 2.5.1.3).

2.5.1.1. Annotation des cas de regroupement et de découpage de phrases

Une stratégie typique appliquée lors de la simplification de textes consiste à regrouper ou découper les phrases techniques lors de la création de phrases simples (Brouwers *et al.*, 2014). Avant le regroupement, les phrases techniques sont raccourcies, ce qui permet leur association en une phrase qui reste lisible dans la version simplifiée. À l'inverse, quand une phrase technique donnée contient plus d'une proposition, il est possible de faire de chaque proposition une phrase indépendante. Il est à noter que, dans certains cas, le découpage d'une phrase complexe en plusieurs phrases simples peut perturber la compréhension des phrases nouvellement créées. En effet, une proposition d'une phrase complexe peut nécessiter des informations cruciales contenues dans une autre proposition pour être compréhensible (Brunato *et al.*, 2014). Dans de tels cas, il convient d'éviter le découpage pour la simplification.

Dans notre corpus de phrases alignées, les cas de phrases regroupées ou découpées sont détectés à l'aide de leur proximité dans le corpus et à leurs alignements multiples. Nous illustrons le propos ici avec un exemple de découpage et un exemple de regroupement :

— Découpage :
  — *Elle impose l' arrêt du traitement et contre-indique toute nouvelle administration de clindamycine.*
  — *Prévenez votre médecin immédiatement car cela impose l' arrêt du traitement.*
  Cette réaction va contre-indiquer toute nouvelle administration de clindamycine.
— Regroupement :
  — *la chimiothérapie par induction peut prolonger la survie de 8 à 20 % et la chimioradiothérapie adjuvante concomitante peut prolonger la survie jusqu' à 16 %.*
  Chez les patients présentant des tumeurs non résécables, la chimioradiothérapie concomitante ou alternée peut prolonger la survie de 10 à 22 %.
  — *L'amélioration de la survie globale avec la chimiothérapie est estimée entre 8 % et 22 %.*

2.5.1.2. Schéma d'annotation sémantique en types de transformation

Les transformations sémantiques liées à la simplification sont annotées dans YAWAT (*Yet Another Word Alignment Tool*) (Germann, 2008). YAWAT permet de

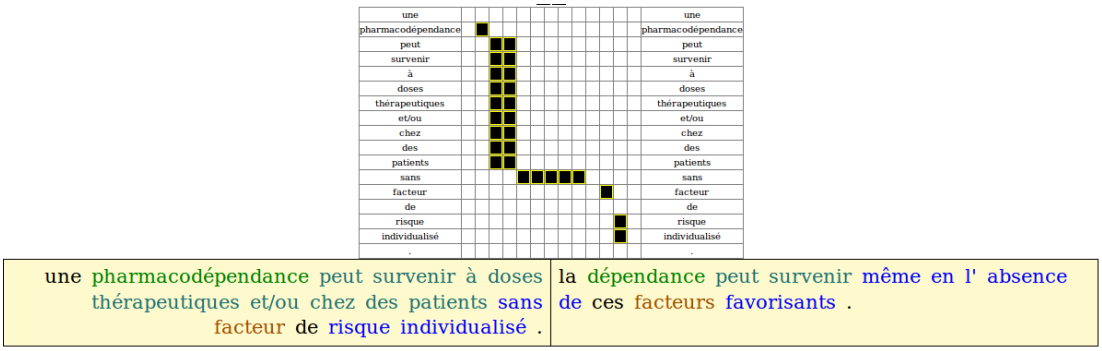


FIGURE 2.1. – Alignement des mots en matrice dans l'interface de YAWAT.

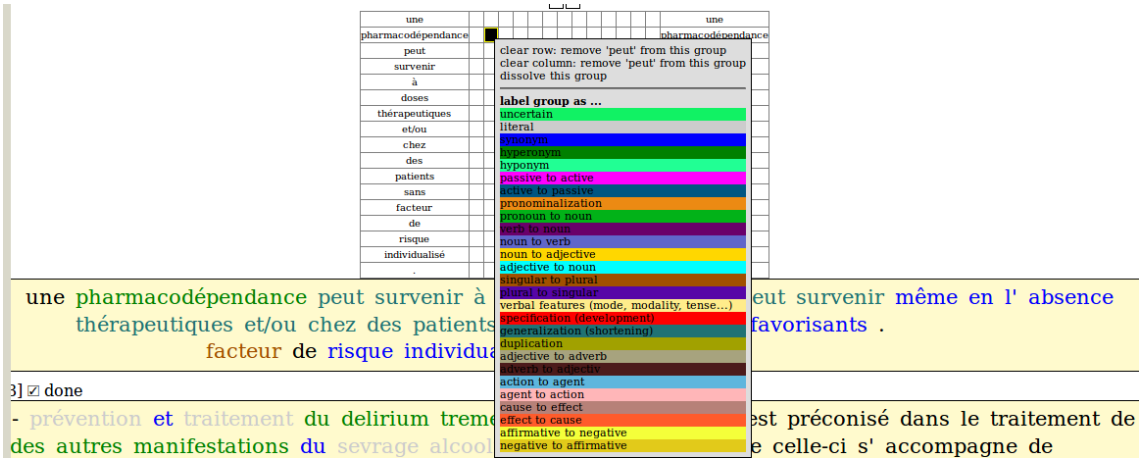


FIGURE 2.2. – Schéma d'annotation dans l'interface de YAWAT.

visualiser et manipuler des textes parallèles. Il a été conçu pour travailler avec des textes bilingues parallèles correspondant à des traductions mutuelles (Yu *et al.*, 2012). Nous proposons de l'exploiter avec des textes parallèles monolingues liés à la simplification. YAWAT affiche les deux phrases alignées côte à côte. L'annotateur peut ensuite aligner les mots à l'aide de la matrice (Figure 2.1) et assigner le type de transformation à chaque couple de segments contextuels. Les types de transformation servent à décrire plus précisément leur nature sémantique. Nous avons défini un ensemble de types de transformation pertinents pour notre corpus en partant de travaux similaires (Brunato *et al.*, 2014) et de l'analyse de notre corpus.

— *littéral*, mots identiques dans les deux phrases. Il s'agit de la valeur par défaut,
— *synonyme*, substitution d'un mot complexe par un synonyme *{pharmacodépendance}{dépendance}*,
— *hyperonyme*, substitution d'un mot technique par un hyperonyme *{clindamicine}{médicament}*,
— *hyponyme*, substitution d'un mot technique par un hyponyme *{manifestations cutanées}{éruptions cutanées}*
— *précision*, ajout de l'explication d'un terme médical *{bêta-lactamines}{bêta-lactamines (pénicilline, céphalosporine)}*. La différence avec *synonymie* et que, au lieu de la substitution, le terme technique reste et une explication (défini-

tion, exemple...) est ajoutée,
- *généralisation*, suppression d'information spécifique *{arrêt du traitement et contre indique toute nouvelle administration du clindamycine}{arrêt du traitement}*,
- *duplication*, deux occurrences ou plus du même terme dans une phrase unique,
- *pronominalisation*, substitution par un pronom *{l'antibioprophylaxie}{elle}*,
- *p2a* (et *a2p*), la voix passive dans la phrase technique est reformulée à la voix active, même sans identité lexicale du verbe *{ne doit jamais être utilisé}{ne prenez jamais}* et l'inverse *{n'a aucun}{n'est pas attendu}*,
- *p2n*, substitution d'un pronom par sa dénomination sous forme de nom *{elles}{ce médicament}*,
- *v2n* (et *n2v*), substitution d'un verbe par un nom *{conduire}{conduite}* et l'inverse *{l'arrêt du traitement}{arrêter brutalement}*,
- *n2a* (et *a2n*), substitution d'un nom par l'adjectif correspondant *{allergies}{allergiques}* et l'inverse *{cardiaque}{du cœur}*,
- *s2p* (et *p2s*), substitution d'une forme au singulier par une forme au pluriel *{de tout antibiotique}{d'antibiotiques}* et l'inverse *{les enfants}{l'enfant}*,
- *adj2adv* (et *adv2adj*), substitution d'un adjectif par un adverbe *{récente}{récemment}* et l'inverse *{tard}{tardif}*,
- *agt2act* (et *act2agt*), substitution de l'agent par l'action *{conducteurs}{conduite}* et l'inverse *{conduite}{conducteurs}*,
- *cau2eff* (et *eff2cau*), substitution de la cause par l'effet *{prescrits}{utilisés}* et l'inverse *{dans le traitement}{chez les patients atteints}*,
- *aff2neg* (et *neg2aff*), substitution de la forme affirmative de l'information par la négation de l'affirmation inverse *{présentant une absence complète}{n'avez aucune}* et l'inverse *{ne pas}{éviter de}*.

En comparaison avec le travail de Brunato *et al.* (2014) pour l'italien, qui fait apparaître 12 transformations, la typologie que nous proposons est plus riche et précise. Notre typologie affiche 23 transformations possibles (Figure 2.3 page 33).

Comme certaines séquences peuvent être caractérisées avec plusieurs types de transformations, nous avons défini des règles de priorité lors de l'annotation. Ces règles de priorité servent à appréhender les cas tels que *{cardiaque}{du cœur}* : dans cet exemple il s'agit d'expressions qui sont synonymes, mais on observe aussi le passage d'un adjectif à un nom. La règle de priorité qui s'applique est *a2n* > *synonyme*. Cela résulte d'un choix de donner la préférence à la règle applicable qui fournit la description la plus précise. Toute autre substitution par une autre partie du discours (*n2v, adj2adv*, etc.) aura donc également la priorité sur la synonymie.

2.5.1.3. Annotation syntaxique

L'analyse syntaxique permet d'annoter les phrases parallèles avec des informations syntaxiques. Nous partons du principe que les informations sémantiques et syntaxiques doivent être étudiées ensemble pour une meilleure description des transformations. L'analyse syntaxique est faite à l'aide de Cordial (Laurent *et al.*, 2009), qui procède à la segmentation, l'étiquetage morpho-syntaxique, la lemmatisation et l'analyse syntaxique en constituants. Le tableau 2.3 montre un exemple d'étiquetage et d'analyse de Cordial pour la phrase en (5). Nous pouvons voir que la séquence

nb.	*forme*	*partie du discours*	*groupe synt.*
1	dalacine	NCI	1
2	n’	ADV	3
3	a	VINDP3S	3
4	aucun	ADJIND	5
5	effet	NCMS	5
6	ou	COO	-
7	qu’	ADV	3
8	un	DETIMS	9
9	effet	NCMS	9
10	négligeable	ADJSIG	9
11	sur	PREP	13
12	l’	DETDFS	13
13	aptitude	NCFS	13
14	à	PREP	15
15	conduire	VINF	15
16	des	DETDPIG	17
17	véhicules	NCMP	17
18	et	COO	-
19	à	PREP	20
20	utiliser	VINF	20
21	des	DETDPIG	22
22	machines	NCFP	22
23	.	PCTFORTE	-

Tableau 2.3. – Exemple de l’annotation syntaxique de Cordial (position du mot dans la phrase, forme du mot, partie du discours et groupe syntaxique).

un effet négligeable appartient à un seul et même groupe syntaxique, comme indiqué dans la colonne *groupe synt.*

(5) *Dalacine n’a aucun effet ou qu’un effet négligeable sur l’aptitude à conduire des véhicules et à utiliser des machines.*

2.5.2. Résultats

Nous présentons maintenant les résultats obtenus avec l’application de la méthode d’annotation sur le corpus CLEAR, pour chacune des trois étapes :

— les cas de regroupement et de découpage de phrases (section 2.5.2.1)
— les transformations lexicales et syntaxiques (section 2.5.2.2)

2.5.2.1. Regroupement et découpage de phrases

Dans l’ensemble, nous avons compté 51 cas dans lesquels au moins deux phrases techniques sont regroupées en une seule phrase simplifiée et 16 cas dans lesquels une

phrase technique est découpée en au moins deux phrases différentes. Dans un travail précédent, il a été observé que le regroupement de phrases lors de la simplification est un phénomène rare (Brouwers *et al.*, 2014). Cependant, dans notre corpus, nous observons la tendance inverse : il y a plus de cas de regroupement que de découpage. Nous pouvons avancer plusieurs raisons à cela :

— Dans le travail évoqué (Brouwers *et al.*, 2014), les auteurs travaillent sur des articles de Wikipédia et Vikidia. Vikidia s'adresse à des enfants de 8 à 13 ans et fournit un jeu de règles strictes pour la rédaction d'articles. L'une des règles demande d'utiliser des phrases courtes et claires. Dans notre travail, Wikipédia et Vikidia correspondent à la partie *Encyclopédie* du corpus. Les deux autres parties du corpus ne suivent pas les mêmes principes de rédaction.

— La particularité des notices de médicaments est qu'elles contiennent beaucoup de coordinations et de listes d'éléments (maladies, effets indésirables, fonctions...). Souvent, le choix est de les présenter comme des listes à points dans les documents techniques, alors que dans les documents simplifiés ces listes apparaissent comme des phrases coordonnées. Lorsqu'il y a plus d'une phrase pour indiquer les éléments, ils sont regroupés par types, comme par exemple : effets indésirables gastro-intestinaux, effets indésirables neurologiques, etc.

— Dans les résumés des revues systématiques, les phrases techniques sont souvent raccourcies lors de la simplification. En conséquence, les phrases peuvent être regroupées en même temps. Contrairement à Vikidia, la rédaction de résumés Cochrane en langage simplifié n'est pas régie par des consignes générales et chaque éditeur peut appliquer ses propres principes.

2.5.2.2. Analyse des transformations lexicales et syntaxique

Dans la figure 2.3, nous présentons la typologie des transformations liées à la simplification. Cette figure présente également la prévalence de chaque transformation à l'aide de sa fréquence en valeur absolue (occurrences brutes) et relative (pourcentage sur l'ensemble des transformations). Nous faisons la distinction entre plusieurs transformations de haut niveau, qui peuvent se retrouver dans les typologies existantes (Brunato *et al.*, 2014; Brouwers *et al.*, 2014) : la substition lexicale, l'insertion lexicale, la suppression lexicale, la substitution syntaxique, la pronominalisation et l'utilisation des formes affirmatives et négatives.

Nos choix présentent des différences par rapport aux typologies existantes :

— nous considérons que les substitutions strictement syntaxiques correspondent au changement de voix du verbe (de la voix active à la voix passive et inversement). Ainsi, le passage du singulier au pluriel et autres traits verbaux sont classés dans la catégorie des substitutions lexicales sans glissement sémantique.

— nous séparons la synonymie de l'hyperonymie car ces deux types de relation lexicale présentent des différences fondamentales (équivalence sémantique ou subsomption) et requièrent des méthodes et ressources différentes lors de la simplification.

— nous ne distinguons pas les transformations lexicales et sémantiques : la sémantique devient un trait de la substitution lexicale.

— nous ne faisons pas la distinction, opérée par Brunato *et al.* (2014), qui consiste à distinguer plusieurs cas d'insertion et de suppression de mots en fonction

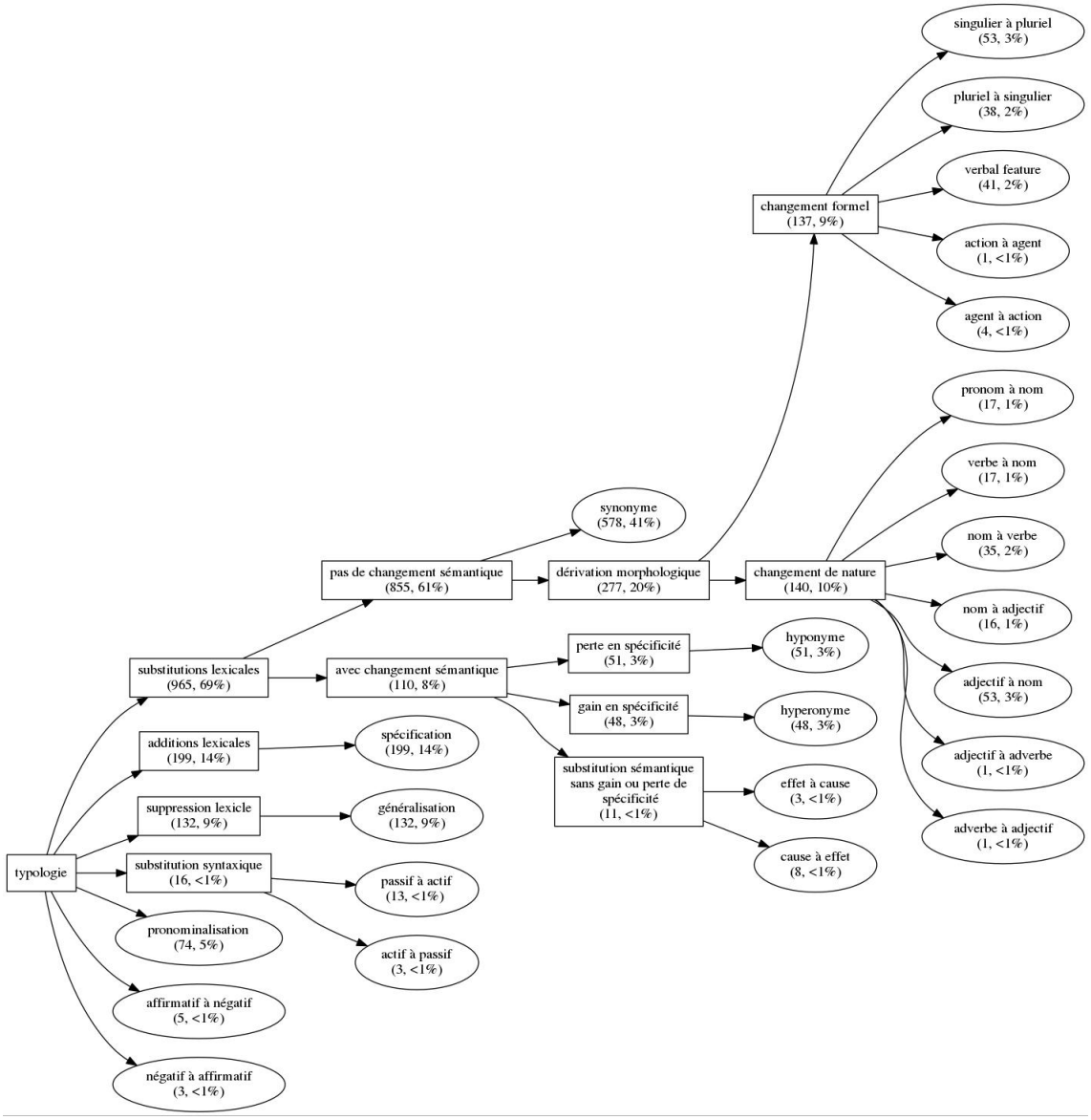


FIGURE 2.3. – Typologie des transformations liées à la simplification.

de leur partie du discours (verbe, nom...). La raison est que le plus souvent, les insertions et les suppressions s'appliquent au niveau de la proposition syntaxique. De plus, nous considérons le changement de partie du discours comme un phénomène de substitution lexicale, que nous décrivons en détails selon les parties du discours impliquées.

— nous nous distinguons de la typologie proposée par Vila *et al.* (2011) car celle-ci est dédiée à la description générale des paraphrases et ne s'intéresse pas spécifiquement aux transformations observables lors de la simplification.

Nous faisons maintenant des observations sur les résultats obtenus. Le plus grand ensemble des transformations (965 occurrences, 69 %) est celui de la substitution lexicale, au sein duquel nous distinguons différentes substitutions avec glissement sémantique (hyponymie et hyperonymie) et sans glissement sémantique (synonymie, transformation morphologique). Ensuite viennent les insertions ou précisions

lexicales (199 occurrences, 14 %), quand des explications sont ajoutées aux termes techniques dans les phrases simplifiées, et les suppressions lexicales ou généralisations (132 occurrences, 9 %), quand les informations sont raccourcies et supprimées lors de la simplification.

tag	*Découpage*	*Regroupement*	*Total*
uncer	0	1	7
syno	24	112	578
hypero	1	10	48
hypo	0	13	51
p2a	1	0	13
pronoun	9	2	74
a2p	0	0	3
p2n	1	3	17
v2n	1	1	17
n2v	2	5	35
n2a	2	0	16
a2n	2	17	53
s2p	0	6	53
p2s	5	3	38
vfea	0	4	41
specif	12	34	199
gener	14	10	132
dupli	0	0	1
adj2adv	0	0	1
adv2adj	0	0	1
act2agt	0	0	1
agt2act	0	0	4
cau2eff	1	1	8
eff2cau	0	0	3
aff2neg	0	1	5
neg2aff	0	0	3

Tableau 2.4. – Fréquence des transformations dans les cas de découpage et de regroupement, et au total.

Le tableau 2.4 indique la fréquence des différents types de transformation selon qu'ils apparaissent dans les phrases découpées, regroupées ou plus généralement dans le corpus (la colonne *Total*). Comme pour la Figure 2.3, les transformations les plus fréquentes sont liées à la synonymie et à la précision ou à la généralisation du contenu. Ces types de transformation sont fréquents dans l'ensemble du corpus et, en conséquence, dans les cas de découpage et de regroupement de phrases. Nous ne voyons pas d'indices d'association entre le découpage ou le regroupement de phrases et l'utilisation de certains types de transformations sémantiques.

Nous pouvons observer quelques substitutions lexicales plus fréquentes dans le processus de simplification :

— Les transformations *a2n* (adjectif → nom) (53 cas) s'expliquent par le fait que les adjectifs sont souvent créés sur des bases d'origine savante (comme *hépat-* dans *hépatique*), alors que le nom, construit sur une base non savante (comme *foie*), est plus commun et donc plus simple à comprendre,
— les cas de substitution par un hyperonyme (48 cas) permettent d'utiliser des mots ayant un sens plus large pour faciliter la compréhension (comme *médicament* pour *clindamicyne*),
— les cas de substitution par un hyponyme (51 cas) peuvent aussi faciliter la compréhension avec une description plus explicite (comme *frissons et tremblements* pour *syndrome pseudo-grippal*),
— les transformations *n2v* (nom → verbe) (35 cas) remplacent un nom abstrait par une action verbale (par exemple *l'arrêt d'un traitement* devient *arrêter un traitement*) et diminuent ainsi le degré d'abstraction de la phrase en la rendant ainsi plus facile à comprendre.

Enfin, notons que les substitutions lexicales dans notre typologie représentent 69 % des transformations (figure 2.3). Les autres travaux évoqués situent la proportion de cette catégorie de transformations autour de 30 ou 40 %. Ceci peut s'expliquer par la spécialisation de notre corpus : le domaine médical utilise un vocabulaire très spécifique qui le rend opaque pour le grand public. S'y rajoute le fait que les changements de nombre sont inclus dans les substitutions lexicales. Cette présence massive des substitutions lexicales représente donc une indication sur la démarche à adopter pour la simplification automatique de textes médicaux : il est nécessaire de disposer de corpus ou de ressources qui permettent de traiter les termes spécifiques au domaine. Cela nous fournit donc une motivation supplémentaire pour la création d'un corpus parallèle spécialisé relevant du domaine médical pour la simplification automatique.

L'analyse syntaxique permet d'associer les informations sémantiques et les informations syntaxiques. Ainsi, dans une minorité de cas (221, 16 %), la nature des groupes syntaxiques reste la même entre la version originale et simplifiée. Dans plusieurs autres cas, le groupe syntaxique de départ est complété par d'autres groupes dans la phrase simple (GN → $[\text{GN GP}]_{\text{GN}}$; GN → $[\text{GN GAdj}]_{\text{GN}}$). 531 transformations visent à modifier des groupes nominaux (*pustulose exanthématique aiguë généralisée > éruption sur la peau pouvant être accompagnée de fièvre*), 190 des groupes prépositionnels (*aux premier et deuxième trimestre de la grossesse > en début de grossesse*) et 174 des groupes verbaux (*peut entraîner l'apparition de > en cas de*). Cela montre que :

— l'analyse syntaxique permet de donner d'importantes indications pour la détection des frontières des séquences à transformer ;
— des mots et expressions de différentes natures syntaxiques peuvent être transformés (noms, verbes, adjectifs...) ;
— les noms et les groupes nominaux, qui correspondent à des concepts, occupent une place importante dans les transformations et sont souvent modifiés lors de la simplification.

2.6. Conclusion

Dans ce chapitre, nous avons décrit la collecte de textes biomédicaux en français et leur compilation en un corpus comparable pour la simplification, le corpus *CLEAR*. L'objectif principal de la constitution de ce corpus comparable est de disposer d'une source de données pour la constitution d'un corpus parallèle. Cette ressource est en effet nécessaire pour les travaux sur la simplification automatique de textes. À cette fin, nous avons procédé à l'alignement manuel d'un échantillon aléatoire pris dans ce corpus pour créer des données de référence. De plus, nous avons annoté les procédés de simplification observés dans ces données de référence avec des informations syntaxiques et sémantiques et proposé une typologie des transformations observées.

Les corpus comparables sont en effet un point de départ incontournable pour les travaux de recherche en simplification. Nous avons ainsi pu voir comment exploiter de tels corpus afin d'analyser la différence de présentation des informations en fonction du public visé. La comparaison de cette analyse et de la typologie des transformations liées à la simplification avec d'autres typologies antérieures nous a permis de révéler des différences liées au domaine médical. Notamment, le point le plus saillant est le taux de substitutions lexicales, qui oscille entre 30 et 40 % dans les typologies des opérations de simplification de langue générale et qui monte à près de 70 % dans le domaine de la santé.

Nos travaux ouvrent des perspectives pour d'éventuels travaux de simplification pour d'autres domaines de spécialité. Nous pouvons illustrer le potentiel d'un tel travail dans d'autres domaines de spécialité avec le domaine juridique. Nous prendrons l'exemple d'une phrase extraite de l'article 203 du code civil belge [15] et sa simplification disponible sur le site de l'entreprise belge Droits Quotidiens [16], qui vise à rendre les informations juridiques accessibles à ses clients. Ainsi, l'exemple (6) montre la présentation de l'information sur ce que couvre la notion d'obligation alimentaire dans le droit belge : la version technique suivie de la version simplifiée. La définition du concept *frais extraordinaires* est plutôt longue. Au niveau syntaxique, la troisième phrase contient des propositions subordonnées juxtaposées et imbriquées, ce que nous n'observons pas dans les données médicales. Nous pouvons également constater que la phrase technique ne présente pas de termes techniques inaccessibles au grand public. Cependant, les notions *hébergement* et *soins de santé* apparaissent dans la version simplifiée mais ne figurent pas dans le code civil original. Nous voyons donc que dans le domaine juridique la simplification nécessite une phase d'interprétation de l'implicite. Ces différentes observations indiquent que les transformations liées à la simplification ont un grand potentiel de variation d'un domaine à l'autre, selon les niveaux linguistiques considérés (syntaxe, lexique, sémantique, pragmatique, etc.). Notre méthode pour l'analyse des procédés de simplification et pour la production d'une typologie des transformations pourrait donc constituer une bonne base pour son adaptation à d'autres domaines de spécialité.

(6) – *Les frais comprennent les frais ordinaires et les frais extraordinaires. Les frais ordinaires sont les frais habituels relatifs à l'entretien quotidien de l'en-*

15. http://www.ejustice.just.fgov.be/cgi_loi/change_lg.pl?language=fr&la=F&cn=1804032130&table_name=loi

16. https://www.droitsquotidiens.be/fr/question/quest-ce-quune-obligation-alimentaire

fant. Par frais extraordinaires, on entend les dépenses exceptionnelles, nécessaires ou imprévisibles qui résultent de circonstances accidentelles ou inhabituelles et qui dépassent le budget habituel affecté à l'entretien quotidien de l'enfant qui a servi de base, le cas échéant, à la fixation des contributions alimentaires.
– On parle d'obligation alimentaire mais l'aide peut couvrir plus que le simple fait de pouvoir se nourrir. Les frais d'hébergement ou de soins de santé, rentrent également dans le cadre de cette aide.

Nous concluons le chapitre par un renvoi vers les tableaux qui présentent les informations quantitatives sur le corpus comparable et sur le corpus parallèle aligné manuellement : tableaux 2.1 et 2.2.

3. Création du corpus de phrases parallèles pour la simplification automatique de textes biomédicaux en français

3.1. Introduction

Ce chapitre présente nos travaux dédiés à la détection automatique de phrases parallèles, avec un contenu sémantique identique ou très similaire, au sein de documents monolingues comparables en français distingués par leur technicité. Un tel ensemble de phrases parallèles alignées est nécessaire pour créer des modèles de simplification automatique de textes. À notre connaissance, le seul travail de ce type en français a été effectué avec un alignement manuel de phrases (Brouwers *et al.*, 2014) mais les phrases alignées et les règles de transformation syntaxiques ne sont pas disponibles.

Dans ce chapitre, nous présentons d'abord les travaux existants autour de l'alignement de phrases à partir de corpus comparables (section 3.2). Nous présentons ensuite notre approche et les résultats que nous obtenons (section 3.3). Nous décrivons également la manière dont nos données ont été valorisées, y compris lors de la compétition DEFT 2020 (sections 3.3.7 et 3.4.3). Nous concluons ce chapitre avec un bilan de ces différents travaux (section 3.5).

Les travaux présentés dans ce chapitre ont fait objet de plusieurs publications dédiées à l'alignement de phrases parallèles à partir de corpus monolingues comparables (Cardon & Grabar, 2018, 2019a,b,c, 2020b,c,d; Cardon *et al.*, 2020b).

3.2. État de l'art

Dans les corpus parallèles, l'alignement de phrases parallèles peut se baser sur des indices de surface comme la longueur relative des phrases (Gale & Church, 1993) ou les informations lexicales (Chen, 1993).

Dans les corpus comparables, en revanche, les phrases ne présentent pas d'identité lexicale mais plutôt une proximité sémantique et, de plus, les informations similaires ne sont pas nécessairement présentées dans le même ordre. Par ailleurs, de tels corpus peuvent contenir des données parallèles à différents niveaux de granularité : documents, phrases, segments sous-phrastiques (Hewavitharana & Vogel, 2011). Dans le domaine de la traduction automatique, les corpus comparables bilingues ont été exploités pour créer des corpus parallèles et alignés. Ces travaux requièrent l'utili-

sation de lexiques bilingues ou de systèmes de traduction automatique et reposent en général sur trois étapes :

1. détection de documents comparables au sein d'un corpus plus large grâce aux métriques de similarité (Utiyama & Isahara, 2003; Fung & Cheung, 2004), ce qui permet de réduire l'espace de recherche de phrases parallèles ;
2. détection de phrases ou de segments candidats à l'alignement en exploitant des systèmes de recherche d'information inter-langue (Utiyama & Isahara, 2003), des arbres d'alignement de séquences (Munteanu & Marcu, 2002) ou des traductions automatiques mutuelles (Yang & Li, 2003; Abdul-Rauf & Schwenk, 2009) ;
3. sélection de bonnes propositions en exploitant des classifieurs binaires (Tillmann & Xu, 2009; Ștefănescu *et al.*, 2012), des mesures de similarité (Fung & Cheung, 2004), le taux d'erreurs (Abdul-Rauf & Schwenk, 2009), des modèles génératifs (Zhao & Vogel, 2002) ou des règles spécifiques (Yang & Li, 2003).

Plus récemment, la recherche de phrases parallèles a été également explorée dans le contexte monolingue : la similarité sémantique textuelle (*semantic text similarity - STS*) est calculée au niveau de phrases ou de segments sous-phrastiques dans une langue donnée. Cette tâche a attiré l'attention des chercheurs car les phrases sémantiquement similaires fournissent des indications précieuses pour la détection du plagiat, l'expansion de requêtes ou les questions-réponses, par exemple. Ainsi, la compétition *SemEval* propose une tâche dédiée au calcul de la similarité sémantique textuelle (Agirre *et al.*, 2013) et poursuit l'objectif suivant : étant donné une paire de phrases, les systèmes automatiques doivent prédire si ces phrases sont similaires sémantiquement et leur assigner un score de similarité allant de 0 (sémantique indépendante) à 5 (sémantique identique). Plusieurs types de méthodes sont exploitées par les participants à cette compétition.

— *Les méthodes basées sur les indices formels*, qui exploitent les chaînes de caractères et de mots (Clough *et al.*, 2002; Zhang & Patrick, 2005; Nelken & Shieber, 2006; Qiu *et al.*, 2006; Zhu *et al.*, 2010a; Zhao *et al.*, 2014). Des descripteurs souvent utilisés sont : l'intersection lexicale globale ou pour une catégorie grammaticale donnée, la longueur des phrases, la distance des chaînes d'édition, les nombres, les entités nommées, la sous-chaîne de caractères commune la plus longue (Clough *et al.*, 2002; Zhang & Patrick, 2005; Nelken & Shieber, 2006; Qiu *et al.*, 2006; Zhu *et al.*, 2010b; Zhao *et al.*, 2014) ;
— *Les méthodes basées sur les ressources lexicales*, qui exploitent des sources lexicales externes, comme WordNet [1] ou la ressource PPDB [2] avec les paraphrases (Miller, 1995; Mihalcea *et al.*, 2006; Fernando & Stevenson, 2008; Lai & Hockenmaier, 2014). Les descripteurs souvent exploités sont : l'intersection avec les ressources, la distance entre les synsets, l'intersection entre les synsets, la similarité sémantique calculée dans les graphes de ces ressources, le paraphrasage connu, les antonymes, les synonymes et les hyperonymes (Mihalcea *et al.*, 2006; Fernando & Stevenson, 2008; Lai & Hockenmaier, 2014) ;

1. `https://wordnet.princeton.edu/`
2. `http://paraphrase.org/#/download`

— *Les méthodes basées sur la syntaxe*, qui exploitent la modélisation syntaxique des phrases. Les descripteurs souvent utilisés sont : les catégories syntaxiques, les dépendances syntaxiques, l'analyse en constituants, les relations prédicat-argument, la distance d'édition entre les arbres syntaxiques (Wan *et al.*, 2006; Severyn *et al.*, 2013; Tai *et al.*, 2015; Tsubaki *et al.*, 2016) ;
— *Les méthodes basées sur les corpus*, qui exploitent les modèles distributionnels, LSA, etc. Les descripteurs souvent exploités sont : les plongements lexicaux ou syntaxiques, calculés sur les corpus traités ou bien sur les corpus de référence et les *topic models* (Barzilay & Elhadad, 2003; Guo & Diab, 2012; Zhao *et al.*, 2014; Kiros *et al.*, 2015; He *et al.*, 2015; Mueller & Thyagarajan, 2016).

Ces différentes méthodes peuvent aussi être combinées pour optimiser les résultats (Bjerva *et al.*, 2014; Lai & Hockenmaier, 2014; Zhao *et al.*, 2014; Rychalska *et al.*, 2016). Notons aussi que l'exploitation des données de référence préparées pour les compétitions *SemEval* constituent une plateforme de test libre et sont également exploitées par des chercheurs en dehors de la compétition (Severyn *et al.*, 2013; Kiros *et al.*, 2015; He *et al.*, 2015; Tsubaki *et al.*, 2016; Mueller & Thyagarajan, 2016), ce qui permet de stimuler les travaux sur cette question de recherche.

Soğancıoğlu *et al.* (2017) ont travaillé sur l'estimation de la similarité sémantique dans le domaine médical. Deux approches principales sont testées : des mesures de similarité sémantique basées sur les indices formels, et des mesures basées sur des ressources externes, en l'occurrence WordNet (Miller, 1995) et l'UMLS (Lindberg *et al.*, 1993). Les mesures de similarité classiques sont exploitables pour le français. Il est plus délicat de reproduire l'approche qui exploite WordNet et l'UMLS, ces deux ressources – et plus particulièrement l'UMLS – n'ayant pas d'équivalent aussi complet en français.

Finalement, parfois les travaux se positionnent dans d'autres contextes et poursuivent des tâches légèrement différentes, comme la détection de phrases parallèles sans évaluer leur degré de similarité (Barzilay & Elhadad, 2003; Nelken & Shieber, 2006; Zhu *et al.*, 2010b).

Nous voyons donc que la détection de phrases parallèles, ou de phrases sémantiquement similaires, dans des corpus monolingues comparables est une tâche qui attire de plus en plus l'attention des chercheurs. Nous nous intéressons à cette tâche car elle permet de construire des ressources (par exemple un lexique ou des règles de transformation) ou d'entraîner des modèles d'apprentissage automatique utilisables en simplification automatique.

3.3. Méthodologie pour l'alignement de phrases parallèles

Nous exploitons le corpus monolingue, dont la création est décrite au chapitre 2, et une liste de mots grammaticaux, qui comprend au total 83 entrées, telles que les déterminants ou les prépositions.

Dans notre corpus, les documents comparables sont déjà associés entre eux. En revanche, comme les textes techniques et simplifiés sont souvent rédigés de manière indépendante, l'ordre des phrases dans les documents n'est pas significatif. L'accent

principal de la méthode est donc mis sur la recherche de phrases parallèles. Dans ce chapitre, nous proposons des méthodes pour effectuer la détection et l'alignement de phrases parallèles.

Notre corpus de départ est un corpus comparable. Cela implique deux contraintes importantes pour la tâche d'alignement automatique de phrases.

1. Les informations ne sont pas identiques : il n'est donc pas certain que chaque phrase d'un document puisse être alignée avec une phrase dans un autre document.
2. L'organisation des informations étant différente dans deux documents comparables, nous ne pouvons pas compter sur l'ordre dans lequel les informations apparaissent pour chercher des alignements.

Ces deux contraintes nous poussent à examiner l'ensemble de la combinatoire des phrases des deux documents pour y rechercher des alignements. Cela résulte en un déséquilibre considérable :

— Selon le tableau 2.2 page 26, à partir de 39 couples de documents, nous avons 266 couples de phrases alignées manuellement ;
— Lorsque nous produisons la combinatoire possible de phrases à partir des 39 couples de documents utilisés pour l'alignement manuel, nous obtenons plus d'un million de paires de phrases ;
— Comme résultat, pour chaque paire alignable nous avons environ 4 400 paires non alignables.

Le déséquilibre des données est une problématique connue (Zhang & Zweigenbaum, 2017), quoique peu traitée dans la littérature. Nous proposons donc de prendre en compte cette problématique de déséquilibre de données dans nos travaux.

Notre méthode se compose de plusieurs étapes : (1) le pré-traitement (section 3.3.1), qui inclut une étape de filtrage de phrases pour réduire le déséquilibre, (2) l'alignement de phrases (section 3.3.2) et (3) l'évaluation des alignements (section 3.3.3).

Le schéma en figure 3.1 présente les différentes étapes de la méthode d'alignement.

Nous décrivons également les différentes expériences effectuées (section 3.3.4) et les résultats obtenus (sections 3.3.5 et 3.3.6). Nous présentons un cas de valorisation de nos données : la tâche 2 de la campagne DEFT 2020 (section 3.3.7). Nous terminons enfin par une discussion sur les limites du travail effectué et sur les perspectives ouvertes (section 3.3.8).

3.3.1. Pré-traitement

Tous les documents sont étiquetés avec TreeTagger (Schmid, 1994), ce qui permet d'en obtenir leurs versions lemmatisées. Les documents sont ensuite segmentés en phrases en exploitant la ponctuation forte (*. ? ! ; :*). Nous procédons également à une étape de filtrage de phrases dans le but d'éliminer un maximum de paires de phrases qui ne représentent pas de bons candidats à l'alignement. Nous exploitons trois méthodes pour le filtrage, basées sur la forme et la syntaxe :

1. Méthode basée sur le nombre de mots dans les phrases. Chaque phrase candidate doit contenir au moins cinq mots, ce qui correspond à la longueur de la phrase la plus courte dans les données de référence ;

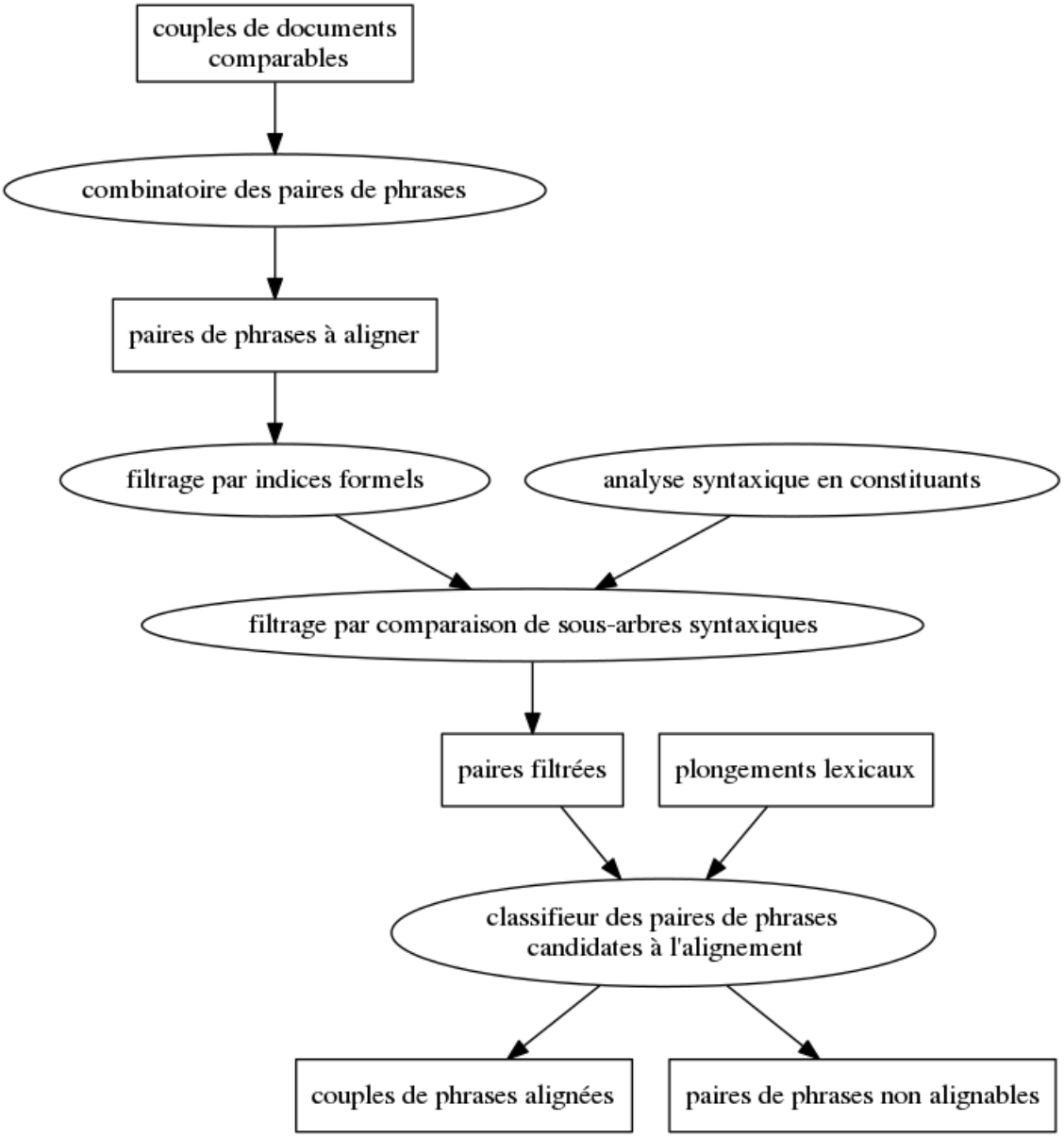


FIGURE 3.1. – Vision d'ensemble de la méthode d'alignement.

2. Suppression de paires avec les phrases identiques ;

3. Exploitation d'informations syntaxiques. Le premier critère pour retenir une paire de phrase est la présence d'un verbe conjugué dans chacune des deux phrases. Nous nous inspirons ensuite d'un travail existant qui mesure la similarité entre les phrases dans un corpus monolingue grâce aux constituants syntaxiques (Duran *et al.*, 2014). Le score de similarité est alors calculé sur la base des nœuds syntaxiques similaires qui contiennent des mots similaires. Il est difficile de transposer cette méthode à notre travail pour deux raisons : (1) la table de similarité entre les constituants est créée pour l'anglais et (2) les auteurs ne fournissent pas d'indications sur les principes de sa création. Nous supposons cependant que l'adoption d'une approche similaire permettra

d'éliminer les paires de phrases indésirables pour l'alignement. Ainsi, au lieu de calculer le score de similarité, nous effectuons un filtrage binaire : garder ou non une paire de phrases candidates. Ainsi, pour une paire donnée, nous calculons l'arbre syntaxique de chacune des phrases. Ensuite, nous comparons les feuilles (i.e. mots) des arbres, à l'exception de celles qui contiennent un mot de la liste des 83 mots grammaticaux. Lorsque nous trouvons deux mots identiques, nous vérifions leurs nœuds pères : s'ils sont identiques, nous gardons la phrase comme candidate à l'alignement. Le processus est illustré par l'algorithme 1 ci-dessous. Nous exploitons également une variante de la méthode : au lieu de nous arrêter lorsque les nœuds pères ne sont pas identiques, nous continuons de remonter l'arbre jusqu'au troisième nœud tant que les nœuds précédents n'ont pas donné de résultats positifs. La comparaison s'arrête lorsque les nœuds sont identiques et la phrase est retenue pour l'alignement ou lorsque la profondeur est supérieure à 3. Cette approche est illustrée par l'algorithme 2, page 45. La considération de nœuds parents de profondeur 3 permet également d'observer comment la profondeur de l'arbre influence le filtrage. L'analyse syntaxique des phrases est obtenue avec le Berkeley Neural Parser et le modèle de langue pour le français avec la librairie python `benepar` (Kitaev & Klein, 2018). Nous utilisons la librairie NLTK `Tree` pour la manipulation des arbres syntaxiques (Bird *et al.*, 2009).

Data: Deux arbres syntaxiques (T_1 et T_2), une liste de *mots grammaticaux* (*SW*)
Result: Booléen
Booléen ← False;
if *au moins un verbe est dans chaque arbre* **then**
  foreach *feuille de* T_1 *(*L_1*) absente de SW* **do**
    foreach *feuille de* T_2 *(*L_2*) absente de SW* **do**
      if L_1 *est identique à* L_2 **then**
        if *l'étiquette du père de* L_1 *est identique à l'étiquette du père de* L_2 **then**
          Booléen ← True ;
        else
          rien ;
        end
      else
        rien ;
      end
    end
  end
else
  rien ;
end
return Booléen ;

Algorithm 1: Méthode de filtrage par la comparaison des pères immédiats des feuilles

```
Data: Deux arbres syntaxiques (T1 et T2), une liste de mots grammaticaux
      (SW)
Result: Booléen
Booléen ← False;
if au moins un verbe est dans chaque arbre then
    foreach feuille de T1 (L1) absente de SW do
        foreach feuille de T2 (L2) absente de SW do
            if L1 est identique à L2 then
                if l'étiquette du père de L1 (P1) est identique à l'étiquette du
                  père de L2 (P2) then
                    Booléen ← True ;
                else
                    if l'étiquette du père de P1 (PP1) est identique à l'étiquette
                      du père de P2 (PP2) then
                        Booléen ← True;
                    else
                        if l'étiquette du père de PP1 est identique à l'étiquette
                          du père de PP2 then
                            Booléen ← True;
                        else
                            rien;
                        end
                    end
                end
            else
                rien ;
            end
        end
    end
else
    rien ;
end
return Booléen ;
```

Algorithm 2: Méthode de filtrage par la comparaison des ancêtres des feuilles, jusqu'à profondeur 3

3.3.2. Alignement de phrases

Nous abordons la recherche de phrases parallèles comme une problématique de catégorisation : pour une paire de phrases conservée après le filtrage, il faut décider si elle relève de la catégorie *aligné* ou non. Nous utilisons plusieurs classifieurs linéaires de `scikit-learn` (Pedregosa *et al.*, 2011) avec leurs paramètres par défaut, s'il n'est pas indiqué autrement : `Perceptron` (Rosenblatt, 1958), `Perceptron multicouche` (`MLP`) (Rosenblatt, 1961), `Random Forest` (`RF`) (Ho, 1995) `Linear discriminant analysis` (`LDA`) (Fisher, 1936) avec le solveur `LSQR`, `Quadratic discriminant analysis` (`QDA`) (Cover, 1965), `Logistic regression` (Berkson, 1944), modèle log-linéaire appris avec `Stochastic gradient descent` (`SGD`) (Ferguson, 1982), et `SVM linéaire` (Vapnik & Lerner, 1963).

Afin d'avoir une méthode assez générique et transposable à d'autres jeux de données, nous utilisons des descripteurs qui sont facilement calculables. Nous exploitons cinq types de descripteurs. Par rapport aux travaux cités dans la section 3.2, ces descripteurs s'appuient essentiellement sur des comparaisons prenant en compte les indices formels ainsi que des critères de ressemblance lexicale. Les descripteurs sont calculés sur les formes et les lemmes :

1. *BL : Descripteurs de base (baseline)* :
   — *Nombre de mots communs, hors mots grammaticaux*, ce qui permet de calculer l'intersection lexicale de base entre les phrases (Barzilay & Elhadad, 2003) ;
   — *Ratio longueur de la phrase la plus courte sur la longueur de la phrase la plus longue*. Ce descripteur suppose que la simplification peut impliquer une association stable avec la longueur des phrases ;
   — *Différence de la longueur moyenne des mots entre les deux phrases* pour estimer l'utilisation de mots longs, jugés spécifiques au langage technique ;
2. *L : Descripteurs issus de la distance de chaînes d'édition* (Levenshtein, 1966) :
   — *Distance d'édition calculée au niveau des caractères*. Il s'agit de l'acception classique de la mesure. Elle prend en compte les opérations d'édition de base (insertion, suppression et substitution). Le coût de chaque opération est de 1 ;
   — *Distance d'édition calculée au niveau des mots*. Ce descripteur est calculé avec des mots comme unité. Il prend en compte les mêmes opérations d'édition avec le coût de 1. Le descripteur permet de calculer le coût de la transformation lexicale ;
3. *S : Descripteurs basés sur les similarités lexicales* avec la *similarité au niveau des mots calculée selon trois scores (cosinus, Dice et Jaccard)*. Ce descripteur fournit une indication plus sophistiquée sur l'intersection lexicale entre les deux phrases. Le poids de chaque mot est de 1 ;
4. *N : Descripteurs basés sur les n-grammes (bigrammes et trigrammes) de caractères en commun*, ce qui permet de prendre en compte la présence de séquences de caractères communs ;
5. *PL : Descripteurs basés sur les plongements lexicaux.* Deux descripteurs sont utilisés :

— *WAVG* (Štajner *et al.*, 2018) : la moyenne des vecteurs de mots de chacune des deux phrases est calculée et ensuite ces vecteurs sont comparés pour attribuer un score de similarité ;
— *CWASA* (Franco-Salvador *et al.*, 2016) pour *continuous word alignment-based similarity analysis*.

Les descripteurs *PL* sont exploités avec des plongements lexicaux entraînés sur le corpus CLEAR à l'aide de Word2Vec [3] (Mikolov *et al.*, 2013a), alors que les scores sont calculés avec l'outil CATS (Štajner *et al.*, 2018).

3.3.3. Évaluation

L'évaluation est effectuée par rapport aux données de référence. L'entraînement du système est effectué sur 70 % de paires de phrases et le test est effectué sur le reste des données. Plusieurs classifieurs et plusieurs combinaisons de descripteurs sont testés. Les mesures d'évaluation classiques suivantes pour un classifieur binaire sont calculées (dans les formules, VP = vrais positifs ; FP = faux positifs ; VN = vrais négatifs, FN = faux négatifs) :

— Vrais positifs : le nombre d'exemples correctement classés
— Précision : $\frac{VP}{VP+FP}$
— Rappel : $\frac{VP}{VP+FN}$
— F-mesure : $2 \times \frac{VP}{VP+FP+FN}$
— Erreur quadratique moyenne (EQM) : $\frac{1}{n}\sum_{i=1}^{n} e_i^2$, où e_i est la différence entre le résultat du classifieur et la valeur de référence (ici 1 pour les phrases alignées et -1 pour les phrases non alignées) pour une paire de phrases i, et n le nombre total d'exemples.

Avec les données déséquilibrées, l'évaluation est effectuée sur 50 tirages différents afin de mieux évaluer les performances de l'alignement des phrases. Nous rapportons uniquement les scores pour la catégorie *aligné* pour deux raisons :

— c'est la catégorie qui nous intéresse,
— avec les données déséquilibrées et une très grande quantité de phrases non alignables, les résultats globaux sont toujours très élevés car les phrases non alignables sont très bien détectées dans leur majorité. Ainsi, rapporter les scores pour les deux catégories ne nous permettrait pas d'interpréter les résultats.

3.3.4. Expériences

Les données de référence fournissent 266 couples de phrases parallèles comme exemples positifs et nous choisissons aléatoirement des exemples négatifs à partir des mêmes documents : 266 paires de phrases non parallèles pour les expériences avec des données équilibrées et d'autres paires de phrases pour des expériences avec des données non équilibrées. Les exemples négatifs sont obtenus en appariant aléatoirement des phrases d'un document technique et de son pendant simple et en vérifiant que ces paires ne font pas partie de la classe positive (équivalence ou inclusion). Il

3. Hyperparamètres de Word2Vec : `-size 300 -window 7 -sample 1e-5 -hs 1 -negative 50 -mincount 20 -alpha 0.025 -cbow 0`

n'y a pas d'intersection entre les phrases alignées et non alignées. Plusieurs expériences sont effectuées, où nous étudions les effets des descripteurs et du déséquilibre. Toutes les expériences sont effectuées sur les données après le pré-traitement décrit en section 3.3.1.

3.3.4.1. Baseline

Notre *baseline* correspond à la combinaison de descripteurs correspondant à des indices formels traditionnellement utilisés pour l'alignement de phrases : la longueur des phrases et l'intersection lexicale entre les phrases. Cette expérience est effectuée avec les données équilibrées.

3.3.4.2. Détection de phrases parallèles avec une distribution équilibrée

Le nombre d'exemples parallèles et non parallèles est comparable, ce qui correspond à une distribution équilibrée des paires de phrases entre les deux catégories. Avec cette expérience, nous testons différents jeux de descripteurs et leurs combinaisons.

3.3.4.3. Détection de phrases parallèles selon la sémantique des paires

Les couples de phrases de référence sont divisées en deux sous-ensembles, selon le lien sémantique qui existe au sein du couple :

— E : 136 couples avec équivalence sémantique,
— I : 130 couples où le contenu de la phrase technique est compris dans la phrase simplifiée ou l'inverse. Ceci représente les cas de découpage ou de regroupement de phrases, ainsi que la suppression ou l'ajout d'information, lors de la simplification.

Nous utilisons tous les descripteurs pour cette expérience.

3.3.4.4. Détection de phrases parallèles avec une distribution déséquilibrée

Comme le montre le tableau 2.2 page 26, les phrases parallèles sont largement minoritaires et il existe beaucoup plus de phrases non alignables. La distribution de phrases parallèles n'est donc pas élevée ni constante : le taux d'alignement varie selon les corpus, les couples de documents et le sens d'alignement. Ainsi, l'objectif de cette expérience est de voir quelles sont les performances du système lorsque les données traitées s'approchent de la distribution naturelle de phrases alignables. Pour chaque sous-ensemble (E, I), nous prenons d'abord autant de couples équivalents que d'exemples négatifs sélectionnés aléatoirement. Ensuite, nous augmentons progressivement le nombre de paires non alignables jusqu'au ratio 200 :1, proche de celui des données réelles après le filtrage. Ceci correspond à l'ensemble déséquilibré D avec les 136 (E) ou 130 I couples alignés et le ratio croissant de paires non alignées. Le ratio et les données changent donc à chaque itération. Nous utilisons aussi l'ensemble réel R, qui comporte toute la combinatoire possible de paires de phrases après filtrage (21 428), alignées et non alignées. L'ensemble R est toujours le même. Nous procédons ainsi en raison du faible nombre d'exemples positifs. Il est donc à noter que le score de rappel en sera artificellement augmenté. Cela dit, le score

de précision évalue la robustesse du modèle à ne pas produire de faux positifs, ce qui nous semble important en raison du grand déséquilibre en faveur des exemples négatifs. À chaque point de déséquilibre de l'ensemble D, nous faisons deux séries d'expériences :

1. DD : entraînement et test au sein de l'ensemble déséquilibré D ;
2. DR : entraînement sur l'ensemble déséquilibré D et test sur les données réelles R (environ 21 428 paires de phrases après filtrage).

Nous utilisons tous les descripteurs pour cette expérience.

3.3.5. Résultats

Nous présentons les résultats des différentes expériences décrites en section 3.3 :
(1) l'effet du filtrage sur les paires à aligner (section 3.3.5.1),
(2) les performances de différents classifieurs binaires sur les données équilibrées (section 3.3.5.2),
(3) la méthode de *baseline* pour l'alignement de phrases avec l'utilisation de descripteurs basiques (section 3.3.5.3),
(4) l'examen des descripteurs utilisés pour la détection de phrases parallèles avec une distribution équilibrée de phrases parallèles et non parallèles pour l'entraînement (section 3.3.5.4),
(5) la détection de phrases parallèles selon la sémantique des paires en distinguant l'équivalence sémantique et l'inclusion (section 3.3.5.5),
(6) la détection de phrases parallèles avec une distribution déséquilibrée s'approchant de la distribution réelle moyenne des phrases alignables (section 3.3.5.6).

3.3.5.1. Pré-traitement

La première colonne du tableau 3.1 indique le nombre de paires de phrases originales, la seconde le nombre de paires qui restent après l'utilisation des indices formels liés à la présence du verbe et l'élimination des paires avec des phrases identiques (IF), et les deux dernières colonnes indiquent le nombre de paires qui restent après l'utilisation du filtre syntaxique, en remontant respectivement au premier (colonne *Syntaxe 1*) et au troisième (colonne *Syntaxe 3*) père. Les indices formels sont appliqués avant les filtres syntaxiques. Les filtres syntaxiques sont appliqués indépendamment l'un de l'autre.

Paires restantes	*Original*	*IF*	*Syntaxe 1*	*Syntaxe 3*
Total	1 164 407	409 530	16 879	21 428
Equivalent	136	136	94	94
Inclusion	130	130	94	100

Tableau 3.1. – Effet du filtrage sur les paires de phrases à aligner.

Nous observons que les indices formels réduisent le nombre total de paires de 65 % : on passe de 1 164 407 à 409 530 paires de phrases. Nous voyons qu'avec ces indices

nous ne perdons aucun exemple positif. À partir des 409 530 paires obtenues après le premier filtre, nous observons une autre grande réduction du volume de paires avec chacun des deux filtres syntaxiques. Le filtre de profondeur 1 laisse 16 879 paires (∼96 % de réduction) et celui de profondeur 3 laisse 21 428 paires (∼95 % de réduction). Le défaut de ce type de filtres est qu'un nombre non négligeable d'exemples positifs est perdu. 42 des 136 (∼30 %) cas d'équivalence ne passent pas les deux filtres syntaxiques. Sur les 130 les cas d'inclusion, 36 (∼27 %) ne passent pas le filtre de profondeur 1, et 32 (∼24 %) ne passent pas le filtre de profondeur 3. Nous présentons deux exemples pour illustrer les résultats du filtrage. L'exemple (1) est correctement conservé alors que l'exemple (2) est rejeté à tort après filtrage IF et Syntaxe 3. Ce rejet s'explique par la différence d'accent entre *capecitabine* et *capécitabine* qui en fait deux mots différents. Cette différence est présente dans les documents d'origine.

(1) *- L'apparition de signes cliniques tels qu'un mal de gorge, une fièvre, une pâleur, un purpura ou un ictère pendant le traitement par la sulfasalazine peut faire suspecter une myélosuppression, une hémolyse ou une hépatotoxicité.*
- L'apparition de signes cliniques tels qu'un mal de gorge, une fièvre, une pâleur, de petites taches rouges sur la peau ou une jaunisse pendant le traitement par la sulfasalazine peut faire suspecter une diminution du nombre de cellules du sang, une destruction des globules rouges ou une toxicité du foie.

(2) *- L'allaitement doit être interrompu en cas de traitement par capécitabine.*
- Vous ne devez pas allaiter si vous êtes traitée par capecitabine eg.

Avec ces deux exemples, nous voyons que les phrases avec des substitutions lexicales, comme *{hémolyse}{destruction des globules rouges}*, peuvent être conservées. Rappelons que le filtrage syntaxique décrit en section 3.3.1 ne conserve que des paires de phrases qui ont au moins un mot en commun dans des sous-arbres syntaxiques de même nature. Bien que ce filtrage laisse de côté les phrases simplifiées sans mot commun avec la phrase d'origine, nous voyons que nous pouvons néanmoins retenir des couples de phrases qui comportent des substitutions lexicales. Ceci est important car nous avons observé que la substitution lexicale représente environ 70 % des transformations dans les textes médicaux (section 2.5.2.2 page 32). En revanche, les transformations syntaxiques, comme la modification de parties du discours *{allaitement}{allaiter}* ou de la voix du verbe *{passive}{active}*, sont plus difficiles à conserver avec le filtrage syntaxique. Pour ces cas, un travail plus poussé sur les structures syntaxiques comparables et la morphologie sera nécessaire.

Tous les traitements décrits ci-après sont effectués sur les données filtrées avec les indices formels (présence d'un verbe et suppression des phrases identiques) car ils ne laissent aucun exemple positif de côté. Nous conservons également le filtre syntaxique car il réduit grandement le déséquilibre des données, ce qui est l'effet recherché. Nous avons préféré garder le filtre syntaxique de profondeur 3. Comme le montre le tableau 3.1, les deux filtres syntaxiques conservent le même nombre d'exemples positifs pour l'équivalence et réduisent le nombre d'exemples négatifs du même ordre de grandeur. Bien que le filtre de profondeur 1 ait conservé un peu moins d'exemples négatifs, nous choissions le filtre de profondeur 3 car nous avons observé qu'il conservait un peu plus d'exemples positifs pour l'inclusion. Même si

nous ne l'avons pas observé avec nos données de test, nous pensons donc que le filtre de profondeur 3 peut potentiellement obtenir plus d'exemples positifs, sans montrer d'inconvénients majeurs par rapport au filtre de profondeur 1.

3.3.5.2. Alignement de phrases parallèles

classifieur	*P*	*R*	*F1*	*EQM*	*VP*
`Perceptron`	0,90	0,93	0,92	0,08	28
`MLP`	0,93	0,93	0,93	0,06	28
`RF`	**1.00**	**0.97**	**0.98**	**0.02**	**29**
`LDA`	0,93	0,87	0,90	0,09	26
`QDA`	0,96	0,90	0,93	0,06	27
`LogReg`	0,97	**0,97**	0,97	0,03	**29**
`SGD`	0,90	0,93	0,92	0,08	28
`LinSVM`	0,97	0,93	0,95	0,04	28

Tableau 3.2. – Résultats d'alignement d'après les différents classifieurs, avec l'ensemble des descripteurs, ratio des classes 1 : 1. Titres de colonnes : précision (P), rappel (R), Erreur Quadratique Moyenne (EQM), Vrais Positifs (VP).

Dans cette expérience, nous exploitons tous les descripteurs sur l'ensemble de test avec le texte non lemmatisé. L'objectif est de vérifier quel algorithme d'apprentissage est le plus performant dans cet environnement pour savoir lequel conserver par la suite des expériences. Cette expérience est effectuée avec des données équilibrées. Les résultats globaux se trouvent dans le tableau 3.2. Les résultats sont présentés en termes de rappel R, précision P, F-mesure F, erreur quadratique moyenne EQM et vrais positifs VP (sur un total de 30 couples de phrases alignées dans l'ensemble de test).

Nous pouvons voir que tous les classifieurs testés sont compétitifs avec une F-mesure entre 0,92 et 0,98. Pour tous les classifieurs, nous indiquons les scores moyens de 20 itérations. La précision et le rappel sont équilibrés entre eux. `Random Forest` semble être le meilleur classifieur : F-mesure de 0,98 (précision 1 et rappel 0,97), le plus grand nombre de vrais positifs (56) et l'erreur quadratique moyenne la plus faible (0,02). `Régression Logistique` est presqu'aussi performant avec 0,97 de précision, rappel et F-mesure. Les expériences qui suivent sont effectuées avec `Random Forest`.

3.3.5.3. Baseline

Pour la *baseline*, nous exploitons les descripteurs le plus souvent utilisés dans les travaux existants : longueur des phrases et intersection lexicale entre les phrases. Les données exploitées sont équilibrées. Les résultats sont présentés dans la première ligne du tableau 3.3 : nous obtenons une F-mesure de 0,95 (avec un rappel de 0,97 et une précision de 0,93), une erreur quadratique moyenne de 0,05 et 28 vrais positifs

sur 30 dans les données de test. Cela indique que les descripteurs traditionnels sont en effet assez efficaces pour cette tâche.

3.3.5.4. Détection de phrases parallèles avec une distribution équilibrée

descripteurs	*R*	*P*	*F1*	*EQM*	*VP*
BL	0,97	0,93	0,95	0,05	28
S	0,97	0,97	0,97	0,03	29
L	0,90	0,93	0,92	0,09	28
N	0,97	0,93	0,95	0,05	28
PL	0,97	0,97	0,97	0,03	29
L+S	1,00	0,93	0,97	0,03	28
L+N	1,00	0,97	0,98	0,02	29
L+PL	0,97	0,97	0,97	0,03	29
S+N	1,00	0,97	0,98	0,02	29
S+PL	1,00	0,97	0,98	0,02	29
BL+L	1,00	0,97	0,98	0,02	29
BL+S	1,00	0,97	0,98	0,02	29
BL+N	1,00	0,97	0,98	0,02	29
BL+PL	1,00	0,97	0,98	0,02	29
N+PL	1,00	0,97	0,98	0,02	29
BL+L+S	1,00	0,97	0,98	0,02	29
BL+L+N	1,00	0,97	0,98	0,02	29
BL+L+PL	1,00	0,97	0,98	0,02	29
BL+S+N	1,00	0,97	0,98	0,02	29
BL+S+PL	1,00	0,97	0,98	0,02	29
BL+N+PL	1,00	0,97	0,98	0,02	29
L+S+N	1,00	0,97	0,98	0,02	29
L+S+PL	1,00	0,97	0,98	0,02	29
L+N+PL	1,00	0,97	0,98	0,02	29
BL+L+S+N	1,00	0,97	0,98	0,02	29
BL+L+S+PL	1,00	0,97	0,98	0,02	29
BL+L+N+PL	1,00	0,97	0,98	0,02	29
BL+S+N+PL	1,00	0,97	0,98	0,02	29
L+S+N+PL	1,00	0,97	0,98	0,02	29
BL+L+S+N+PL	1,00	0,97	0,98	0,02	29

TABLEAU 3.3. – Résultats d'alignement : différents ensembles de descripteurs, `Random Forest`, ratio des classes 1 : 1. Titres de colonnes : précision (P), rappel (R), Erreur Quadratique Moyenne (EQM), Vrais Positifs (VP). Titres de rangées : baseline (BL), similarité (S), Levenshtein (L), plongements lexicaux (PL).

L'objectif de cette expérience est d'évaluer la performance de différents jeux de descripteurs dans la tâche d'alignement de phrases, ainsi que de leurs combinaisons.

Le tableau 3.3 présente les résultats de la détection de phrases parallèles avec une distribution équilibrée des données. Les meilleurs résultats sont obtenus par l'ensemble S (mesures de similarité) avec une F-mesure de 0,97. Les moins bons résultats sont obtenus avec l'ensemble L (distance de Levenshtein) avec une F-mesure de 0,92. Ils restent néanmoins élevés. Les différentes combinaisons de descripteurs permettent d'améliorer ces résultats, ce qui indique que chaque type de descripteurs apporte des informations complémentaires. La plupart des combinaisons atteignent les résultats les plus élevés : rappel 1, précision 0,97, F1 0,98, erreur quadratique moyenne 0,02 et 29 vrais positifs.

3.3.5.5. Détection de phrases parallèles selon la sémantique des couples avec des données équilibrées

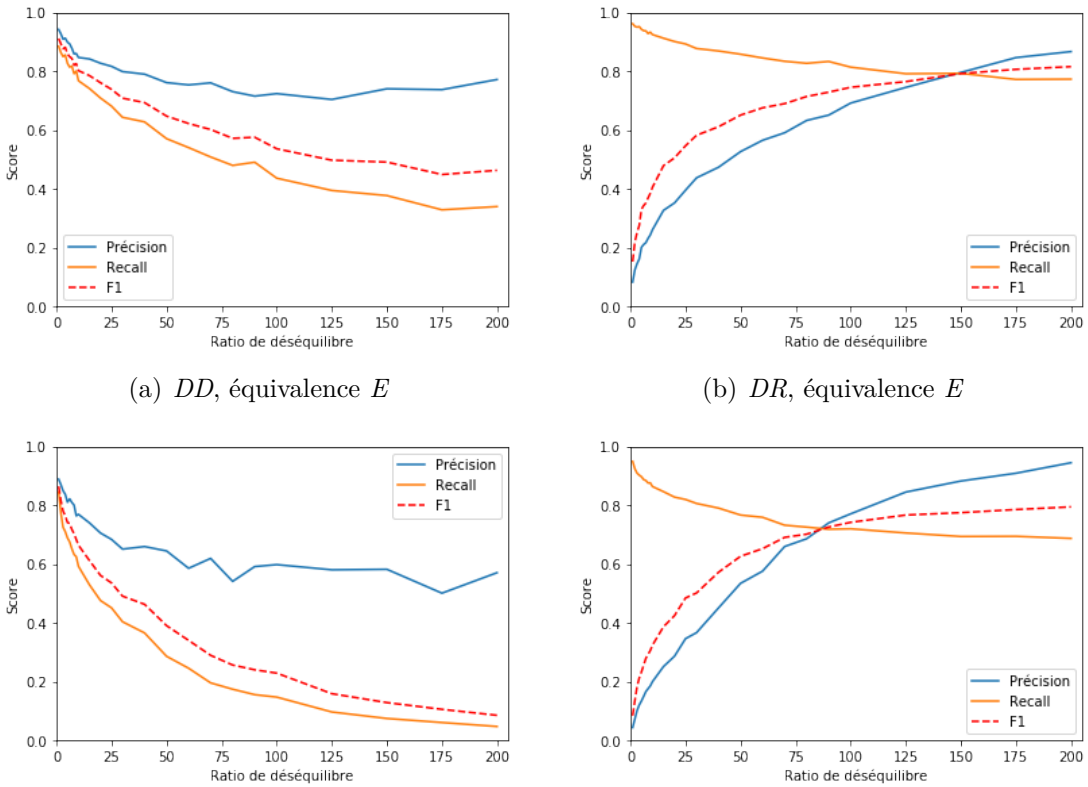
Dans cette série d'expériences avec des données équilibrées, nous voulons voir s'il existe une différence selon le type de relation sémantique des couples de phrases. Dans l'ensemble de test, nous comptons 30 couples en équivalence et 31 en inclusion. Selon le tableau 3.4, il existe une légère différence : il est un peu plus facile de détecter les phrases en relation d'équivalence que les phrases en relation d'inclusion. Nous supposons que les couples d'inclusion couvrent une plus grande variété de situations, ce qui est plus difficile à modéliser avec le faible volume de données dont nous disposons.

Ensemble	*P*	*R*	*F1*	*EQM*	*VP*
Équivalence E	1,00	0,97	0,98	0,02	29
Inclusion I	1,00	0,94	0,97	0,03	29

Tableau 3.4. – Résultats d'alignement : les deux ensembles de données équilibrées (l'équivalence sémantique et les inclusions), ensemble de test, tous les descripteurs, `Random Forest`, ratio des classes 1 : 1. Titres de colonnes : précision (P), rappel (R), Erreur Quadratique Moyenne (EQM), Vrais Positifs (VP).

3.3.5.6. Détection de phrases parallèles avec une distribution déséquilibrée

Comme les documents comparables peuvent contenir un taux variable de phrases parallèles, nous faisons des tests avec des données déséquilibrées. Nous testons différents taux de déséquilibre. Les résultats sont présentés à la figure 3.2 : l'axe x représente l'augmentation du déséquilibre (seule la première position 1 correspond aux données équilibrées), l'axe y représente les scores de précision, rappel et F-mesure. Les résultats pour les deux ensembles sont présentés : équivalence (figures 3.2(a) et 3.2(b)) et inclusion (figures 3.2(c) et 3.2(d)). Les figures de gauche présentent les résultats DD : l'entraînement et le test sont effectués sur des données avec le même rapport de déséquilibre D. Les figures de droite présentent les résultats DR obtenus par les mêmes modèles entraînés sur les données déséquilibrées D mais testés sur l'ensemble des données R (toutes les paires de phrases possibles). Les résultats présentés sont les moyennes de 50 itérations.



(a) *DD*, équivalence *E*

(b) *DR*, équivalence *E*

(c) *DD*, inclusion *I*

(d) *DR*, inclusion *I*

FIGURE 3.2. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, avec l'intégralité des descripteurs.

Nous pouvons faire plusieurs observations sur ces expériences avec des données déséquilibrées. Comme indiqué dans la section précédente, les paires équivalentes (figures 3.2(a) et 3.2(b)) sont plus faciles à catégoriser que les inclusions. Les scores de précision et de rappel sont alors plus élevés à différents points de déséquilibre. Ce résultat est positif car les phrases équivalentes fournissent les informations les plus utiles et complètes sur les transformations requises lors de la simplification. Sans surprise, l'augmentation du déséquilibre mène vers des performances réduites durant l'entraînement. Cela signifie que le déséquilibre crée de la confusion entre les paires alignables et non alignables. Cependant, pour atteindre notre objectif, qui consiste à identifier le peu d'exemples positifs présents dans une masse de paires non alignables, il vaut mieux utiliser un modèle plus robuste face au déséquilibre des données, même s'il fonctionne moins bien à l'entraînement.

En annexe A page 121, nous présentons les figures correspondant aux mêmes expériences pour chacun des ensembles de descripteurs. Cela montre le comportement des descripteurs en situation de déséquilibre, là où le tableau 3.3 le montrait en situation d'équilibre. Nous pouvons ainsi voir que les observations sont comparables : les descripteurs de type *L* (figure A.3, page 122) obtiennent les moins bonnes performances, avec une F-mesure qui monte puis stagne autour de 0,35 à partir du

déséquilibre 1 : 75. Les descripteurs S ainsi que les descripteurs N ont le même comportement avec une F-mesure respectivement légèrement supérieure à 0,60 (figure A.2, page 121) et légèrement inférieure à 0,60 (figure A.4, page 122). Les meilleurs descripteurs sont de types BL (figure A.1, page 121) et PL (figure A.5), page 121 avec une F-mesure qui monte à environ 0,75. Toujours à l'instar des observations faites sur les données équilibrées, les combinaisons de descripteurs améliorent systématiquement les résultats, y compris lorsque l'on inclut les descripteurs les moins performants. Ainsi, la combinaison de deux jeux de descripteurs les moins performants, L et S, monte jusqu'à une F-mesure de 0,75. Toutes les autres combinaisons se rapprochent de 0,80 de F-mesure.

Pour comparer les algorithmes de classification traditionnels avec des méthodes plus récentes de l'état de l'art, nous avons mené la même tâche avec plusieurs approches neuronales.

— Nous avons utilisé un réseau à propagation avant avec la fonction d'activation `ReLU`, optimiseur `ADAM`, fonction d'erreur `BCEWithLogitsLoss` (sigmoïde + `BCELoss`). Nous avons effectué différentes expériences faisant varier le nombre de couches cachées et leur taille ainsi que le nombre d'*epochs*. Les résultats étaient plus élevés lors de l'entraînement sur les données équilibrées : jusqu'à 0.98 de précision, rappel et F-mesure. Cependant, à partir du déséquilibre 1 : 5, toutes les phrases étaient systématiquement classées dans la catégorie *non-aligné*. Cela montre les limites de l'architecture neuronale utilisée et l'impact du petit volume de données de référence disponibles.
— Nous avons fait des essais avec un système d'alignement existant qui utilise un réseau de neurones récurrent bidirectionnel (Grégoire & Langlais, 2018), mais les résultats obtenus avec nos données n'étaient pas exploitables. En effet, les résultats étaient très élevés (une f-mesure de 0,99) lors des expériences avec les données équilibrées, mais à partir du ratio 1 : 5 tout était systématiquement classé dans la classe majoritaire.
— Nous avons ajouté des descripteurs à base de plongements lexicaux supplémentaires, comme proposés par Kajiwara & Komachi (2016), mais cette expérience n'a pas montré d'évolution dans les résultats.

Nous pensons que le faible volume d'exemples à notre disposition (136 exemples d'équivalence et 130 exemples d'inclusion) représente un frein à l'utilisation de méthodes neuronales. Un travail plus avancé sur ce type de méthodes et l'accroissement du volume des données de référence font partie de nos perspectives.

Notons également que nous avons mené les mêmes expériences en rajoutant des descripteurs basés sur les connaissances. Nous avons ainsi exploité la ressource *UMLS* (Lindberg *et al.*, 1993). Il s'agit d'une ressource terminologique du domaine médical, dont une version française est accessible. Les termes médicaux sont associés à des identifiants alphanumériques, appelés *CUI* (*Concept Unique Identifier*). Un CUI regroupe une série de termes qui désignent le même concept. Ainsi, nous avons recherché les termes dans les phrases candidates à alignement pour détecter les CUI, ce qui nous a permis d'avoir un nouveau descripteur : le nombre de CUI en commun entre deux phrases. La méthode de détection des CUI consiste à rechercher les chaînes de caractères correspondantes dans le texte lemmatisé. Cette expérience n'a eu aucune incidence sur nos résultats, que ce soit en négatif ou en positif.

Pour nos descripteurs basés sur les plongements lexicaux, nous avons essayé d'utiliser des vecteurs pré-entraînés avec Fast Text [4] (Grave *et al.*, 2018), au lieu de ceux que nous avons entraînés sur CLEAR, et n'avons pas noté de différences significatives.

Enfin, nous avons cherché à traiter le déséquilibre avec un algorithme dédié à ce problème : SMOTE (Chawla *et al.*, 2002). Le principe de SMOTE consiste à produire des exemples synthétiques qui relèvent de la classe minoritaire, afin de réduire ou effacer le déséquilibre entre les classes. Cela n'a mené à aucune différence notable dans les résultats.

3.3.6. Analyse des erreurs

	équivalence	inclusion	intersection	faux positifs
Nb d'alignements	75	15	2	8

Tableau 3.5. – Analyse de 100 alignements par le modèle entraîné sur les couples équivalents avec un ratio de 125 : 1, appliqué sur un ensemble aléatoire de paires non vues pendant l'entraînement.

Nous avons sélectionné aléatoirement 100 alignements proposés par le classifieur, qui ne font pas parti de l'ensemble d'entraînement. Le tableau 3.5 montre la répartition de ces alignements selon le type de relation entre les deux phrases : équivalence, inclusion, intersection et faux positifs (absence de lien sémantique). Le modèle utilisé est entraîné sur un ratio de 1 : 125 avec les couples équivalents comme classe positive, car ce ratio montre le meilleur équilibre rappel/précision sur les données réelles. Nous pouvons observer que 75 % des alignements correspondent en effet à l'équivalence, ce qui est en adéquation avec la précision observée sur nos données de test, comme indiqué à la figure 3.2. Nous remarquons également que 15 % des alignements relèvent de l'inclusion. Nous trouvons enfin que 2 % des alignements correspondent aux intersections. Nous ne recherchons pas spécifiquement ce type de couples de phrases pour la simplification car nous le considérons plus difficile à exploiter pour la simplification. Cependant, de tels alignements peuvent être utiles pour l'identification de paraphrases, par exemple. Nous avons ainsi 90 à 92 % d'alignements exploitables et 8 à 10 % de bruit, selon que les intersections sont acceptées ou non.

Certains couples de phrases issues des corpus *Cochrane* et *Médicament* sont plus difficiles à aligner car ces corpus combinent les transformations lexicales spécifiques du domaine, des changements quant à l'emplacement de la négation et quant à sa nature, grammaticale ou lexicale (sous forme de préfixe), ainsi que des transformations syntaxiques liées au style injonctif/prescriptif adressé au patient prenant la médication, comme dans les exemples (3) et (4). De tels couples ne seront pas détectées comme équivalents par notre classifieur : elles font partie du silence. Pour aider l'alignement de tels couples de phrases, il serait nécessaire de capter la similarité

4. `https://fasttext.cc/docs/en/crawl-vectors.html`

lexicale et sémantique avec des ressources et connaissances complémentaires. Elles peuvent venir de ressources externes ou bien être acquises sur le corpus. Nous avons effectué un premier pas dans cette direction avec le repérage des CUI de l'UMLS, sans changement notable des résultats, (cf. section 3.3.5.6). Cependant un travail plus approfondi dans ce sens fait partie des perspectives ouvertes dans notre travail.

(3) *- Les médicaments inhibant le péristaltisme sont contre-indiqués dans cette situation.*
Dans ce cas, ne prenez pas de médicaments destinés à bloquer ou ralentir le transit intestinal.

(4) *- Aucune preuve n'indique que les agents gonflants sont efficaces dans le traitement du SCI*
- Nous avons observé que les agents gonflants n'étaient pas efficaces dans le traitement du SCI

3.3.7. Valorisation des données : tâche 2 de DEFT 2020

Les données que nous avons produites lors de ces travaux ont été valorisées dans le cadre d'une tâche de la campagne DEFT 2020 (Cardon *et al.*, 2020b). Tous les détails de la campagne sont disponibles dans les actes de l'atelier (Cardon *et al.*, 2020a).

La tâche proposée aux participants est la suivante : pour une phrase source donnée, trois phrases candidates à l'alignement sont proposées. Pour chaque phrase source, un seul bon alignement est possible. Le but consiste à identifier le bon candidat à l'alignement. Le cadre de cette tâche enlève ainsi deux contraintes majeures que nous avons rencontrées lors de notre travail d'alignement :

1. nous ne savions pas si chaque phrase trouverait un alignement,
2. nos données présentaient un déséquilibre beaucoup plus important entre phrases parallèles et non parallèles : 1 : 4 000 en conditions réelles au lieu de 1 : 2 lors de la tâche.

Le corpus d'entraînement comprend 572 ensembles de phrases sources et cibles tandis que le corpus d'évaluation en compte 530. Cette tâche est évaluée avec une mesure de précision classique. Le tableau 3.6 fournit des exemples de phrases sources et cibles sur différents sujets. Nous présentons un cas unique où la même phrase a été utilisée comme phrase source et comme l'une des phrases cibles (exemple 2) et un autre cas où des indices numériques tels que les dates ne correspondent pas forcément alors que les phrases sont à associer (exemple 3).

4 équipes ont participé à cette tâche. Le tableau 3.7 présente les scores de précision obtenus par les participants. Sur l'ensemble des soumissions, la moyenne est de 0,9822 et la médiane se situe à 0,9868. Nous voyons que les résultats sont très élevés. Nous pouvons en déduire que la suppression des deux caractéristiques auxquelles nous avons été confronté – le déséquilibre et l'absence de certitude quant à la présence d'un alignement possible pour chaque phrase – rend la tâche très aisée. Nous avons déjà montré la difficulté que le déséquilibre entraîne et avons proposé une méthode de filtrage des données pour la réduire. L'incertitude de trouver un alignement est un problème qui introduit des difficultés encore plus grandes. Si chaque phrase peut

Type	Phrases proposées
Source (1)	Arrivé en France en 1972, ce chat reste méconnu en dehors de son pays d'origine.
Cibles	Les principaux matériaux sont le grès et la latérite.
	Ce chat est apparu dans une portée d'American Shorthairs, en 1966, dans l'État de New-York.
	Bien qu'il soit apparu en France dès 1972, ce chat reste méconnu hors de son État d'origine.
Source (2)	En Suède, le taux légal est de 0,2 g par litre de sang.
Cibles	en suisse, le taux légal est de 0,5 g/l de sang ou 0,22 mg d'alcool par litre d'air expiré, depuis 2005.
	En Suède, le taux légal est de 0,2 g par litre de sang.
	En Belgique, le taux légal est de 0,5 g/l de sang ou 0,22 mg d'alcool par litre d'air expiré.
Source (3)	En 1534, il est appelé comme maître d'œuvre par le comte Giangiorgio Trissino pour diriger le chantier de la villa Cricoli.
Cibles	*En 1537, il est appelé par le comte Giangiorgio Trissino pour diriger le chantier de la villa Cricoli.*
	Trissino est un humaniste, poète, philosophe et diplomate au service de la curie romaine (le gouvernement pontifical) ; c'est aussi un passionné d'architecture .
	Le théâtre Olympique, achevé après 1580, est l'œuvre ultime de Palladio, terminée après sa mort par son fils Silla et son disciple Scamozzi.

Tableau 3.6. – Exemples de phrases sources et cibles pour la tâche 2 de DEFT 2020. La phrase cible la plus parallèle de la phrase source apparaît en italiques

être alignée avec une autre phrase, comme dans la tâche 2 de DEFT, il est possible de produire des candidats à l'alignement et la tâche devient une tâche de sélection du bon candidat.

3.3.8. Limites et perspectives

Nous revenons sur les limites de notre méthode d'extraction de phrases parallèles à partir de corpus comparables et sur les perspectives que notre travail ouvre.

La limite principale des expériences présentées est liée aux descripteurs exploités :

— Parmi les quatre types de descripteurs distingués dans les travaux existants (les descripteurs basés sur le lexique, l'exploitation de ressources externes, les descripteurs basés sur la syntaxe et les descripteurs basés sur le distributionnalisme)), nous exploitons les descripteurs principalement basés sur le lexique et le distributionnalisme. Cependant, cet aspect doit évoluer car les phrases équivalentes différenciées par leur degré de technicite peuvent présenter de fortes différences lexicales et syntaxiques, ce qu'a pu montrer notre observation des faux négatifs lors de l'analyse des erreurs d'alignement (section 3.3.6). Une

Soumission	EDF R&D		Reezocar			Sorbonne			Synapse		
	1	2	1	2	3	1	2	3	1	2	3
Précision	0,9830	0,9868	0,9868	0,9811	0,9849	0,9887	0,9887	0,9887	**0,9906**	0,9849	0,9396

Tableau 3.7. – Evaluation des prédictions en précision. Le meilleur résultat est en gras

meilleure intégration de plongements lexicaux, de ressources sémantiques et d'une architecture neuronale font partie de nos perspectives.

— Comme nous l'avons vu, dans les données de référence, la distance lexicale entre les phrases techniques et simplifiées est assez élevée. En conséquence, d'autres descripteurs doivent être utilisés pour mieux cerner les phrases alignables. Par exemple, nous envisageons d'utiliser des connaissances externes de manière plus approfondie, comme les terminologies médicales (Côté *et al.*, 1993; Lindberg *et al.*, 1993) et le lexique ReSyf (Billami *et al.*, 2018), ou des ressources que l'on peut acquérir à partir de corpus. Nous comptons également mettre à profit les modèles de langues neuronaux contextuels comme FlauBERT (Le *et al.*, 2020) et CamemBERT (Martin *et al.*, 2020).

— Une étude plus poussée des descripteurs pourrait également être intéressante : d'une part en comparant les performances selon les types d'alignement (équivalence et inclusion), d'autre part en étudiant l'impact des descripteurs individuellement et non pas par sous-ensembles. Nous introduisons également une étape préalable d'analyse syntaxique en constituants pour le filtrage. Cette étape a pour but de limiter les effets négatifs du déséquilibre, une caractéristique naturelle des données que nous traitons. Le travail à venir pourra consister à enrichir cette étape pour éliminer un maximum de phrases non alignables.

Une autre limite méthodologique est liée à la catégorisation binaire des paires de phrases selon qu'elles sont alignables ou non. Cette catégorisation est motivée par la tâche poursuivie, où nous avons besoin de couples de phrases parallèles pour induire des règles de transformation nécessaires et décrire ainsi la simplification. Cependant, comme dans les données STS de la campagne SemEval, nous pouvons aussi viser de caractériser les paires de phrases sur une échelle de similarité et disposer ainsi de données de référence plus fines. Cet objectif demande un effort d'annotation plus conséquent. Nous avons effectué un travail de ce type sur des données issues du corpus CLEAR et des articles de Wikipédia/Vikidia en langue générale (Cardon & Grabar, 2020b). Ces données ont été exploitées lors de la compétition DEFT en 2020 [5], que nous présentons dans la section suivante (section 3.4).

Nous avons utilisé notre modèle le modèle offrant l'équilibre optimal entre précision et rappel, celui entraîné avec un ratio de 1 : 125. Pour enrichir l'ensemble de phrases parallèles. En plus des corpus comparables liés au domaine médical, nous exploitons également un corpus semblable de la langue générale qui regroupe des articles comparables de Wikipédia et de Vikidia, ce qui représente 3 494 couples de documents. La ressource constituée grâce au travail dont nous parlons ici, un corpus de 10 942 couples de phrases alignées, sera mise à disposition pour la recherche. En

5. `https://deft.limsi.fr/2020/`

dehors de la simplification automatique, les phrases parallèles peuvent aussi être intéressantes pour d'autres applications de TAL, comme l'étude de la similarité textuelle, les systèmes de question-réponse, la recherche d'information ou l'implication textuelle.

3.4. Étude de la similarité sémantique

La similarité sémantique est une tâche du traitement automatique des langues. Elle consiste à évaluer le degré de proximité de sens entre deux énoncés donnés sur une échelle continue. Plusieurs compétitions dédiées à la similarité sémantique ont eu lieu dans le cadre de la campagne d'évaluation SemEval entre 2012 et 2017. Ainsi, des données pour plusieurs langues (Anglais, Espagnol et Arabe) ont été créées (Cer *et al.*, 2017) et rendues disponibles pour la recherche. De la même manière, des données pour le portugais ont été proposées lors de l'atelier ASSIN (Feitosa & Pinheiro, 2017). De telles données sont très utiles et recherchées, mais il n'en existe pas pour le français. Cela est un manque que nous proposons de combler.

Dans cette section, nous présentons un corpus de similarité sémantique pour le français. Les données utilisées sont issues du travail présenté plus haut (section 3.3). Ces données sont annotées manuellement grâce à l'assignation du degré de similarité entre deux phrases (section 3.4.1). Ensuite, nous présentons la ressource produite : un corpus de similarité sémantique pour le français (section 3.4.2). Nous décrivons également des expériences de reproduction automatique des annotations manuelles (section 3.4.3). Nous terminons par un bilan de ces différentes étapes 3.4.4.

3.4.1. Annotation manuelle de la similarité sémantique

Dans cette section, nous présentons d'abord les données fournies aux annotateurs, nous décrivons ensuite le processus d'annotation et analysons les critères d'annotation définis par les annotateurs.

3.4.1.1. Données

Un jeu de 1 010 paires de phrases a été fourni à cinq annotateurs. Les paires de phrases proviennent d'un corpus de langue générale qui contient des phrases extraites d'articles de Wikipédia et Vikidia, ainsi que de textes liés au domaine médical extraits du corpus CLEAR (Grabar & Cardon, 2018b). Les paires de phrases soumises à jugement ont été aléatoirement sélectionnées à partir d'alignements proposés par différentes versions intermédiaires du classifieur, présenté à la section précédente (section 3.3), et quelques appariements strictement aléatoires. Le fait de prendre des alignements de classifieurs moyennement performants nous paraît intéressant pour deux raisons :

— éviter d'avoir trop de paires de phrases non liées, en raison du déséquilibre créé par l'appariement de toutes les phrases possibles entre deux documents ;
— avoir une distribution satisfaisante de paires avec des scores variées : le fait que les paires de phrases non liées correspondent aux faux positifs implique que ces phrases représentent un minimum de similarité.

3.4.1.2. Processus d'annotation

Les cinq annotateurs ont un niveau scolaire allant de la licence au doctorat. Parmi ces cinq annotateurs, deux sont formés au traitement automatique des langues, un est médecin. Tous les annotateurs sauf un sont de langue française maternelle. L'auteur de ce travail n'a pas participé à l'annotation. Les instructions données aux annotateurs étaient simples et concises :

— attribuer un score de 0 quand les phrases n'ont rien à voir entre elles,
— assigner un score de 5 quand les phrases ont le même sens,
— définir soi-même une échelle et les critères pour les valeurs intermédiaires,
— donner une brève description des critères d'annotation établis.

Des travaux similaires comme ceux de SemEval (Cer *et al.*, 2017) ou sur la similarité sémantique en domaine biomédical (Soğancıoğlu *et al.*, 2017) donnent des instructions plus précises aux annotateurs. Cependant, nous avons choisi de ne pas donner d'instructions de ce type à nos annotateurs. Notre motivation était de nous baser sur la compétence linguistique des annotateurs, de ne pas biaiser les annotations de similarité sémantique entre les phrases et de pouvoir procéder à une comparaison des différents jugements sémantiques des annotateurs. De plus, les instructions présentées dans les travaux cités plus haut ne nous semblaient pas empêcher des différences dans leur interprétation par les annotateurs. Nous pouvons illustrer ce propos avec des descriptions comme "*The two sentences are roughly equivalent*" ou "*completely or mostly equivalent*", ou encore "*The two sentences share some details*".

Les annotateurs ont estimé que la tâche d'annotation manuelle des 1 010 paires de phrases leur a pris entre sept et quinze heures.

3.4.1.3. Échelles et critères d'annotation des annotateurs

Les annotateurs A1, A3, A4 et A5 ont assigné des scores entiers [0, 1, 2, 3, 4, 5] aux paires de phrases, alors que l'annotateur A2 a également utilisé des valeurs intermédiaires [0.5, 1.5, 2.5, 3.5, 4.5]. De plus, A3 a dit avoir pris les phrases strictement comme elles étaient : le contexte inconnu était considéré comme non-existant. Cela implique par exemple que les pronoms n'étaient jamais présumés comme se référant à un élément explicitement mentionné dans l'autre phrase, ce qui réduit les cas de similarité élevée entre les phrases. Nous illustrons ceci avec l'exemple (5). Les principes d'annotation de A3 l'ont mené à attribuer la similarité 0 à cette paire de phrases : il considérait que rien ne dit que "cette substance" et "ce latex" désignent la même substance. À titre d'information, les autres annotateurs ont attribué les scores 3 (A4), 4 (A2 et A5) et 5 (A1) à cette paire de phrases. La moyenne est donc de 3,2 : nous voyons ici que la moyenne permet d'atténuer les écarts entre les annotateurs.

(5) *- C'est de ce latex, une fois séché, que l'on extrait la morphine qui sert de base à l'héroïne.*
- C'est de cette substance qu'on extrait la morphine, dont est dérivée l'héroïne.

Les échelles et les critères de jugement utilisés par les annotateurs sont indiqués dans le tableau 3.8. Nous pouvons observer les différences et les similarités entre les différents critères d'annotation fournis par les annotateurs. En premier lieu, les

	A1	A2	A3	A4	A5
0.5		Quelques segments identiques			
1	Même sujet, peu de relation	Une phrase résume l'autre	Peu d'information partagée	Travail d'inférence possible	Presque aucun rapport de sens
1.5		Information principale manquante d'un côté et information secondaire absente			
2	Même sujet, information différente	Information principale incomplète d'un côté	Même fonction, peu d'information partagée	Niveau intermédiaire	Même sujet, information différente
2.5		Même sens, expression radicalement différente			
3	Même sujet, information faiblement partagée	Même sens, expression différente	Information supplémentaire d'un côté	Concept principal d'une phrase absent de l'autre	Information supplémentaire d'un côté
3.5		Même sens, des paraphrases			
4	Presque le même contenu, information supplémentaire d'un côté	Même sens, légères reformulations	Même fonction et information presque identique	Information supplémentaire d'un côté	Une légère différence dans l'information donnée
4.5		Même sens, légère différence syntaxique			

Tableau 3.8. – Critères d'annotation définis par les annotateurs

échelles élaborées par A2 et A3 tendent à donner des scores plus bas que les trois autres. Cependant ces deux échelles sont très différentes l'une de l'autre :

- A2 est le seul annotateur qui s'est intéressé à la formulation : pour qu'une paire de phrase atteigne le score le plus haut sur cette échelle, les deux phrases doivent être identiques.
- L'échelle élaborée par A3 est plus semblable aux autres, que celle d'A2, mais elle est plus contraignante en raison de son exclusion du contexte (dont les anaphores) dans le jugement de similarité.

Voilà quelques principes d'affectation des scores par les annotateurs, tout en sachant que l'annotateur A2 a utilisé une échelle différente :

- le score 5 implique que l'information est la même dans les deux phrases ;
- le score 4 implique que l'information est presque la même dans les deux phrases, par exemple avec des substitutions lexicales ;
- le score 3 implique que l'information est présente dans les deux phrases mais qu'au moins une phrase exprime de l'information absente de l'autre (les cas d'intersection sémantique, par exemple) ;
- le score 2 implique que les phrases ont quelque chose qui les différencie tout en parlant du même sujet. Nous illustrons ceci avec l'exemple en (6). Cet exemple reçoit le score de similarité 2 par A2, alors que les autres annotateurs lui attribuent des scores plus élevés : 3 pour A3 et A4, 4 pour A1 et 5 pour A5. Ici, selon les principes d'annotation de A2, la différence de dénomination des vitamines dont l'absorption est réduite semble être une information suffisamment différente pour contrebalancer la similarité. Ainsi nous pouvons observer deux raisonnements complémentaires : d'un côté il est possible de considérer la similarité et augmenter le score grâce aux informations similaires, de l'autre côté il est possible de considérer la différence entre les phrases et diminuer le score grâce aux informations dissimilaires ;
- le score 1 est plus difficile à décrire de manière synthétique pour l'ensemble annotateurs. A1 et A4 mentionnent quelque chose en commun (le domaine ou la possibilité de faire une inférence) mais disent aussi que rien de plus n'établit le lien entre les deux phrases. Quant aux A3 et A5, ils mettent l'accent sur l'absence de relation entre les phrases.

(6) *- L'utilisation prolongée de l'huile de paraffine est susceptible de réduire l'absorption des vitamines liposolubles (A, D, E, K).*
- L'utilisation prolongée de l'huile de paraffine est susceptible de diminuer l'absorption de certaines vitamines.

Pour résumer, nous pouvons voir que les annotateurs se sont attachés à plusieurs critères pour porter le jugement sur la similarité sémantique des phrases :

- intersection du sens (information manquante, incomplète ou supplémentaire d'un côté ou de l'autre),
- utilisation de paraphrases et d'expressions différentes,
- possibilité de faire de l'inférence,
- importance des différences entre les informations données dans les phrases.

De manière générale, nous observons que la complétude de l'information commune entre les phrases technique et simple est le critère le plus fréquemment utilisé par

tous les annotateurs.

3.4.1.4. Scores agrégés

À partir des scores de tous les annotateurs, nous calculons deux valeurs agrégées pour les paires de phrases :

— le score moyen arrondi pour chaque paire (*Moy* ci-après),

— le score assigné le plus souvent à chaque paire (*Vote* ci-après).

Comme nous l'avons vu, cela permet d'atténuer les différences entre les annotateurs et d'obtenir des scores plus consensuels.

3.4.2. Analyse des annotations

Dans cette section, nous analysons les annotations plus en détail : leur répartition par score et la corrélation des scores des cinq annotateurs.

3.4.2.1. Répartition par score

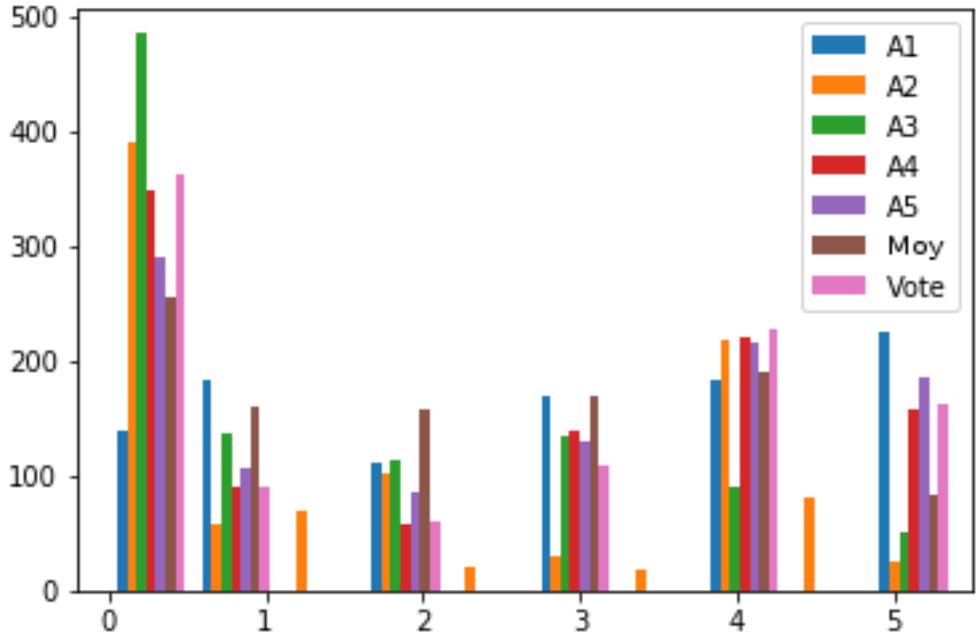


FIGURE 3.3. – Répartition par score et par annotateur

La figure 3.3 montre la répartition par score et par annotateur. L'axe horizontal montre les différents scores et l'axe vertical montre le nombre de paires. Les barres isolées en couleur orange sont dues à l'utilisation des valeurs intermédiaires (par pas de 0,5) par A2. Nous incluons aussi les scores Moy et Vote.

Nous pouvons observer que 0 est le score attribué le plus souvent par tous les annotateurs sauf un (A1). L'annotateur A3 a assigné le score 0 à quasiment la moitié des paires, ce qui est en accord avec les critères donnés par cet annotateur :

il ne fait aucune présomption sur le contexte pour la résolution de coréférences et a ainsi l'approche la plus contraignante.

Nous pouvons également voir que chaque annotateur sauf un (A1) a assigné le score 4 plus souvent que le score 5. Cela peut s'expliquer par la nature des paires de phrases. Comme indiqué dans la section 3.4.1.1, le corpus d'origine est prévu pour la simplification et les paires de phrases proviennent de couples de documents dont l'un est plus technique que l'autre. Ainsi, on peut s'attendre à ce qu'il y ait une plus grande quantité de phrases presqu'identiques que de phrases complètement identiques. En effet, les textes ne sont pas rédigés par les mêmes personnes ni pour les mêmes destinataires et donc elles ne fournissent par les mêmes informations de la même manière.

En observant les scores Moy et Vote, nous voyons que les scores 3 et 4 sont les plus consensuels. Le score 2 semble être le moins consensuel, avec une moyenne bien au-dessus de ce que les annotateurs ont attribué individuellement et un vote très bas.

Pour illustrer cette ressource de paires de phrases associées aux scores de similarité, nous présentons dans le tableau 3.9 quelques exemples de paires de phrases avec les moyennes des scores des cinq annotateurs.

3.4.2.2. Coefficients de corrélation

Nous avons calculé l'α de Krippendorff (Krippendorff, 1970) pour évaluer le coefficient de corrélation global des annotations. Pour les cinq annotateurs, l'α est de 0,69. Cette valeur est supérieure au seuil de fiabilité généralement utilisé ($\alpha = 0{,}67$). Cependant, ce score reste assez bas. Quand nous ajoutons Moy et Vote au calcul, l'α monte à 0,77. Afin d'explorer ces résultats plus en profondeur, nous avons calculé la corrélation entre les annotateurs pris deux par deux.

Le tableau 3.10 montre le coefficient de corrélation de Pearson pour chaque paire d'annotateurs. Les observations que nous pouvons faire sont cohérentes avec la figure 3.3 et les critères décrits en section 3.4.1.3 :

— Le coefficient de corrélation le plus bas (0.64) est observé entre A2 et A3 : A2 est l'annotateur qui a utilisé des pas de 0,5 dans son échelle, alors que l'exclusion de la résolution des coréférences par A3 l'a mené à attribuer 0 à presque la moitié des paires. Ainsi, ces deux annotateurs ont appliqué des échelles et des critères qui les différencient beaucoup. Les coefficients entre ces deux annotateurs et les trois autres vont de 0,70 à 0,77 ;

— Les coefficients de corrélation entre les trois autres annotateurs (A1, A4 et A5) sont les plus hauts : 0,84 pour A1 et A4, 0,81 pour A1 et A5 et 0,80 pour A4 et A5. Les échelles et critères de ces annotateurs sont en effet plus comparables.

Globalement, les coefficients de corrélation montrent une fiabilité satisfaisante pour ce jeu de données, avec des variations en fonction de différentes échelles élaborées par les annotateurs. Nous avons vu que les deux échelles qui se distinguent le plus, A2 et A3, ont le coefficient de corrélation le plus faible entre elles. Elles montrent cependant de bons coefficients de corrélation avec les trois autres annotateurs : entre 0.70 et 0.77 (rappelons qu'un accord est qualifié de substantiel à partir de 0,67). Les trois autres annotateurs montrent des coefficients de corrélation très élevés entre eux : de 0,80 à 0,84.

Phrases source et cible	*Score*
Il commence par s'intéresser à la résistance à la faim, la soif et à la fatigue en 1951.	0,4
Pour prouver qu'on pouvait vivre sans eau ni nourriture, il traversa en solitaire l'Atlantique sans autres ressources que les poissons, le plancton, l'eau de pluie et de petites quantités d'eau de mer durant 65 jours.	
Comparativement à un traitement d'une durée standard, les études portant sur le traitement de courte durée ont présenté des périodes plus courtes de fièvre (différence moyenne (DM) -0,30 jour, intervalle de confiance (IC) à 95 % -0,45 à -0,14) et de mal de gorge (dm -0,50 jour, ic à 95 % -0,78 à -0,22).	1,4
Le traitement de courte durée a entraîné une meilleure observance mais davantage d'effets secondaires.	
Deux essais (106 participants) comparaient l'héparine de bas poids moléculaire à un placebo ou à l'absence de traitement.	2
Deux essais (259 participants) comparaient l'héparine à l'absence de traitement.	
Les études examinées ont été réalisées à trop petite échelle pour déterminer si le remplissage était bénéfique aux femmes recevant une analgésie régionale au cours du travail avec les anesthésiques ou les opioïdes locaux à plus faible dose.	3,2
Les preuves sont insuffisantes pour déterminer si le remplissage est bénéfique aux femmes recevant une analgésie régionale au cours du travail avec les agents à plus faible dose ou aux femmes ayant des complications de grossesse.	
Les sondes gastriques sont couramment utilisées pour administrer des médicaments ou une alimentation entérale aux personnes ne pouvant plus avaler.	4
Les sondes gastrique sont couramment utilisées pour administrer des médicaments et de la nourriture directement dans le tractus gastro-intestinal (un tube permettant de digérer les aliments) pour les personnes ne pouvant pas avaler.	
Les études incluses ne rapportaient pas les coûts des soins de santé comme critère de jugement.	4,8
Aucune des études incluses ne rapportaient les coûts des soins de santé comme critère de jugement.	

Tableau 3.9. – Exemples de paires de phrases avec la moyenne des scores attribués par les annotateurs.

3.4.3. Calcul automatique de la similarité des paires de phrases

Nous avons mené une expérience pour vérifier à quel point les annotations peuvent être reproduites automatiquement. Dans cette section, nous décrivons notre méthode

	A1	A2	A3	A4	A5
A1	1	0,77	0,72	0,84	0,81
A2	0,77	1	0,64	0,75	0,74
A3	0,72	0,64	1	0,75	0,70
A4	0,84	0,75	0,75	1	0,80
A5	0,81	0,74	0,70	0,80	1

Tableau 3.10. – Coefficients de corrélation de Pearson entre les annotateurs

pour le calcul automatique de la similarité des paires de phrases et les résultats obtenus. La tâche 1 de la campagne d'évaluation DEFT 2020 était effectuée sur les mêmes données. Nous pouvons donc comparer nos résultats avec ceux des participants à cette tâche.

Nous exploitons la méthode précédemment décrite pour la détection de phrases parallèles dans des corpus comparables (section 3.3). Cependant, pour projeter les scores sur une échelle continue, nous utilisons l'algorithme `Random Forest Regressor` au lieu du classifieur. Nous rappelons ici les descripteurs utilisés :

1. *Nombre de mots pleins en commun* : Ce descripteur permet de calculer le recouvrement lexical basique entre les phrases technique et simplifiée (Barzilay & Elhadad, 2003) ;
2. *Pourcentage de mots d'une phrase dans l'autre, dans les deux directions* : Ce descripteur représente l'inclusion lexicale et sémantique entre les deux phrases ;
3. *Ratio de longueur en mots entre les deux phrases* : Ce descripteur présume que la simplification implique une association stable avec la longueur de la phrase ;
4. *Différence de la longueur moyenne des mots entre les phrases* : Ce descripteur est similaire au précédent mais prend en compte la longueur des mots ;
5. *Nombres totaux de bigrammes et trigrammes communs* : Ce descripteur est calculé sur les n-grammes de caractères ;
6. *Mesures de similarité basées sur le lexique (cosinus, Dice et Jaccard)* : Ce descripteur fournit une indication plus sophistiquée sur le recouvrement lexical entre les phrases. Le poids pour chaque mot est de 1 ;
7. *Distance d'édition minimale en caractères* (Levenshtein, 1966) : Ce descripteur est une acception classique de la distance d'édition. Il prend en compte les opérations d'édition basiques (insertion, suppression et substitution) au niveau des caractères. Le coût de chaque opération est de 1 ;
8. *Distance d'édition minimale en mots* (Levenshtein, 1966) : Ce descripteur est calculé avec le mot pour unité. Il prend en compte les mêmes opérations d'édition avec le même coût à 1.
9. *WAVG* : Ce descripteur utilise le plongement lexical. La moyenne des vecteurs de chaque mot est calculé pour les deux phrases, et le score de similarité résulte de la comparaison de ces deux moyennes (Štajner *et al.*, 2018) ;
10. *CWASA* : Ce descripteur est l'analyse de similarité continue basée sur l'alignement de mots, comme décrite par Franco-Salvador *et al.* (2016).

Nous rappelons aussi que, pour les deux derniers descripteurs, nous avons entraîné les plongements lexicaux sur le corpus CLEAR avec l'algorithme Word2Vec (Mikolov *et al.*, 2013b), alors que les scores ont été calculés avec l'outil CATS (Štajner *et al.*, 2018).

Nous avons effectué l'expérience sur les scores de chaque annotateur, de même que sur les scores Moy et Vote. Les données sont aléatoirement segmentées en deux ensembles : 60 % pour l'entraînement et 40 % pour le test. Le tableau 3.11 présente le nombre et le pourcentage d'annotations pour chaque degré de similarité dans les corpus d'entraînement et d'évaluation.

Corpus	Degrés de similarité											
	0		1		2		3		4		5	
Entraînement (600 paires)	216	36,0%	56	9,3%	29	4,8%	66	11,0%	136	22,7%	97	16,2%
Evaluation (410 paires)	147	35,9%	37	9,0%	28	6,8%	44	10,7%	90	22,0%	64	15,6%

Tableau 3.11. – Nombre et pourcentage d'annotations par degré de similarité dans les corpus de la tâche 1

A1	A2	A3	A4	A5	Moy	Vote
0,80	0,74	0,72	0,79	0,76	0,85	0,79

Tableau 3.12. – Coefficient de corrélation de Pearson pour les expériences de régression

Le tableau 3.12 montre les résultats obtenus avec notre modèle. Selon les annotateurs, les coefficients de corrélation vont de 0.72 (A3) à 0.80 (A1). Cela montre que les différentes échelles peuvent être reproduites automatiquement assez aisément. Le score sur la moyenne est à 0,85 et sur le vote à 0,79. L'observation la plus intéressante est que les meilleurs résultats (coefficient de corrélation Pearson à 0,85) sont obtenus sur la moyenne des scores. Cela indique que la moyenne des scores et la perception collective de la similarité sémantique restent cohérentes, malgré les différences observées lors de l'analyse des grilles d'annotation élaborées individuellement par les annotateurs.

Le tableau 3.13 reprend les corrélations entre les annotateurs (tableau 3.10) et y ajoute le score du modèle pour reproduire les annotations de chacun d'entre eux. Nous voyons que le modèle parvient à un accord comparable à celui qui est observé entre les annotateurs humains. Le score du modèle automatique se situe en quatrième place pour A2, en troisième place pour A1 (avec 0,82), A4 (avec 0,79) et A5 (avec 0,78) et en première place pour A3 (avec 0,80).

Nous avons fourni les données pour la tâche 1 de la campagne d'évaluation DEFT 2020. Comme lors de l'annotation pour la création des données de référence, aucune définition des degrés de similarité de 0 à 5 n'a été fournie aux participants.

5 équipes ont participé à cette tâche. Nous proposons ici un aperçu des méthodes utilisées par les participants :

	A1	A2	A3	A4	A5
A1	1	0,77	0,72	0,84	0,81
A2	0,77	1	0,64	0,75	0,74
A3	0,72	0,64	1	0,75	0,70
A4	0,84	0,75	0,75	1	0,80
A5	0,81	0,74	0,70	0,80	1
Modèle	0,82	0,73	0,80	0,79	0,78

Tableau 3.13. – Coefficients de corrélation de Pearson calculés entre les annotateurs et entre les annotateurs et le modèle de régression.

— L'équipe UASZ (Drame *et al.*, 2020) a exploité les mesures de similarité Dice, Jaccard, Ochiai (OCHIAI, 1957) et Q-Gram (Ukkonen, 1992) ainsi que des représentations vectorielles TF×IDF et à base de plongements lexicaux ;
— Synapse (Belkacem *et al.*, 2020) a proposé une méthode se basant sur des modèles de plongements lexicaux dérivés de BERT, en particulier Sentence M-BERT et MUSE ;
— L'équipe Sorbonne (Buscaldi *et al.*, 2020) s'est concentrée sur les mesures de similarité Dice, Jaccard, Cosinus et Bray-Curtis (Bray & Curtis, 1957) pour étudier cette forme de *baseline* ;
— L'équipe Reezocar (Tapi Nzali, 2020) a utilisé l'algorithme XGBoost (Chen & Guestrin, 2016) avec des vecteurs de poids TF×IDF (Sparck Jones, 1988) ;
— L'équipe EDF (Cao *et al.*, 2020) a mis en œuvre deux types de méthodes : une méthode à base de graphes sémantiques et deux classifieurs basés sur la régression logistique. L'un des classifieurs utilise des descripteurs similaires à ceux que nous avons utilisés : Dice, nombre de mots communs entre les deux phrases et la distance de Levenshtein. D'autres descripteurs sont utilisés : le nombre de mots de chacune des deux phrases, la différence entre le nombre de mots des deux phrases et la valeur absolue de cet écart. L'autre classifieur reproduit explicitement les descripteurs que nous avons proposés dans un travail publié avant l'intégration des descripteurs à base de plongements lexicaux (Cardon & Grabar, 2018).

Le tableau 3.14 montre les résultats obtenus avec la corrélation de Spearman sur les prédictions des participants, ainsi que sur nos résultats. La première ligne montre les résultats que nous obtenons avec cette métrique. Nous voyons ensuite que les résultats des participants à la tâche vont de 0,71 (EDF 2, Reezocar 1 et 3) à 0,78 (UASZ 3). Bien que les méthodes soient variées, nous voyons que les mesures de similarité classiques ainsi que la similarité cosinus pour comparer les représentations de phrases à base de plongements lexicaux sont performantes pour cette tâche, en accord avec ce que nous avons également montré dans nos travaux. Notons que la reproduction par l'équipe EDF (soumission 3 dans le tableau 3.14) de notre méthode sans les plongements lexicaux obtient une corrélation de Spearman de 0,72 alors que nous obtenons 0,75 avec les plongements lexicaux en plus.

Soumission	*Corrélation de Spearman*
Nos travaux	0,75
EDF R&D, 1	0,73
EDF R&D, 2	0,71
EDF R&D, 3	0,72
Reezocar, 1	0,71
Reezocar, 2	0,74
Reezocar, 3	0,71
Sorbonne, 1	0,75
Sorbonne, 2	0,73
Sorbonne, 3	0,75
Synapse, 1	0,75
Synapse, 2	0,74
Synapse, 3	0,77
UASZ, 1	0,75
UASZ, 2	0,77
UASZ, 3	**0,78**

Tableau 3.14. – Evaluation des prédictions avec la corrélation de Spearman. Le meilleur résultat est en gras

3.4.4. Bilan

Nous avons proposé un travail inspiré d'une tâche récurrente dans le traitement automatique des langues : celle de la similarité sémantique. Aucune ressource dédiée à cette tâche n'existait pour le français, nous venons combler ce manque. Notre ressource consiste en 1 010 paire des phrases annotées manuellement par cinq personnes sur une échelle de 0 à 5. Nous avons laissé les annotateurs juger des principes à appliquer pour l'attribution des scores sur cette échelle. Nous avons analysé les échelles d'annotation qu'ils ont utilisées et nous ont transmises, ainsi que les scores qu'ils ont attribués. Cette analyse nous a permis d'en produire une synthèse, qui comprend une description des principes adoptés par les annotateurs ainsi que des scores agrégés. Les expériences ultérieures que nous avons menées, ainsi que celles des participants de la tâche 1 de DEFT 2020, ont permis de montrer que les données étaient exploitables pour la reproduction automatique de ces annotations.

3.5. Conclusion

Nous avons proposé une série d'expériences d'alignement de phrases parallèles à partir de corpus monolingues comparables en français. La dimension comparable est due à la technicité des documents et contraste les versions techniques et simplifiées des documents et des phrases. Nous exploitons un corpus comparable lié au domaine biomédical et contenant des documents de trois genres (encyclopédique, scientifique et notices de médicaments). Les données de référence sont construites manuellement. La recherche de phrases parallèles est abordée comme une problématique de catégorisation : nous devons décider si une paire de phrases peut être alignée ou non.

Plusieurs classifieurs et descripteurs sont exploités. Nos résultats atteignent une F-mesure de 0,98 sur les données équilibrées avec un bon équilibre entre la précision et le rappel. Les meilleurs résultats sont obtenus avec le classifieur `Random Forest`.

Les pistes d'amélioration que nous envisageons sont les suivantes :

— Un travail plus approfondi sur le filtrage syntaxique. En effet, bien que cette étape soit utile pour traiter le problème du déséquilibre, elle limite la représentativité des opérations syntaxiques de simplification présentes dans le corpus. Notre hypothèse est qu'il est possible d'extraire des règles plus complexes à partir des exemples laissés de côté par le filtre actuel. Ces règles serviraient à enrichir l'étape de filtrage. Un filtrage enrichi permettrait d'une part d'obtenir un corpus plus fourni. D'autre part, il permettrait de mieux connaître les caractéristiques linguistiques des opérations de transformations syntaxiques non identifiées par le filtre dans le corpus. Cette traçabilité est souhaitable pour une meilleure explicabilité de la suite du traitement, notamment lorsque les corpus servent par la suite à entraîner des systèmes d'apprentissage profond.

— Une intégration plus poussée de ressources et connaissances complémentaires. Telles quelles, les ressources terminologiques du domaine médical ne sont pas adaptées à notre tâche. En effet, l'intégration de la la détection des CUI de l'UMLS dans nos descripteurs (section 3.3.5.6 page 53) n'a pas amélioré nos résultats. Cependant, ces ressources sont riches en informations et un travail d'adaptation de ces ressources à la tâche d'alignement peut être envisagé.

— Une sélection plus étendue de descripteurs. Notamment, de nouveaux modèles de langue pour les plongements lexicaux ont récemment vu le jour, basés sur l'algorithme BERT (Devlin *et al.*, 2019). Des modèles pour le français ont été entraînés, FlauBERT (Le *et al.*, 2020) et CamemBERT (Martin *et al.*, 2020). Utiliser de tels modèles pour obtenir des représentations de phrases et les comparer pourrait être un ajout intéressant.

— L'utilisation de méthodes neuronales. Notre méthode nous a permis de faciliter le travail d'extraction de phrases alignées dans le corpus comparable. Le volume de données augmentant, nous envisageons à terme de pouvoir utiliser des réseaux de neurones pour l'alignement de phrases.

La ressource créée à l'aide de cette méthode contient actuellement 10 942 couples de phrases équivalentes relevant du domaine biomédical.

Deux autres expériences s'intéressent aux types de relations au sein des couples de phrases (les couples de phrases avec la relation d'équivalence sont plus faciles à aligner que les phrases en relation d'inclusion) et à l'équilibre entre les paires alignables et non alignables.

Nous avons aussi présenté une expérience supplémentaire effectuée pour combler l'absence de corpus de similarité sémantique en français. Au cours de cette expérience, un corpus de 1 010 phrases extraites du corpus biomédical et d'articles divers de Wikipédia et Vikidia a été annoté manuellement par cinq annotateurs. Nous avons demandé à ces annotateurs de décider quelle est la similarité sémantique entre les phrases, en suivant une échelle de 0 à 5. Le corpus est fourni avec des scores manuellement attribués par cinq annotateurs. Avec leurs scores, les annotateurs ont également fourni les schémas d'annotation qu'ils ont utilisés. Nous avons analysé les données produites et montré qu'il y a des écarts entre les différentes échelles

d'annotation, ce qui indique une nécessité d'expliciter les critères de (dis)similarité associée à chaque point de l'échelle dans de futures expériences de ce type. Malgré ces écarts, les accords inter-annotateurs sont globalement élevés. Les écarts les plus importants peuvent s'expliquer par les choix faits lors de l'annotation de la similarité sémantique. Nous avons ensuite utilisé ces données pour prédire automatiquement les scores de similarité pour les paires de phrases. Ces expériences montrent que les scores peuvent être assez bien reproduits avec des approches automatiques : en comparaison avec différents coefficients de corrélation entre les différents annotateurs humains, le résultat du calcul automatique du degré de similarité présente dans tous les cas un coefficient plus élevé de corrélation que l'annotateur humain le plus déviant. Ainsi le calcul automatique de la similarité sémantique et les données créées manuellement se confirment mutuellement peuvent tous deux être utilisés pour des travaux autour de la similarité sémantique.

La tâche d'attribution d'un score de similarité sémantique sur nos données, avec une échelle de 0 à 5, a été proposée à la compétition DEFT 2020. Nous avons ainsi pu comparer notre approche avec celles des participants. Notre méthode a obtenu une corrélation de Spearman à 0,75. Ces résultats sont comparables à ce que d'autres équipes ont proposé : les scores de corrélation des participants vont de 0,71 à 0,78.

Ces expériences nous ont permis de construire deux ressources qui manquaient jusqu'alors pour le français :

— un corpus de similarité sémantique de 1 010 paires de phrases annotées sur une échelle de 0 (sémantique indépendante) à 5 (sémantique identique) ;
— un corpus de 10 942 couples de phrases parallèles et alignées utilisable pour la simplification automatique de textes.

4. Expériences en simplification automatique

4.1. Introduction

Dans les chapitres prédédents, nous avons décrit nos travaux sur la constitution des données qui peuvent servir de matériau pour la simplification automatique. Le chapitre actuel décrit nos expériences en simplification automatique de textes médicaux en français. Nous débutons ce chapitre en présentant les travaux existants dans le domaine de la simplification automatique de textes, y compris les travaux qui concernent l'évaluation de textes simplifiés (section 4.2). Ensuite, nous décrivons une série d'expériences de simplification automatique que nous avons menée à l'aide de ressources spécifiques et d'une méthode neuronale (section 4.3). Enfin, nous concluons ce chapitre (section 4.4).

Ce chapitre est partiellement basé sur deux publications (Cardon, 2018; Cardon & Grabar, 2020a).

4.2. État de l'art

Cette section présente un état de l'art de la simplification automatique de textes ainsi que des méthodes pour l'évaluation de la simplification. La simplification automatique de textes est une tâche relativement récente dans le domaine du traitement automatique des langues. Elle a connu un grand succès auprès des chercheurs. Ainsi, plusieurs travaux établissant un panorama du domaine de la simplification automatique ont été publiés ces dernières années (Siddharthan, 2014; Shardlow, 2014; Saggion, 2017; Alva-Manchego *et al.*, 2020b).

Nous terminons par une présentation des méthodes d'évaluation des systèmes de simplification automatique.

La simplification automatique peut être effectuée dans deux cadres très différents :

— *Simplifier des textes à destination des machines.* Historiquement, c'est la première perspective explorée en simplification automatique, avec pour objectif de faciliter l'analyse syntaxique pour des tâches ultérieures (Chandrasekar *et al.*, 1996). Plus récemment, la simplification automatique est explorée pour des tâches de TAL comme l'étiquetage sémantique (Evans & Orasan, 2019) ou la traduction automatique de langues peu dotées (Štajner & Popović, 2019), par exemple ;

— *Simplifier des textes à destination des humains.* La simplification peut également aider les utilisateurs humains à mieux comprendre les textes. Ainsi, plusieurs types d'utilisateurs ont été ciblés dans les travaux existants, comme

par exemple : les lecteurs aphasiques (Devlin & Unthank, 2006), les enfants (De Belder & Moens, 2010) ou le grand public non spécialiste d'un domaine de spécialité (Shardlow & Nawaz, 2019).

Notre travail s'inscrit dans ce deuxième cadre, simplifier les documents à destination des utilisateurs humains. De plus, nous visons à simplifier les textes techniques du domaine médical.

4.2.1. Simplification automatique de textes

Traditionnellement, on positionne la simplification automatique à deux niveaux linguistiques essentiellement : la simplification syntaxique (section 4.2.1.1) et la simplification lexicale (section 4.2.1.2). Ces deux niveaux sont d'habitude abordés avec des méthodes et outils spécifiques. Un autre niveau linguistique est le niveau discursif, où il s'agit de travailler au niveau du texte et non pas de la phrase, comme c'est fait avec la simplification lexicale et syntaxique. Dans notre travail, nous nous intéressons surtout à la simplification lexicale et syntaxique car, comme nous l'avons vu au chapitre 2, section 2.5.2.2 (page 32), ces deux types de transformations sont extrêmement fréquents dans les documents médicaux. Bien que des travaux explorent la dimension discursive au sein du TAL (Braud & Denis, 2016; van Noord *et al.*, 2018) les recherches portant sur cette dimension en simplification n'en sont qu'à leurs débuts (Alva-Manchego *et al.*, 2019b). Avec les développements récents de méthodes d'apprentissage dans le traitement automatique des langues, y compris en simplification automatique, les méthodes neuronales sont prometteuses. Nous présentons donc les travaux exploitant de telles méthodes pour la simplification syntaxique et lexicale dans la section 4.2.1.3.

4.2.1.1. Simplification syntaxique

Les premiers travaux en simplification automatique ont été effectués au niveau syntaxique des phrases (Chandrasekar *et al.*, 1996). L'objectif principal consiste alors à simplifier des phrases complexes en les découpant en plusieurs phrases simples. Ainsi, la méthode proposée dans ce travail exploite un analyseur syntaxique afin de reconnaître les propositions relatives et les appositions, et de les extraire pour en faire des phrases indépendantes. Dans l'exemple en (1) tiré de (Chandrasekar *et al.*, 1996), une simplification effectuée manuellement illustre l'objectif poursuivi et le résultat obtenu. Nous voyons que la proposition *who masterminded the Kanishka crash in 1984* est extraite de la phrase d'origine pour devenir une phrase indépendante, avec une reprise du sujet *Talwinder Singh* dans la nouvelle phrase.

(1) *Talwinder Singh, who masterminded the Kanishka crash in 1984, was killed in a fierce two-hour encounter.*
Talwinder Singh was killed in a fierce two-hour encounter. Talwinder Singh masterminded the Kanishka crash in 1984.

Typiquement, un processus en trois étapes permet d'effectuer la simplification syntaxique (Shardlow, 2014) :

1. effectuer l'analyse syntaxique de la phrase ;

2. appliquer des règles, le plus souvent créées manuellement (Brouwers *et al.*, 2014; Evans & Orăsan, 2019) et plus rarement apprises sur un corpus (Chandrasekar & Srinivas, 1997; Seretan, 2012) ;
3. produire la simplification syntaxique.

4.2.1.2. Simplification lexicale

La simplification lexicale consiste à détecter les termes (mots, expressions, abréviations, etc.) difficiles à comprendre dans un texte et à les remplacer de façon à ce qu'ils soient accessibles au lectorat visé. Il s'agit d'une spécialisation de la tâche de substitution lexicale, tâche qui vise à remplacer un mot par un autre qui convient mieux dans un contexte donné (McCarthy & Navigli, 2007; Fabre *et al.*, 2014; Zhou *et al.*, 2019). Les premiers travaux visant à accomplir la simplification lexicale (Carroll *et al.*, 1998; Devlin & Tait, 1998) se basent sur une ressource lexicale en anglais, comme le réseau sémantique WordNet (Miller, 1995). WordNet est une base de données lexicales qui renseigne sur les relations sémantiques entre les mots, et en particulier sur les relations de synonymie ou d'hyperonymie. Pour un mot à simplifier donné, la méthode consiste typiquement à rechercher ses synonymes dans WordNet. Le mot et ses synonymes sont ensuite associés à un indice de fréquence, tiré de la liste de fréquences de Kučera-Francis dans la *Oxford Psycholinguistic Database* (Quinlan, 1992). Le mot dont la fréquence est la plus haute est conservé pour effectuer la simplification.

Depuis cette première méthode, la simplification lexicale s'est considérablement développée. Le consensus actuel décrit cette tâche comme une succession de quatre étapes entre l'entrée et la sortie d'un système de simplification automatique. Chacune de ces étapes donne lieu à des travaux spécifiques :

1. *Identification de mots ou termes qui peuvent présenter des difficultés de compréhension.* Une première méthode consiste à mettre à profit des ressources lexicales de référence (Shardlow, 2013; Paetzold & Specia, 2016) et les fréquences associées aux mots. Une deuxième méthode consiste à détecter les difficultés de compréhension avec des techniques d'oculométrie, qui permettent d'observer les mouvements de l'œil lors de la lecture d'un texte (Grabar *et al.*, 2018). En effet, il a été observé que les lecteurs ont tendance à fixer le regard plus longtemps sur les passages les plus difficiles à comprendre. Les travaux plus récents utilisent des méthodes neuronales pour l'identification de termes complexes, ce qui nécessite une annotation manuelle préalable (Gooding & Kochmar, 2019; Yimam *et al.*, 2018; Finnimore *et al.*, 2019; Yimam *et al.*, 2017; Bingel & Bjerva, 2018; Pylieva *et al.*, 2019).
2. *Production d'une liste de candidats à la substitution* (Saggion *et al.*, 2013). Cette étape repose souvent sur la disponibilité d'un dictionnaire d'expressions synonymiques, comme WordNet pour l'anglais (Miller, 1995) ou ReSyf pour le français (Billami *et al.*, 2018), où les potentiels candidats à la simplification ont la même catégorie lexicale que le terme à substituer.
3. *Ordonnancement des candidats à substitution.* Lorsque plusieurs candidats à substitution sont disponibles, il est nécessaire de les ordonner par rapport à leur niveau de difficulté pour être en mesure de sélectionner les candidats les

plus faciles à comprendre (François *et al.*, 2016). Cette étape était au cœur d'une tâche proposée lors de la compétition *SemEval 2012* (Specia *et al.*, 2012), où les organisateurs donnaient aux participants des phrases à simplifier et une liste de candidats à la substitution. L'objectif de la tâche consistait à sélectionner le meilleur candidat. Les participants ont exploité plusieurs critères pour effectuer cette tâche : lexique d'un corpus oral et de Wikipedia, n-grammes de Google, WordNet (Sinha, 2012) ; longueur de mots, nombre des syllabes, information mutuelle, fréquences (Jauhar & Specia, 2012) ; fréquences dans Wikipedia, longueur de mots, n-grammes, complexité syntaxique des documents (Johannsen *et al.*, 2012) ; n-grammes, fréquences dans Wikipedia, n-grammes de Google (Ligozat *et al.*, 2012) ; WordNet, fréquences (Amoia & Romanelli, 2012). Comme pour l'identification de mots complexes, des méthodes neuronales sont maintenant de plus en plus utilisées pour cette étape (Paetzold & Specia, 2017).

Nous illustrons ces étapes avec un exemple donné par les organisateurs de SemEval 2012. La phrase à simplifier est *Hitler committed terrible atrocities during the second World War.* Les organisateurs évoquent un système qui identifierait *atrocities* comme un mot complexe. Ils donnent la liste des candidats qui seraient renvoyés par une ressource adéquate : *abomination*, *cruelty*, *enormity* et *violation*. Il est attendu des participants qu'ils sélectionnent le plus simple parmi ces candidats, ici *cruelty*, et qu'ils procèdent à la substitution pour obtenir la phrase *Hitler committed terrible cruelties during the second World War.*

4.2.1.3. Méthodes d'apprentissage

Les premières méthodes d'apprentissage pour la simplification se basent sur des méthodes statistiques.

— (Zhu *et al.*, 2010a) proposent un système qui apprend des règles de transformation syntaxique sur un corpus pré-traité par un analyseur syntaxique.
— (Woodsend & Lapata, 2011) se placent dans le cadre des grammaires quasi-synchrones (Smith & Eisner, 2006). Ils proposent un système qui génère plusieurs simplifications pour une phrase donnée et sélectionnent la meilleure à l'aide d'une méthode d'optimisation linéaire en nombres entiers.
— (Coster & Kauchak, 2011b) et (Wubben *et al.*, 2012) utilisent différentes méthodes de traduction statistique pour la simplification automatique.

Les travaux les plus récents utilisent des méthodes neuronales pour la simplification automatique (Zhang & Lapata, 2017; Nisioi *et al.*, 2017; Sulem *et al.*, 2018b; Shardlow & Nawaz, 2019; Abdul Rauf *et al.*, 2020; Cooper & Shardlow, 2020). Avec ces méthodes, la simplification lexicale et la simplification syntaxiques sont effectuées en une seule fois. Les règles de transformation sont apprises automatiquement par le modèle. Le modèle fonctionne comme une boîte noire et les règles ne sont accessibles ni au concepteur ni à l'utilisateur.

L'architecture typiquement utilisée est celle de l'encodeur-décodeur avec un mécanisme d'attention, initialement conçue pour la traduction automatique (Bahdanau *et al.*, 2014). Cette architecture fonctionne avec trois éléments principaux.

1. Un réseau de neurones, l'encodeur, qui produit un vecteur à partir de la phrase

à simplifier.

2. Le vecteur produit par l'encodeur, qui a pour rôle de représenter le sens de la phrase d'origine.
3. Un réseau de neurones, le décodeur, qui décode le vecteur produit par l'encodeur et en fait une nouvelle phrase.

L'encodeur et le décodeur sont entraînés en même temps. Pour une phrase technique donnée, lors de l'entraînement d'un système de simplification avec cette architecture, l'encodeur la convertit en vecteur et le décodeur produit une nouvelle phrase à partir de ce vecteur. Ensuite, l'écart entre la nouvelle phrase et la simplification de référence est calculé avec une fonction d'erreur. Les paramètres de l'encodeur et du décodeur sont ajustés en fonction de l'erreur calculée et le processus est répété jusqu'à la fin de l'entraînement. Nous décrivons plus en détails les deux premiers travaux à utiliser l'architecture encodeur-décodeur pour la simplification automatique de textes (Zhang & Lapata, 2017; Nisioi *et al.*, 2017). Les deux travaux ont été publiés à quelques mois d'intervalle.

Nous citons d'abord l'étude de Nisioi *et al.* (2017) car elle a servi de base à d'autres travaux (Shardlow & Nawaz, 2019; Abdul Rauf *et al.*, 2020; Cooper & Shardlow, 2020). Les auteurs utilisent l'outil OpenNMT (Klein *et al.*, 2017), un outil libre et gratuit conçu pour la traduction automatique. L'encodeur et le décodeur utilisent des LSTM *long short-term memory* (Hochreiter & Schmidhuber, 1997). Plus de détails sur l'implémentation qu'ils utilisent sont disponibles dans l'article. Les travaux en simplification neuronale ayant eu lieu par la suite sont le plus souvent des variations autour de cette architecture (Alva-Manchego *et al.*, 2017; Scarton & Specia, 2018; Vu *et al.*, 2018; Shardlow & Nawaz, 2019; Abdul Rauf *et al.*, 2020; Cooper & Shardlow, 2020).

Zhang & Lapata (2017) utilisent aussi des LSTM pour l'encodeur et le décodeur. Plus de détails sur l'implémentation sont disponibles dans l'article. Avec ce modèle encodeur-décodeur de base, les auteurs disent obtenir un taux très élevé de reproduction de la phrase d'entrée vers la phrase de sortie (73 % sur Newsela et 83 % sur les jeux de données basés sur Wikipedia). Pour remédier à cela, les auteurs proposent d'utiliser un cadre d'apprentissage par renforcement (Williams, 1992). Cela signifie que pendant l'entraînement, une fois que le décodeur a terminé la production de la phrase de sortie, l'algorithme de renforcement intervient dans la mise à jour des paramètres de l'encodeur et du décodeur. Les auteurs utilisent trois critères pour l'algorithme de renforcement :

1. simplicité : les auteurs utilisent la mesure d'évaluation SARI (voir section **? ?**). Cependant, ils font le constat que, leurs données d'entraînement n'ayant qu'une référence pour chaque phrase et leurs données étant bruitées, seulement utiliser SARI ne convient pas. Ainsi, ils calculent SARI de la manière attendue mais également en inversant les places de la simplification de référence et de la sortie du système dans le calcul. Le critère de simplicité correspond à la somme pondérée de ces deux calculs de SARI.
2. préservation du sens : les auteurs utilisent un encodeur LSTM qui convertit la phrase d'entrée et la phrase de sortie en deux vecteurs, dont ils calculent la similarité cosinus.

3. grammaticalité : les auteurs entraînent au préalable un modèle de langue sur les phrases simples à l'aide d'un LSTM. Le critère de grammaticalité est la probabilité assignée à la phrase de sortie par ce modèle de langue.

Les auteurs ne rapportent pas le taux de reproduction entre la phrase d'entrée et la phrase de sortie avec la mise en place de l'algorithme de renforcement.

4.2.2. Méthodes et outils d'évaluation de la simplification automatique

Dans cette section, nous nous intéressons à l'évaluation de la simplification automatique car il s'agit d'une question de recherche extrêmement importante. Deux types d'évaluation sont distingués : (1) évaluation automatique ou évaluation quantitative (section 4.2.2.1) et (2) évaluation qui se base sur le jugement humain ou évaluation qualitative (section 4.2.2.2). Nous concluons cette section par une discussion sur la problématique de l'évaluation de la simplification automatique (section 4.2.2.3).

4.2.2.1. Évaluation automatique

Parmi les métriques le plus fréquemment utilisées dans les travaux de simplification automatique, nous pouvons mentionner trois métriques :

1. BLEU (Papineni *et al.*, 2002) est une métrique conçue à l'origine pour évaluer les résultats de traduction automatique. Elle est également utilisée pour évaluer la simplification, qui est alors vue comme une tâche de traduction monolingue. BLEU compare la sortie du système avec les données de référence. Cette métrique donne une indication approximative de la performance d'un système, surtout concernant la grammaticalité et la préservation du sens (Martin *et al.*, 2018). Cependant, comme l'ont observé Sulem *et al.* (2018a), elle est moins efficace pour évaluer la simplicité.
2. SARI (Xu *et al.*, 2016) est la métrique la plus commune exploitée dans les travaux de simplification automatique. Cette métrique a été créée spécifiquement pour cette tâche. Le score SARI est calculé en comparant la sortie du système non seulement avec les données de référence, mais aussi avec les données source. Il est à noter que SARI est plus fiable lorsque plusieurs références sont disponibles (Alva-Manchego *et al.*, 2020b; Zhang & Lapata, 2017).
3. Flesch (Flesch, 1948) est un index de lisibilité. Cet index ne se base que sur des indicateurs formels comme la longueur des phrases et le nombre de syllabes par mot. Il est conçu pour l'anglais. Sa formule est :

$$206,835 - 1,015(\frac{\text{nombre de mots}}{\text{nombre de phrases}}) - 84,6(\frac{\text{nombre de syllabes}}{\text{nombre de mots}})$$

Une adaptation pour le français existe également : l'index Kandel (Kandel & Moles, 1958). La valeur absolue de cet index n'est pas une information en soi : la mesure est décrite comme étant pertinente par le biais des comparaisons.

Ainsi, un score plus élevé qu'un autre signale une meilleure lisibilité. Sa formule est :

$$207 - 1,015(\frac{\text{nombre de mots}}{\text{nombre de phrases}}) - 73,6(\frac{\text{nombre de syllabes}}{\text{nombre de mots}})$$

Ces trois métriques sont implémentées dans un outil dédié à l'évaluation de la simplification automatique, EASSE (Alva-Manchego *et al.*, 2019a).

4.2.2.2. Évaluation humaine

Lors de l'évaluation humaine, les sorties d'un système de simplification sont soumises au jugement d'utilisateurs humains. Nous illustrons l'évaluation humaine par trois exemples issus de la littérature :

1. Dans le travail de Nisioi *et al.* (2017), les instructions données aux juges humains correspondent à trois types de critères.
   — *Correction et nombre de modifications* : les juges comptent les modifications opérées dans une phrase par un système de simplification. Pour qu'une modification ou suppression soit marquée comme correcte, elle doit présever le sens et la grammaticalité de la phrase, tout en rendant la phrase plus simple à lire. Deux anglophones natifs et deux anglophones non-natifs ont participé à cette évaluation ;
   — *Grammaticalité et préservation du sens* : les juges notent chacun de ces deux critères sur une échelle de 1 (très mauvais) à 5 (très bon). Les échelles ne sont pas détaillées et nous ne connaissons pas le degré de liberté sur l'interprétation des critères laissé aux annotateurs. Trois annotateurs anglophones natifs ont participé à cette évaluation. L'accord inter-annotateurs utilisé est le κ de Cohen : il est de 0,78 pour la grammaticalité, ce qui est bon, et de 0,63 pour la préservation du sens, ce qui indique un accord faible (Artstein & Poesio, 2008).
   — *Simplicité* : les juges évaluent la simplicité sur une échelle allant de +2 pour une meilleure simplicité à -2 pour une moins bonne simplicité. Il existe donc cinq valeurs possibles : +2 (beaucoup plus simple), +1 (plutôt plus simple), 0 (difficulté égale), -1 (plutôt plus difficile), -2 (beaucoup plus difficile). Le tableau 4.1 reprend cette échelle. Trois annotateurs anglophones non-natifs ont participé à cette évaluation. L'accord inter-annotateurs utilisé est le κ de Cohen : il est à 0,66, ce qui correspond à un accord assez faible.

Score	Signification
+2	beaucoup plus simple
+1	plutôt plus simple
0	difficulté égale
-1	plutôt plus difficile
-2	beaucoup plus difficile

Tableau 4.1. – Échelle de notation pour la simplicité utilisée par Nisioi *et al.* (2017)

2. L'étude de Zhang & Lapata (2017) offre une description bien moins détaillée : les auteurs indiquent seulement avoir demandé aux juges humains de noter la grammaticalité, la préservation du sens et la simplicité sur une échelle de Likert à cinq points. Les annotateurs ont été recrutés sur la plate-forme Mechanical Turk [1] et se présentaient comme des anglophones natifs. Leur nombre est inconnu. Les auteurs ne rapportent aucune information sur l'accord inter-annotateurs.

3. L'étude de Shardlow & Nawaz (2019) propose deux types d'évaluation. Le premier type est mené par l'un des auteurs et consiste à catégoriser les erreurs par type. La typologie proposée est la suivante :
   — type 1 : modification sans perte ou altération du sens ;
   — type 2 : aucune modification ;
   — type 3 : réduction significative de l'information ;
   — type 4 : une seule substitution lexicale opérée, avec perte ou altération du sens ;
   — type 5 : paraphrase ou reformulation incorrecte ;
   — type 6 : répétition multiple d'un mot du texte d'origine.

   Les auteurs précisent que cette typologie est une échelle allant de l'absence d'erreur à l'erreur la plus grave. Un seul type d'erreur est attribué à chaque phrase simplifiée. Une phrase, qui peut recevoir plus d'un type d'erreur, est marquée avec le type d'erreur le plus grave parmi les types disponibles.

   Le deuxième type d'évaluation mené fait appel à dix annotateurs humains recrutés sur la plate-forme Figure Eight [2]. Comme le travail fait une comparaison entre quatre modèles, les annotateurs reçoivent quatre simplifications (une par modèle) pour une phrase d'origine donnée. Ils doivent classer ces simplifications de la plus simple à la moins simple. Les auteurs ne rapportent pas d'accord inter-annotateurs. Cette approche d'évaluation mène à une évaluation comparative.

Ces illustrations d'approches de l'évaluation humaine nous semblent représentatives des pratiques actuelles en évaluation de la simplification automatique. Nous voyons qu'il existe des différences sur les aspects que l'on mesure : certains travaux mesurent la grammaticalité, la préservation du sens et la simplicité, alors que d'autres s'attachent à caractériser les types d'erreurs. Il peut aussi être difficile de vérifier le détail de ce qui a été demandé aux annotateurs. Même si on dispose de détails sur l'échelle, comme c'est le cas avec le tableau 4.1, reproduire le même type d'évaluation avec d'autres annotateurs reste difficile. Par exemple pour la simplicité, les critères "plutôt plus difficile" ou "plutôt plus simple" (dans l'article d'origine, *somewhat more difficult* et *somewhat simpler*) laissent beaucoup de place à l'interprétation et résultent en un accord inter-annotateurs assez faible. Ainsi, en l'absence de standards consensuels, utiliser le jugement humain dans un but de comparaison des résultats d'un travail à l'autre reste difficile actuellement.

1. https ://www.mturk.com
2. Cette plate-forme a été rachetée par l'entreprise Appen en 2019, https ://appen.com/

4.2.2.3. Discussion

Les travaux dédiés à l'évaluation de la simplification automatique sont peu nombreux actuellement. De plus, certaines métriques d'évaluation, comme BLEU et SARI, sont plus populaires que d'autres. Ainsi, aucun travail n'est dédié à l'étude des mesures de lisibilité comme Flesch ou Kandel. Dans ce qui suit, nous proposons une discussion autour des métriques BLEU et SARI, qui sont assez incontournables actuellement. Plusieurs chercheurs montrent ainsi leurs limites et essayent de proposer d'autres métriques ou approches qui seraient plus justes dans l'évaluation de la simplification automatique.

Martin *et al.* (2018) tentent de proposer une évaluation automatique sans recours à des données de référence, en comparant des métriques simples comme la longueur des mots, leur nombre de syllabes, la longueur des phrases, etc. avec des métriques automatiques de traduction automatique dont BLEU. Les auteurs précisent que le périmètre de leur étude est assez étroit en raison du faible volume de données servant de base à leur travail. Ils concluent que le sujet doit être exploré plus en profondeur. Cependant, leur travail dégage une plutôt bonne corrélation de BLEU avec la grammaticalité et la préservation du sens. Sulem *et al.* (2018a) partagent ce dernier constat, en y ajoutant une nuance : cette bonne corrélation ne s'applique pas aux cas de découpage de phrases. Ils observent également des corrélations négatives de BLEU avec le critère de simplicité. Alva-Manchego *et al.* (2020a) ont trouvé de faibles corrélations entre le jugement humain et BLEU.

Les concepteurs de SARI mettent en avant de bonnes corrélations avec le jugement humain pour la simplicité (Xu *et al.*, 2016). Cependant, Alva-Manchego *et al.* (2020a) trouvent de faibles corrélations entre le jugement humain et SARI sur ce critère.

À l'heure actuelle, l'évaluation de la simplification automatique de textes est donc une question ouverte, sur laquelle aucun progrès substantiel n'a été réalisé récemment. Pourtant, la nécessité d'une méthode d'évaluation fiable est d'autant plus pressante que les méthodes actuelles s'appuient sur l'apprentissage profond, ce qui implique que les modifications apportées aux textes ne peuvent pas être expliquées d'une manière précise. Il est donc difficile d'évaluer, mais surtout de comparer, les performances de différents systèmes de simplification automatique. Les travaux décrits en section 4.2.1.3 s'attachent à optimiser les scores SARI et BLEU mais les écarts entre leurs différents résultats sont assez ténus. À cela s'ajoute le fait que la comparaison est compliquée par les données de référence utilisées : le corpus le plus souvent utilisé actuellement pour l'anglais, Newsela, n'est pas librement accessible et ne propose pas un découpage standard en ensembles d'entraînement, de test et de validation.

L'évaluation des systèmes est une problématique actuelle qui n'est pas limitée à la simplification automatique, mais touche plusieurs domaines voisins :

— la communauté de recherche en traduction automatique remet en question la pertinence de BLEU pour l'évaluation (Mathur *et al.*, 2020) ;
— la communauté de recherche en génération automatique de texte remet en question ses pratiques sur le jugement humain (Belz *et al.*, 2020b,a).

En parallèle, de nouvelles métriques basées sur les avancées récentes du traitement automatique des langues commencent à être développées. Nous prenons l'exemple

de BERTScore, un score basé sur les plongements lexicaux de BERT (Zhang *et al.*, 2020). Il s'agit d'une métrique destinée à la génération automatique de textes, et plus spécifiquement à l'évaluation de l'équivalence sémantique, sur le modèle de BLEU. Les auteurs produisent une représentation de la phrase de référence et de la phrase à évaluer à l'aide des plongements contextuels de BERT, puis les comparent en utilisant la similarité cosinus. À notre connaissance, au moment de la rédaction de ce travail, cette métrique n'a pas encore été utilisée dans un travail de simplification automatique.

4.3. Expériences en simplification basées sur le modèle de traduction neuronale

En accord avec les derniers travaux en simplification automatique, nous proposons une série d'expériences en simplification en exploitant une approche neuronale. Nous avons travaillé à la construction d'un corpus parallèle dans les chapitres précédents. Dans le chapitre actuel, notre objectif consiste donc à utiliser ce corpus pour la simplification automatique de textes médicaux. La possibilité de disposer de telles données nous permet d'utiliser des méthodes neuronales. Notons que ce type d'approches en simplification automatique est utilisé sur des textes en anglais (Cooper & Shardlow, 2020; Nisioi *et al.*, 2017) et a également été utilisé sur le français (Abdul Rauf *et al.*, 2020).

La série d'expériences que nous avons menées poursuit plusieurs objectifs :

1. mesurer l'importance du volume de données pour la simplification automatique neuronale ;
2. mesurer l'influence de données spécialisées de bonne qualité sur les performances d'un système de simplification de textes de spécialité ;
3. étudier l'utilité d'un lexique qui établit la correspondance entre des termes spécialisés et leurs équivalents grand public.

Nous décrivons d'abord les données utilisées (section 4.3.1). Nous présentons ensuite la série d'expériences menées (section 4.3.2), ainsi que les méthodes d'évaluation adoptées (section 4.3.3). Nous analysons ensuite les résultats obtenus (section 4.3.4). Nous terminons en faisant le bilan des travaux effectués en simplification automatique (section 4.4).

4.3.1. Données linguistiques

Nous utilisons deux corpus parallèles créés pour la simplification :

— Le premier corpus est construit à partir du corpus CLEAR, à l'aide de la méthode décrite au chapitre précédent (section 3.3, page 41). Ce corpus comptait 4 596 couples de phrases équivalentes lors de la réalisation des expériences présentées ici.

— Le deuxième corpus est obtenu grâce à la traduction automatique de WikiLarge (Zhang & Lapata, 2017) de l'anglais vers le français. WikiLarge est une compilation de trois corpus publiés précédemment pour la simplification automatique de textes en anglais (Zhu *et al.*, 2010a; Woodsend & Lapata, 2011;

Kauchak, 2013), tous issus de Wikipedia. La traduction de ce corpus vers le français a été effectuée par nous en utilisant `OpenNMT-py` (Klein *et al.*, 2017) avec les paramètres par défaut et le modèle anglais-français. Ce corpus contient près de 300 000 couples de phrases. Nous utilisons ce corpus pour avoir un volume de données assez conséquent pour l'utilisation d'une méthode neuronale. Cette utilisation d'une traduction automatique a été explorée pour le français par Abdul Rauf *et al.* (2020) et a montré des résultats encourageants. Cependant, cette étude utilise une traduction de Newsela, en conséquence le corpus n'est pas disponible.

Le tableau 4.2 montre le volume des données (couples de phrases et nombres d'occurrences) de ces deux corpus.

	Total		*Entraînement*		*Validation*		*Test*	
Corpus	*Paires*	*Occurrences*	*Paires*	*Occurrences*	*Paires*	*Occurrences*	*Paires*	*Occurrences*
WikiLarge FR	297 494	12 753 567	296 402	12 695 192	992	42 676	100	4 302
CLEAR	4 596	226 149	4 196	206 500	300	7 381	100	4 965

Tableau 4.2. – Taille des deux corpus parallèles exploités, WikiLarge FR et CLEAR

Nous avons divisé le corpus CLEAR en trois ensembles : entraînement, validation et test. Le découpage est le suivant : 100 exemples pour le test, trois fois plus pour la validation, et le reste pour l'entraînement. WikiLarge FR est déjà divisé en ces trois ensembles. Nous avons seulement réduit aléatoirement l'ensemble de test de 359 exemples à 100 pour une meilleure comparaison des résultats d'un corpus à l'autre. Comme ces deux corpus contiennent des données de Wikipedia, nous avons vérifié s'il n'y avait pas de couples de phrases identiques dans les deux corpus et nous n'en avons trouvé aucun.

Nous utilisons par ailleurs un lexique qui propose des paraphrases grand public pour des termes techniques médicaux (Koptient & Grabar, 2020), comme *{hypotension}{baisse de la tension artérielle}*. Ce lexique a été constuire à l'aide de terminologies médicales (Lindberg *et al.*, 1993) et de corpus en français (corpus CLEAR et différents forums médicaux de discussions en ligne). Lors de la réalisation de nos expériences, le lexique comptait 7 580 paraphrases pour 4 516 termes médicaux.

4.3.2. Protocole expérimental

L'objectif du protocole expérimental consiste à évaluer sur deux aspects :

1. *Impact des corpus de langue générale et médicale.* Pour évaluer cet impact, nous utilisons différents ratios de données issues des deux corpus (WikiLarge FR et CLEAR). Comme notre but est la simplification de textes biomédicaux et que le corpus CLEAR est de petite taille, nous utilisons toujours l'intégralité des ensembles d'entraînement et de validation de CLEAR. Par ailleurs, nous ajoutons graduellement des exemples de WikiLarge FR à l'ensemble d'entraînement : nous commençons avec un modèle entraîné sur le même nombre d'exemples des deux corpus (ratio 1 : 1) et ajoutons des exemples de WikiLarge FR pour atteindre successivement les ratios 1 : 5, 1 : 10, 1 : 25, 1 : 50 et

1 : 75. Le ratio 1 : 75 est une approximation et correspond au cas où l'intégralité du corpus d'entraînement de WikiLarge FR est utilisé. Chaque ensemble d'exemples supplémentaires de WikiLarge FR est sélectionné aléatoirement et ajouté à l'ensemble du ratio précédent. Ces expériences permettent d'observer d'éventuelles différences de robustesse des modèles lorsqu'ils sont testés soit sur des données en langue biomédicale soit sur des données en langue générale. Elles permettent aussi d'observer si ces exemples supplémentaires améliorent les résultats.

2. *Impact du lexique*. Nous menons les mêmes expériences avec les différents ratios des deux corpus et nous y ajoutons le lexique. Le lexique est exploité de deux manières :
   — pendant l'entraînement, où nous ajoutons le lexique à l'ensemble d'entraînement ;
   — pendant la phase de simplification directement, avec le paramètre `--phrase_table` de OpenNMT-py qui est prévu pour le traitement des mots inconnus. Ce paramètre a été utilisé de la même manière dans un travail sur la simplification de lettres cliniques en anglais (Shardlow & Nawaz, 2019).

   Nous effectuons donc trois séries d'expériences :
   — SL : sans utilisation de lexique (figure 4.1),
   — LE : utilisation du lexique pendant l'entraînement (figure 4.2),
   — LS : utilisation du lexique pendant la simplification (figure 4.3).

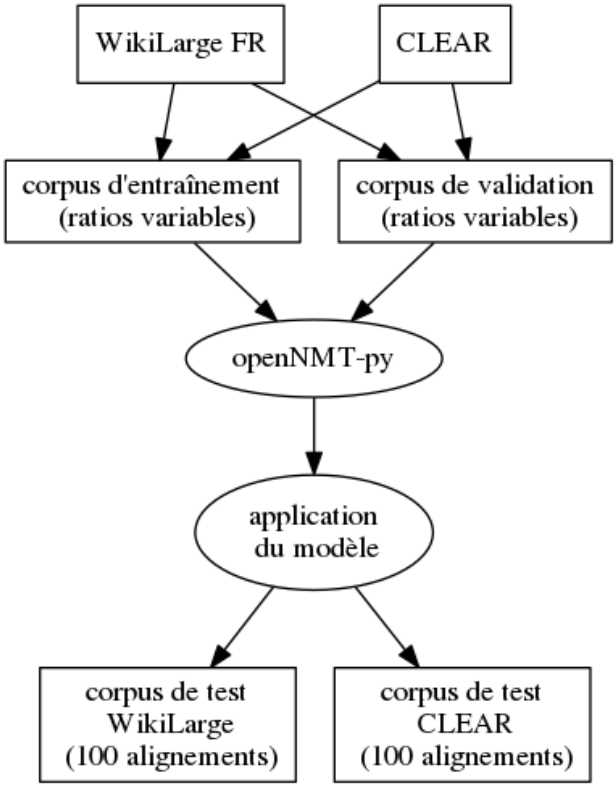


FIGURE 4.1. – Étapes des expériences *SL*.

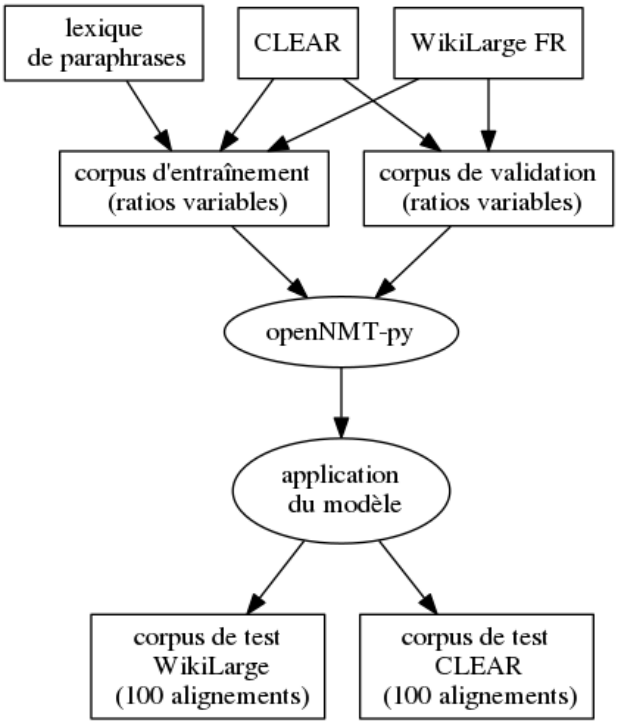


FIGURE 4.2. – Étapes des expériences *LE*.

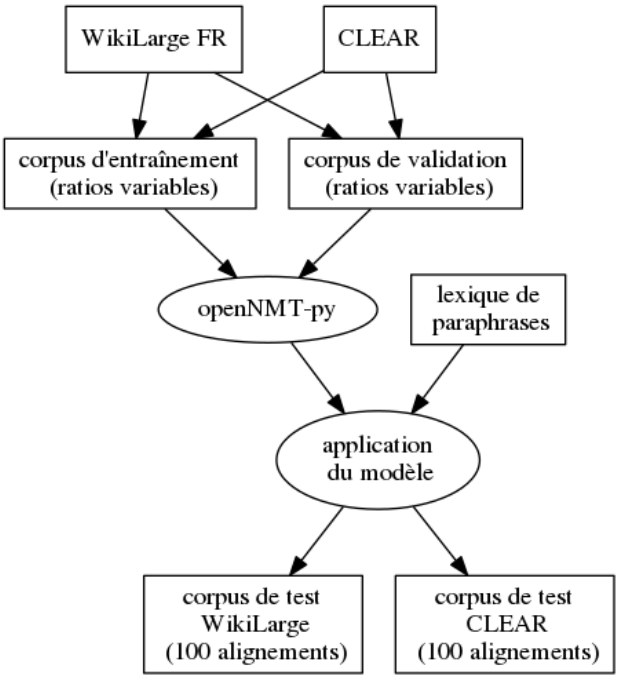


FIGURE 4.3. – Étapes des expériences *LS*.

Lors de ces différentes expériences, les ensembles de validation et de test restent les mêmes. Les ensembles d'entraînement varient selon les ratios et l'utilisation du lexique.

À l'étape de simplification, nous utilisons OpenNMT-py avec la configuration suivante : deux couches de LSTM bidirectionnels de 500 unités pour l'encodeur et le décodeur, optimiseur `ADAM`, taux d'apprentissage 0,001, probabilité de *dropout* 0,3, probabilité de *dropout* pour l'attention de 0,2. Chaque entraînement distinct a pris environ cinq heures sur un GPU Geforce RTX 2070. Lors de la phase de simplification, nous utilisons le paramètre `--replace_unk` pour forcer le programme de copier les mots inconnus de l'entrée vers la sortie, sauf pour la série d'expériences *LS* où ce paramètre est remplacé par `--phrase_table` avec pour valeur le fichier contenant le lexique. Les hyperparamètres utilisés ont été inspirés par un travail existant sur la simplification automatique de textes en français (Abdul Rauf *et al.*, 2020).

Nous rapportons également les résultats de trois *baselines* :

— un modèle entraîné sur CLEAR uniquement,
— un modèle entraîné sur WikiLarge FR uniquement,
— une *baseline* appelée *identique*, où les sorties sont identiques aux entrées. Cette *baseline* n'est pas caractérisée par une simplification, mais sert plutôt de base de comparaison.

4.3.3. Évaluation

En nous appuyant sur l'état de l'art, nous utilisons deux types de méthodes pour l'évaluation : (1) l'évaluation quantitative ou automatique et (2) l'évaluation qualitative ou manuelle.

Nous évaluons les résultats de manière quantitative avec les trois métriques automatiques : BLEU, SARI et Kandel, présentées en section 4.2.2.1. Bien que ces scores fassent l'objet d'une remise en question, il n'existe aucune méthode d'évaluation plus performante à l'heure actuelle. Comme l'objectif de notre travail n'ambitionne pas de proposer de nouvelles formes d'évaluation de la simplification automatique, nous nous limitons à utiliser ces trois métriques existantes. Nous calculons les deux premières métriques avec la suite d'évaluation EASSE spécifiquement conçue pour la simplification automatique de textes (Alva-Manchego *et al.*, 2019a). Nous avons modifié notre installation de EASSE pour remplacer la formule de Flesch par la formule de Kandel.

Nous menons également une évaluation qualitative : nous analysons manuellement les exemples produits par différents modèles en les comparant avec les *baselines*. Nous évaluons aussi les résultats obtenus sur les données de test de CLEAR avec le meilleur modèle généré selon les métriques automatiques. Cette analyse qualitative est menée avec les trois critères que sont la grammaticalité, la préservation du sens et la simplicité. Comme nous l'avons vu en section 4.2.2.2, il n'y a pas de pratique standard pour le jugement humain. Nous avons donc élaboré des échelles de jugement que nous utilisons dans cette étude. Chaque phrase produite par le système de simplification reçoit un score allant de 1 à 5 pour chacun de ces trois critères. Les tableaux 4.3, 4.4 et 4.5 résument les échelles proposées et utilisées :

— *Grammaticalité* : 5 est attribué à une phrase correcte ; 4 est attribué en cas de présence d'une erreur mineure ; 3 en cas de présence de plus d'une erreur mineure ; 2 est attribué en cas d'erreur de syntaxe ; 1 est attribué si la phrase

est incompréhensible ;
— *Préservation du sens* : 5 correspond à l'équivalence sémantique ; 4 signale la perte ou la modification d'une petite information ; 3 signale la perte ou la modification d'une information importante ; 2 signale la perte du sens principal ; 1 signale une absence de lien avec la phrase d'origine ou une phrase incompréhensible ;
— *Simplicité* : 5 signale la présence de plus d'un bon phénomène de simplification (un bon phénomène de simplification est par exemple une substitution lexicale correcte ou une suppression pertinente) ; 4 signale la présence d'un bon phénomène de simplification ; 3 signale une phrase identique à l'originale ou qui ne présente pas de transformations notables ; 2 signale l'introduction d'une difficulté ; 1 signale une phrase bien plus difficile à lire.

Score	Signification
1	phrase incompréhensible
2	erreur de syntaxe
3	présence de plus d'une erreur mineure
4	présence d'une erreur mineure
5	aucune erreur

Tableau 4.3. – Échelle de notation pour la grammaticalité

Score	Signification
1	phrase incompréhensible
2	perte du sens principal
3	perte ou modification d'une information importante
4	perte ou modification d'une information mineure
5	équivalence sémantique

Tableau 4.4. – Échelle de notation pour la préservation du sens

Score	Signification
1	phrase beaucoup plus difficile à lire
2	introduction d'une difficulté de lecture
3	phrase identique à l'originale ou sans transformation notable
4	présence d'un bon phénomène de simplification
5	plus d'un bon phénomène de simplification

Tableau 4.5. – Échelle de notation pour la simplicité

Une sortie du système de simplification qui serait identique à la phrase d'origine obtient donc 5 en grammaticalité, 5 en préservation du sens et 3 en simplicité. Cette évaluation manuellem est menée par l'auteur.

4.3.4. Résultats

Nous présentons d'abord les résultats quantitatifs (section 4.3.4.1), calculés avec trois métriques automatiques : BLEU provenant de la traduction automatique, SARI spécifiquement conçu pour la simplification automatique et Kandel, qui est une adaptation au français de l'indice de lisibilité Flesch. Ces métriques sont détaillées en section 4.2.2.1.

Nous présentons ensuite une évaluation qualitative, effectuée grâce à une évaluation manuelle des trois critères que sont la grammaticalité, la préservation du sens et la simplicité (section 4.3.4.2).

4.3.4.1. Évaluation quantitative

	WikiLarge FR			*CLEAR*		
Modèle	*BLEU*	*SARI*	*Kandel*	*BLEU*	*SARI*	*Kandel*
Identique	60,02	25,05	81,15	55,00	23,73	76,67
WikiLarge FR	39,08	37,61	89,71	9,72	30,97	95,58
CLEAR	0,15	20,52	94,32	21,59	22,07	84,15
SL 1 :1	5,83	25,60	**98,20**	26,23	38,10	**84,26**
SL 1 :5	14,82	30,38	96,23	29,86	39,20	80,43
SL 1 :10	33,74	35,01	92,97	41,05	38,32	80,02
SL 1 :25	25,88	34,44	92,26	37,24	**40,34**	78,12
SL 1 :50	44,48	**38,93**	90,52	49,16	35,36	79,09
SL 1 :75	**49,67**	38,02	89,71	**50,23**	33,91	79,11
LS 1 :5	15,06	30,28	**103,29**	30,00	39,10	82,06
LS 1 :10	33,70	35,12	102,17	40,29	38,32	79,42
LS 1 :25	26,16	34,44	99,84	37,17	**40,09**	79,04
LS 1 :50	44,49	**39,05**	100,63	**48,16**	35,33	**89,08**
LS 1 :75	**49,70**	38,26	97,76	47,61	34,27	78,43
LE 1 :5	23,98	33,68	95,56	39,07	**40,94**	87,36
LE 1 :10	30,94	34,05	94,61	38,17	36,38	86,72
LE 1 :25	**37,29**	34,74	91,40	42,92	39,14	88,22
LE 1 :50	32,68	**36,73**	**98,81**	**49,72**	37,52	90,60
LE 1 :75	34,20	36,47	89,05	40,16	38,58	**92,35**

Tableau 4.6. – Scores des métriques d'évaluation obtenus avec les différentes expériences sur les ensembles de test de WikiLarge FR et CLEAR. *SL* = sans lexique, *LS* = lexique pendant la simplification, *LE* = lexique pour l'entraînement

Le tableau 4.6 montre les scores SARI, BLEU et Kandel obtenus avec les différents modèles sur les données de test de WikiLarge FR et CLEAR. Pour ces trois métriques, plus un score est élevé, plus il correspond à de bonnes performances. Nous rappelons ici que le score Kandel est décrit par son concepteur comme une mesure qui s'interprète par la comparaison : il ne s'agit pas d'une valeur absolue.

Nous rapportons les scores pour le ratio 1 : 1 seulement une fois dans le tableau, pour la série d'expériences SL. Nous ne jugeons pas utile de rapporter ces scores pour tous les modèles car ces scores sont extrêmement faibles en raison du faible volume de données lors de l'entraînement.

Nous faisons d'abord des observations sur les scores SARI et BLEU. Nous commençons par les modèles *baselines* (section 4.3.4.1.1), puis nous commentons les résultats sur ces deux scores sous l'angle de l'impact du volume des données (section 4.3.4.1.2) et de l'impact de la spécialisation des données (section 4.3.4.1.3). Enfin, nous commentons les scores obtenus avec l'index Kandel (section 4.3.4.1.4).

4.3.4.1.1. Baselines

Les trois premières lignes du tableau rapportent les résultats des *baselines*. Nous pouvons voir que l'entraînement sur les données de CLEAR seulement donne de mauvais résultats sur la langue générale avec un score BLEU inférieur à 1 et un score SARI de 20,52. Les performances sont également mauvaises sur la langue médicale bien que BLEU soit plus élevé que sur la langue générale (21,59 *vs* 0,15). Le modèle entraîné sur WikiLarge FR seulement montre d'assez bonnes performances sur les données de test de WikiLarge FR (BLEU à 39,08) et de très mauvaises performances sur CLEAR (BLEU à 9,72). Cette observation montre que les textes de domaines de spécialité doivent être traités avec des données (corpus, lexiques) spécifiques. Les résultats de ces *baselines* indiquent également que pour la simplification de textes spécialisés avec des méthodes neuronales deux critères sont requis :

— disposer de grands volumes de données,
— disposer de données du domaine.

Dans les deux section qui suivent (4.3.4.1.2 et 4.3.4.1.3), nous cherchons à identifier si les données doivent satisfaire ces deux critères ou si un petit volume de données spécialisées peut être compensé par un grand volume de données non spécialisées.

4.3.4.1.2. Impact du volume de données et du lexique

Quand les données des deux corpus sont utilisées pour l'entraînement, les scores vont jusqu'à 49,7 pour BLEU (modèle *LE* 1 : 75) et 39,05 pour SARI (*LS* 1 : 50) sur les données de WikiLarge FR. Sur les données de CLEAR, les scores vont jusqu'à 50,23 pour BLEU (*SL* 1 : 75), alors que tous les scores SARI sont supérieurs à 35. Cela indique que le volume de données est un paramètre crucial avec ce type de méthodes. Nous continuerons à analyser les résultats pour BLEU et SARI uniquement sur les modèles avec les deux ratios les plus élevés, 1 : 50 et 1 : 75. En effet, en-deçà les résultats n'ont pas une grande signification en raison du plus faible volume de données. En ce qui concerne le lexique, il a peu d'impact lorsqu'il est utilisé directement lors de la phase de simplification. Cependant, selon les résultats de plusieurs modèles (*LE* 1 : 50, *LE* 1 : 75), le lexique s'avère efficace lorsqu'il est intégré aux données d'entraînement.

4.3.4.1.3. Impact de la spécialisation des données d'après BLEU et SARI

Sans surprise, les modèles *SL* et *LS* obtiennent des scores semblables lorsqu'ils sont appliqués aux données de test de WikiLarge FR : le lexique n'intervient que lorsque

des termes médicaux sont rencontrés dans les phrases de test. Nous rapportons les scores relatifs à WikiLarge FR obtenus pour BLEU et SARI ici :

— BLEU : 44,48 et 44,49 pour *SL* et *LS* à 1 : 50, 49,67 et 49,70 pour *SL* et *LS* à 1 : 75 ;
— SARI : 38,93 et 39,05 pour *SL* et *LS* à 1 : 50, 38,02 et 38,26 pour *SL* et *LS* à 1 : 75.

Cependant, nous pouvons observer que les performances sont dégradées lorsque l'on ajoute le lexique spécialisé aux données d'entraînement :

— BLEU descend à 32,68 (1 : 50) et 34,20 (1 : 75) pour *LE*,
— SARI descend à 36,73 (1 : 50) et à 36,47 (1 : 75) pour *LE*.

Cette dégradation des performances semble indiquer qu'il est difficile pour un modèle de traiter à la fois la langue générale et la langue spécialisée : en augmentant le volume de données spécialisées, les performances sur la langue générale diminuent.

En ce qui concerne les résultats sur les données de CLEAR, selon les modèles, les résultats sont assez comparables aux résultats obtenus sur les données WikiLarge FR. Une exception à cela est le score BLEU de 40,16 pour *LE* 1 : 75, que nous avons des difficultés à expliquer, car le modèle *LE* 1 : 50 ne connaît pas cette diminution du score par rapport aux modèles *SL* et *LS* appliqués aux mêmes données avec le même ratio. Par ailleurs, nous pouvons noter une légère augmentation des scores de SARI pour les modèles *LE*, ce qui semble indiquer l'utilité du lexique.

4.3.4.1.4. Kandel

À propos du score Kandel, nous pouvons observer qu'aucun modèle ne produit une sortie avec une valeur inférieure à celle de la *baseline* appelée *identique*. Le modèle qui obtient le score Kandel le plus élevé (103,29 contre 81,15 pour les phrases d'origine) est *LS* 1 : 5. L'exemple (2) montre que la phrase transformée est en effet plus courte et les mots ont moins de syllabes. Cependant, nous voyons que ce modèle ne simplifie pas réellement la phrase d'origine : il ne fait que proposer une séquence de mots qui n'a pas de sens. Nous voyons donc que les mesures quantitatives présentent des limites dans l'évaluation de la simplification automatique : les substitutions lexicales correctes peuvent baisser les scores, alors que l'ordonnancement de mots sans une vraie sémantique peut fournir des scores élevés. Il apparaît ainsi qu'avec cette métrique, aucun modèle ne produit de sorties qui détériorent la lisibilité de la phrase au niveau formel, sans se prononcer sur la sémantique cependant.

(2) *Sydney est la plus grande ville et capitale de la Nouvelle-Galles du Sud.*
L'Aquarium de la Nouvelles-Galles du Sud.

4.3.4.2. Évaluation qualitative

Dans cette section, nous menons une évaluation qualitative. Nous prenons d'abord deux exemples réussis, un provenant de WikiLarge FR et un de CLEAR, et commentons les transformations appliquées par les différents modèles utilisés (section 4.3.4.2.1). Puis, à partir de nos observations, nous choisissons un modèle pour mener une évaluation manuelle supplémentaire sur trois critères (la grammaticalité, la préservation du sens et la simplicité) et nous en analysons les réussites et les échecs (section 4.3.4.2.2).

Modèle	*Test sur WikiLarge FR*
Source	Le 14 octobre 1960, le candidat à la présidence John F. Kennedy a proposé le concept de ce qui est devenu le Peace Corps sur les marches de l'Union du Michigan.
Référence	John F. Kennedy, un candidat à la présidence, a proposé l'idée de ce qui devint le Peace Corps sur les marches de l'Union du Michigan le 14 octobre 1960.
Wikilarge	*Aucun changement*
CLEAR	le cancer est de la médecine
SL 1 :25	En 1960, le candidat du président John F. Kennedy a suggéré le Peace Corps sur les marches de l'Union du Michigan.
SL 1 :50	En 1960, le candidat à la présidence John F. Kennedy a proposé l'idée du Peace Corps sur les marches de l'Union du Michigan.
SL 1 :75	*Aucun changement*
LE 1 :25	Le 14 octobre 1960, le candidat au président John F. Kennedy a proposé le concept de ce qui est devenu le Peace Corps sur les marches de l'Union du Michigan.
LE 1 :50	En 1960, le candidat à la présidence John F. Kennedy a proposé l'idée de ce qui est devenu le Peace Corps sur les marches de l'Union du Michigan.
LE 1 :75	John F. Kennedy a proposé le concept de ce qui est devenu le Peace Corps sur les marches de l'Union du Michigan.

Tableau 4.7. – Exemples de simplification d'une phrase de WikiLarge FR

4.3.4.2.1. Exemples commentés

Le tableau 4.7 montre des exemples de simplification d'une phrase de l'ensemble de test de WikiLarge FR, alors que le tableau 4.8 montre des exemples de simplification d'une phrase de l'ensemble de test de CLEAR.

Dans le tableau 4.7, nous prenons pour exemple la phrase source *Le 14 octobre 1960, le candidat à la présidence John F. Kennedy a proposé le concept de ce qui est devenu le Peace Corps sur les marches de l'Union du Michigan.* et la phrase de référence tirée de WikiLarge FR *John F. Kennedy, un candidat à la présidence, a proposé l'idée de ce qui devint le Peace Corps sur les marches de l'Union du Michigan le 14 octobre 1960.* La *baseline* WikiLarge FR n'apporte aucun changement. Le modèle *baseline* entraîné avec CLEAR et appliqué aux données de WikiLarge FR propose une phrase grammaticale, mais qui n'a pas la moindre relation sémantique avec l'exemple : « *Le cancer est de la médecine* ». Cela attire l'attention sur le fait que cette méthode de simplification est très sensible au contenu des données d'entraînement. Nous voyons en effet que le modèle *baseline* entraîné seulement sur les données de CLEAR produit seulement des phrases composées de mots issus du domaine médical, quelle que soit la phrase à simplifier. Nous observons de vraies améliorations de la qualité des simplifications obtenues avec les autres modèles (dont le remplacement de *le candidat du/au président* proposé par les modèles *SL* et *LE* au ratio 1 : 25 par *le candidat à la présidence* proposé par les mêmes modèles au ratio 1 : 50). En effet, à l'exception de la *baseline* CLEAR, les autres exemples de simplification de WikiLarge FR sont comparables à ce que l'on peut observer dans les travaux représentant l'état de l'art en anglais, comme l'atteste l'exemple de simplification neuronale en (3) et extrait d'un travail existant (Cooper & Shardlow, 2020).

(3) *On October 14, 1960, Presidential candidate John F. Kennedy proposed the concept of what became the Peace Corps on the steps of Michigan Union.*
On October 14, 1960, President John F. Kennedy suggested what became the Peace Corps on the steps of Michigan Union.

En ce qui concerne les résultats sur CLEAR (tableau 4.8), nous illustrons les transformations sur une phrase issue du corpus *Médicaments*, dans l'exemple (4). Le document source est destiné aux professionnels de santé, alors que la simplification de référence provient de la notice du médicament correspondant. Les seuls changements dans la simplification de référence sont la troncation de *minutes* en *min* et la suppression de la mention d'une autre rubrique du document.

Les exemples *SL* sont proches de la référence à l'exception de la troncation de *minutes*, alors que les modèles *LS* et *LE* s'en éloignent : *LS* 1 : 75 transforme *hypotension* en *tension inférieure à la normale* par le remplacement mécanique. Cependant, bien que la substitution soit correcte, le résultat n'est pas grammatical : le terme d'origine étant suivi de *artérielle*, le résultat *tension inférieure à la normale artérielle* présente un ordre des mots incorrect pour le français. En revanche, le modèle *LE 1 : 75* propose une simplification plus réussie : la phrase obtenue est grammaticale, son sens est préservé et le terme technique est explicité. Nous reproduisons cette simplification à l'exemple (5). Notons que *diminution de la tension*

Modèle	*Test sur CLEAR*
Source	une hypotension artérielle peut être observée en cas d'administration intraveineuse trop rapide, inférieure à 60 minutes (voir rubrique 4.2)
Référence	une hypotension artérielle peut être observée en cas d'administration intraveineuse trop rapide, inférieure à 60 min
WikiLarge	Une artérielle artérielle peut être observée en cas de crise intraveineuse trop rapide et inférieure à 60 minutes
CLEAR	le traitement de la naissance de la naissance de l' repos [...] de l' repos de la médecine [...] de la médecine de la peau [...] de la peau de la
SL 1 :50 & 1 :75	une hypotension artérielle peut être observée en cas d'administration intraveineuse trop rapide, inférieure à 60 minutes
LS 1 :75	une tension inférieure à la normale artérielle peut être observée en cas d' administration intraveineuse trop rapide, inférieure à 60 minutes
LE 1 :50	une hypotension artérielle peut être observée en cas d'administration intraveineuse trop rapide, inférieure à 60 minutes (voir rubrique « 3)
LE 1 :75	une diminution de la tension artérielle peut être observée en cas d' administration intraveineuse trop rapide, inférieure à 60 minutes

Tableau 4.8. – Exemples de simplification d'une phrase de CLEAR

Grammaticalité	*Préservation du sens*	*Simplicité*
3,16	2,82	3,09

Tableau 4.9. – Scores de grammaticalité, préservation du sens et simplicité sur les 100 simplifications produites par le modèle *LE* sur l'ensemble de test du corpus CLEAR

artérielle est une des équivalences pour *hypotension* dans le lexique que nous utilisons. Ainsi, lorsque le lexique est incorporé dans les données d'entraînement, les substitutions et transformations lexicales sont plus naturelles et grammaticales par comparaison avec la méthode qui exploite le lexique directement lors de l'étape de simplification.

(4) *Une hypotension artérielle peut être observée en cas d'administration intraveineuse trop rapide, inférieure à 60 minutes (voir rubrique 4.2).*
Une hypotension artérielle peut être observée en cas d'administration intraveineuse trop rapide, inférieure à 60 min.

(5) *Une hypotension artérielle peut être observée en cas d'administration intraveineuse trop rapide, inférieure à 60 minutes (voir rubrique 4.2).*
Une diminution de la tension artérielle peut être observée en cas d'administration intraveineuse trop rapide, inférieure à 60 minutes.

4.3.4.2.2. Grammaticalité, préservation du sens et simplicité

Nous évaluons les trois critères du jugement humain sur les sorties du modèle *LE*

1 : 75. Ce modèle a en effet montré les scores quantitatifs parmi les plus élevés au sein de modèles dont les sorties sont exploitables. De plus, nous avons vu avec l'exemple du tableau 4.8 qu'il s'agit d'un modèle qui exploite le plus efficacement le lexique spécialisé. Le lexique médical étant particulièrement opaque, la substitution réussie de termes par leurs paraphrases est une caractéristique cruciale pour la tâche que nous cherchons à accomplir. Le tableau 4.9 montre les résultats de notre évaluation manuelle. Notons que seulement cinq des cent simplifications de l'ensemble de test étaient identiques à la phrase source. Nous proposons ici notre interprétation des scores d'évaluation manuelle du tableau 4.9 : 3,16 pour la grammaticalité, 2,82 pour la préservation du sens et 3,09 pour la simplicité :

— Le *score de grammaticalité* est plutôt faible. En effet, nous avons observé un phénomène récurrent tout au long de l'évaluation : le bégaiement de la méthode, comme la production de sorties contenant des suites de type *de la de la de la...* tel qu'illustré dans les exemples (6), (7) et (8). Le bégaiement révèle la présence de mots inconnus pour le modèle dans la phrase source. Cela veut dire que ces mots ne figurent ni dans le corpus ni dans le lexique. Le lexique spécialisé contient un volume non négligeable de termes (7 580 paraphrases pour 4 516 termes). Cependant, il n'est pas suffisant pour traiter les documents du domaine, car les corpus médicaux contiennent une terminologie beaucoup plus variée. Une solution pourrait être d'avoir recours à des terminologies médicales existantes, comme la SNOMED (Côté, 1996) qui contient environ 150 000 entrées. Cependant, de telles terminologies nécessiteraient un travail d'adaptation pour la tâche de simplification automatique car elles ne renseignent pas sur le degré de lisibilité des termes.
— La *préservation du sens* est plutôt faible également. Cela est la conséquence directe du bégaiement de la méthode en réaction aux mots inconnus : les termes importants sont souvent des termes spécialisés et ceux-ci sont perdus lors de la simplification. La couverture du corpus d'entraînement et du lexique s'avère insuffisante pour répondre aux besoins de la tâche de simplification.
— Le *score de simplicité* est proche du score de 3. Bien que ce soit une moyenne, il reflète le peu de phénomènes de simplification trouvés dans la sortie du modèle. Les mots inconnus en sont la cause : le modèle ne peut pas procéder à une substitution s'il n'a pas rencontré un terme pendant l'entraînement et si le terme est absent du lexique. En revanche, lorsqu'un terme est présent dans le lexique, la substitution peut être réalisée correctement. Au niveau de la syntaxe, nous avons observé peu de transformations également. Une exception à cela est que les incises sont quasiment systématiquement supprimées, comme le montre l'exemple (9), qui représente de surcroît un cas de simplification lexicale réussie : *l'heptaminol* est remplacé par *ce médicament*, en plus d'autres transformations.

(6) *Si un méningiome est diagnostiqué chez un patient traité par androcur, le traitement devra être arrêté (voir rubrique 4.3).*
Si un un est diagnostiqué chez un patient traité par par, le traitement devra être arrêté.

(7) *L'infection par l'hépatite B chez le nouveau-né augmente le risque d'infection*

chronique, de cirrhose ou de cancer du foie (carcinome hépatocellulaire) chez l'adulte.
L'infection par l'hépatite B chez le nouveau-né augmente le risque d'infection chronique, de cirrhose ou de cancer du foie (carcinome carcinome) chez l'adulte.

(8) *Ce médicament contient du benzoate de benzyle (contenu dans le parfum composé) et de l'acide benzoïque (provenant du benzoate de benzyle) et peut provoquer une irritation de la peau, des yeux et des muqueuses.*
Ce médicament contient du du de de (contenu dans le parfum composé) et de l'acide acide (provenant du du de de) et peut provoquer une irrtation de la peau, des yeux et des muqueuses.

(9) *Il est préférable, par mesure de précaution, de ne pas utiliser l'heptaminol au cours de la grossesse.*
Il est préférable de ne pas utiliser ce médicament pendant la grossesse.

Plusieurs des phénomènes décrits peuvent se retrouver dans une même phrase simplifiée par le modèle. C'est notamment le cas pour la suppression d'incises et le bégaiement, les phénomènes les plus courants dans les exemples analysés. Nous illustrons ceci avec l'exemple en (10).

(10) *L'Arum titan, Amorphophallus titanus, également appelé le « phallus de titan » est une espèce de plantes de la famille des Aracées.*
L' L' Arum est une espère de plantes de la famille des des.

Malgré les limites observées, comme le bégaiement de la méthode, nous considérons cependant que nous avons obtenu de très bonnes simplifications avec le modèle *LE 1 : 75*, comme nous avons pu l'observer dans le tableau 4.8 ou dans l'exemple (9). Il nous semble donc que, déjà dans son état actuel, ce modèle pourrait servir de base pour un outil d'aide à la simplification manuelle. Les sorties d'un tel outil devraient être retravaillées lors d'une étape de post-édition, comme cela se fait en traduction automatique. En témoigne l'exemple (11) qui produit du bégaiement (*notamment notamment*, *et et*, *recherches recherches*), mais qui opère également des suppressions et substitutions réussies :

— la suppression des noms de bases de données bibliographiques (*medline*, *embase*, *cinahl*),
— le remplacement du segment *des recherches manuelle d'actes de conférences pertinents et l'examen des références bibliographiques dans les études identifiées et d'autres revues* par le segment plus succinct *ainsi que par des recherches recherches d'actes de conférences*,

Notons que l'apparition d'un mot incongru dans cet exemple (*enfant* en fin de phrase), est un phénomène que nous n'avons pas retrouvé ailleurs.

(11) *Les essais pertinents ont été identifiés par des recherches électroniques dans de nombreuses bases de données de littérature, notamment medline, embase et cinahl, ainsi que par des recherches manuelles d'actes de conférences pertinents et l'examen des références bibliographiques dans les études identifiées et d'autres revues.*

Les essais inclus ont été identifiés par des recherches électroniques dans de nombreuses bases de données, notamment notamment, , et et, ainsi que par des recherches recherches d'actes de conférences enfant.

4.4. Conclusion

Nous avons présenté une série d'expériences autour de la simplification de textes médicaux en français à l'aide d'une méthode neuronale issue de la traduction automatique. Cette série d'expériences poursuit trois objectifs principaux :

— vérifier l'impact du volume des corpus,
— vérifier l'impact de la spécialisation des corpus,
— vérifier l'impact d'une ressource externe, en l'occurrence d'un lexique qui associe des termes médicaux à des paraphrases plus facilement compréhensibles.

Il a été montré que les deux premiers aspects étaient cruciaux. En effet, l'entraînement avec un faible volume de données spécialisées donne de très mauvaises performances sur la langue générale et la langue médicale, alors que l'entraînement avec un grand volume de données non spécialisées donne de bonnes performances sur la langue générale mais de très mauvaises performances sur la langue spécialisée. Enfin, avec toutes les données, spécialisées et non spécialisées, à notre disposition, les performances sont bonnes sur la langue spécialisée mais se dégradent sur la langue générale.

Le lexique n'a pas d'incidence sur les résultats quantitatifs. Cependant, lors de l'évaluation qualitative, nous avons pu remarquer que l'utilisation du lexique lors de l'entraînement conduit à une meilleure intégration des paraphrases.

Cette étude s'est heurtée à un manque de ressources de deux types. En premier lieu, le faible volume de données en français nous a amené à nous servir également d'un corpus traduit automatiquement de l'anglais vers le français. Il serait plus intéressant de pouvoir mener ces expériences avec des corpus français originaux et volumineux. Cependant, de telles ressources n'existent pas encore pour le français. La seconde limitation est liée au manque de données spécialisées. Comme nous l'avons évoqué en section 4.3.4.2.2, une terminologie médicale comme SNOMED contient environ 150 000 termes. Bien que les 150 000 termes ne figurent pas tous nécessairement dans le corpus, nous pouvons émettre l'hypothèse qu'avec un corpus de 4 000 couples de phrases spécialisées et un lexique de 7 580 paraphrases pour 4 516 termes médicaux, la couverture du domaine de ce lexique reste insuffisante pour un processus de simplification de langue médicale intégralement automatisé et réussi. Il n'en reste pas moins que les phénomènes que nous avons pu observer lors de l'évaluation qualitative, comme les bonnes substitutions ou la suppression d'informations superflues, sont intéressants et peuvent représenter des suggestions de simplifications qui peuvent ensuite être reprises manuellement. Pour conclure, nos résultats montrent que la méthode que nous proposons ouvre des perspectives sur une tâche de simplification de textes médicaux en français et de textes de langue générale en français assistée par ordinateur.

Enfin, nous avons vu que l'évaluation de la simplification automatique était une question encore très ouverte. Il n'existe pas de stabilité au niveau des corrélations observées entre les métriques automatiques et le jugement humain. Il n'existe pas

de stabilité non plus au niveau des pratiques de l'évaluation par le jugement humain. Bien que ces évaluations donnent des informations sur les performances des systèmes de simplification automatique, la comparaison des systèmes est difficile. Cela complique la poursuite de la recherche dans le domaine.

5. Conclusion

Nous avons proposé un travail sur la simplification de textes médicaux en langue française. Aucune ressource n'existait pour mener ce travail lorsque nous l'avons commencé. Nous avons donc réalisé notre travail en plusieurs étapes indispensables, depuis la collecte et la constitution de corpus jusqu'à la réalisation d'expériences avec des méthodes issues de l'état de l'art et leur évaluation.

Nous faisons le bilan des contributions apportées et décrivons leurs limites ainsi que les perspectives qu'elles ouvrent. Les contributions se positionnent sur trois axes : (1) la collecte et l'analyse des données, (2) la méthode d'extraction de corpus parallèle et, plus largement, le travail sur la similarité sémantique et (3) les premières expériences en simplification automatique de textes médicaux en français.

5.1. Collecte et analyse du corpus comparable

Nous avons collecté et compilé des textes de diverses sources qui ont le point commun de présenter des informations sur les mêmes sujets mais en deux versions différentes : une version pour un public de spécialistes et une version pour un public néophyte. Ce corpus contient trois types de documents :

— des informations sur les médicaments, qui sont créées à destination des professionnels de santé (les RCP ou les Résumés des caractéristiques du produit) et des patients (les notices de médicaments) ;
— les résumés des revues systématiques Cochrane faites à partir de la littérature médicale. Ces résumés sont créés à destination des professionnels de santé et du grand public ;
— des articles encyclopédiques à destination du grand public (Wikipedia) et des enfants (Vikidia).

Nous avons aligné manuellement un sous-ensemble de documents de ce corpus comparable avec pour objectif la création des données de référence. Ces données de référence permettent de travailler sur l'alignement automatique afin de créer un corpus de phrases parallèles et alignées plus conséquent.

À partir de l'alignement manuel effectué, nous avons également construit une typologie des transformations observées lors de la simplification. Une comparaison de notre typologie avec des typologies existantes montre que la substitution lexicale et l'explicitation de termes sont les procédés les plus fréquents lors de la simplification de documents du domaine médical, par comparaison avec le processus de simplification de textes de la langue générale.

Le corpus comparable constitué est librement disponible pour la recherche [1].

1. `http://www.remicardon.eu`

La perspective principale qui s'ouvre à cette étape consiste à utiliser une approche similaire pour la constitution de corpus comparables propres à d'autres domaines de spécialité.

5.2. Constitution du corpus parallèle

Nous avons constitué un corpus parallèle aligné à partir du corpus comparable. Ce travail est abordé comme une tâche de classification binaire, où il s'agit de décider si une paire de phrases doit être classifiée comme alignable ou non alignable. Lors de cette étape, nous avons rencontré deux défis majeurs :

1. Comme nous devons produire toute la combinatoire possible des phrases de chaque couple de document, il existe un très grand déséquilibre entre les deux classes, *aligné* et *non-aligné*.
2. Avec le peu de données annotées à notre disposition, les méthodes de l'état de l'art, basées sur les réseaux de neurones, ne peuvent pas être exploitées.

Pour relever le premier défi, nous avons mis en place un filtre dont l'objectif consiste à éliminer un maxium d'exemples négatifs (phrases non alignables). Ce filtre se base en partie sur des indices formels simples (comme la présence d'un verbe ou la longueur des deux phrases). Il se base également sur la comparaison de sous-arbres syntaxiques obtenus suite à une analyse syntaxique en constituants. Ce filtre nous permet de réduire grandement le déséquilibre, au prix de la perte d'une petite partie des exemples positifs. Une des perspectives de ce travail consiste à affiner la partie syntaxique du filtre afin de rendre le filtre plus souple et de mitiger ainsi la perte de paires de phrases alignables.

Comme, avec le peu de données disponibles pour l'apprentissage, notre méthode ne peut pas se baser sur un réseau de neurones. Nous avons donc testé des classifieurs binaires. Après avoir comparé divers classifieurs, nous avons choisi d'exploiter `Random Forest` qui se montre efficace et robuste avec nos données. Nous exploitons ce classifieur avec une série de descripteurs basés sur des indices formels, des mesures de distance et de similarité, ainsi que sur les plongements lexicaux. Nous avons étudié l'apport de chaque type de descripteurs et avons choisi de tous les conserver dans notre méthode. Ces descripteurs apportent en effet des indices supplémentaires : leur combinaison permet d'obtenir de meilleurs résultats.

Le résultat principal de notre travail lors de cette étape consiste en un corpus constitué de 10 942 couples de phrases alignées. Cette ressource est disponible pour la recherche [2].

Parmi les perspectives principales de cette étape, nous pouvons mentionner (1) l'exploitation de descripteurs qui s'appuyent sur des connaissances externes ou des terminologies, (2) l'intégration de nouveaux modèles de langue pour les plongements lexicaux (comme FlauBERT ou CamemBERT) et (3) l'exploitation de méthodes neuronales grâce à la création d'un corpus d'apprentissage plus riche (10 942 couples de phrases alignées actuellement).

Les travaux effectués lors de cette étape ont été valorisés de plusieurs manières :

2. http ://www.remicardon.eu

— Nous avons soumis des paires de phrases, de langue médicale et de langue générale, à cinq annotateurs. Nous leur avons demandé d'annoter ces paires de phrases selon leur degré de similarité sémantique sur une échelle allant de 0 (aucun lien) à 5 (équivalence sémantique). Les annotateurs étaient libres de décider de leurs propres critères pour l'attribution des valeurs intermédiaires entre 0 et 5. Nous avons collecté les résultats et produit une synthèse des principes d'annotation ainsi que des scores attribués. Le résultat de cette démarche est un corpus de 1 010 paires de phrases annotées en similarité sémantique et disponibles pour la recherche. Le corpus contient les paires de phrases, les notes de chacun des annotateurs, la moyenne des scores des cinq annotateurs pour chaque paire de phrases, ainsi que la note le plus souvent attribuée.
— Les données que nous avons produites lors de cette étape (une partie du corpus parallèle et le corpus de similarité sémantique) ont été exploitées pour deux tâches lors de la campagne d'évaluation DEFT 2020. Ainsi, plusieurs équipes francophones ont exploité ces données. Cela nous a permis, entre autre, de mieux positionner nos travaux par rapport à d'autres approches et méthodes.

5.3. Expériences en simplification automatique

Finalement, nous avons mené des expériences en simplification automatique. Pour cela, nous avons utilisé le corpus parallèle aligné de langue médicale que nous avons produit, une traduction automatique du corpus WikiLarge utilisé par les travaux en simplification en anglais, de même qu'un lexique qui apparie des termes médicaux complexes avec des paraphases accessibles au grand public. La méthode utilisée se base sur un réseau de neurones de type encodeur-décodeur avec mécanisme d'attention. La traduction automatique du corpus anglais vers le français permet de répondre au besoin de disposer d'un plus grand volume de données, nécessaires pour l'utilisation d'une méthode neuronale. Ces expériences ont pu montrer que des données synthétiques (corpus parallèle et aligné traduit de l'anglais vers le français) associées à des données authentiques (le corpus parallèle aligné constitué à partir du corpus comparable) et dédiées à la tâche spécifique de simplification de langue médicale donnent des résultats encourageants. Notamment, il est envisageable d'utiliser les simplifications produites moyennant une étape de post-édition, comme cela se fait communément en traduction automatique.

Le corpus parallèle aligné constitué en français et le corpus WikiLarge FR, traduit automatiquement de l'anglais vers le français, sont librement disponibles pour la recherche.

Cette étape du travail ouvre plusieurs perspectives :

— Il serait préférable de travailler uniquement avec des données authentiques au lieu d'avoir recours à une traduction automatique du corpus constitué en anglais. Bien que les résultats obtenus soient encourageants, nous pensons néanmoins que la création d'un corpus original plus volumineux en français et dédié à la simplification en langue générale et médicale constitue une étape nécessaire pour améliorer nos résultats.
— Une comparaison des performances entre le système que nous avons utilisé et d'autres types d'architectures neuronales représente une autre perspective.

— Nous avons également vu que les approches pour l'évaluation montrent des limites. De vives discussions au sein de la communauté en simplification automatique ont en effet lieu à leur propos. D'une part, les métriques automatiques requièrent de multiples simplifications de référence pour être plus fiables, ce qui est difficile à obtenir. Ces métriques montrent également une certaine instabilité dans leurs corrélations avec le jugement humain. D'autre part, les pratiques d'évaluation humaine ne sont pas consensuelles, car aucun standard d'annotation et de jugement n'existe. Pour ces raisons, la comparaison des performances de différents systèmes de simplification ne permet pas de tirer des conclusions claires et fiables. De plus, comme les méthodes actuelles font intervenir l'apprentissage profond, cela rend plus difficile l'identification précise des transformations apportées lors de la simplification automatique et donc la comparaison entre différents systèmes. La question de l'évaluation des systèmes automatiques est donc plus cruciale que jamais. Il est donc nécessaire que le stade de scepticisme soit dépassé et qu'un progrès substantiel soit réalisé dans le domaine de l'évaluation.

Bibliographie

Abdul Rauf, S., Ligozat, A.-L., Yvon, F., Illouz, G. & Hamon, T. (2020). Simplification automatique de texte dans un contexte de faibles ressources. In C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla & S. Schneider, Eds., *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pp. 332–341, Nancy, France : ATALA.

Abdul-Rauf, S. & Schwenk, H. (2009). On the use of comparable corpora to improve SMT performance. In *European Chapter of the ACL*, pp. 16–23.

Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs, A. M. A. (1999). Health Literacy Report of the Council on Scientific Affairs. *JAMA*, **281**(6), 552–557.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A. & Guo, W. (2013). *SEM 2013 shared task : Semantic textual similarity. In **SEM*, pp. 32–43.

Alva-Manchego, F., Bingel, J., Paetzold, G., Scarton, C. & Specia, L. (2017). Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pp. 295–305, Taipei, Taiwan : Asian Federation of Natural Language Processing.

Alva-Manchego, F., Martin, L., Bordes, A., Scarton, C., Sagot, B. & Specia, L. (2020a). ASSET : A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4668–4679, Online : Association for Computational Linguistics.

Alva-Manchego, F., Martin, L., Scarton, C. & Specia, L. (2019a). EASSE : Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) : System Demonstrations*, pp. 49–54, Hong Kong, China : Association for Computational Linguistics.

Alva-Manchego, F., Scarton, C. & Specia, L. (2019b). Cross-sentence transformations in text simplification. In *Proceedings of the 2019 Workshop on Widening NLP*, pp. 181–184, Florence, Italy : Association for Computational Linguistics.

Alva-Manchego, F., Scarton, C. & Specia, L. (2020b). Data-driven sentence simplification : Survey and benchmark. *Computational Linguistics*, **46**(1), 135–187.

Amoia, M. & Romanelli, M. (2012). SB : mmSystem - using decompositional semantics for lexical simplification. In **SEM 2012*, pp. 482–486, Montréal, Canada.

Artstein, R. & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4), 555–596.

Audiau, A. (2009). *L'information pour tous. Règles européennes pour une information facile à lire et à comprendre*. Technical report, Nous aussi, UNAPEI.

Bahdanau, D., Cho, K. & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*.

Barbu, E., Martín-Valdivia, M. T. & Ureña-López, L. A. (2013). Open book : a tool for helping ASD users' semantic comprehension. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pp. 11–19, Atlanta, Georgia : Association for Computational Linguistics.

Barzilay, R. & Elhadad, N. (2003). Sentence alignment for monolingual comparable corpora. In *EMNLP*, pp. 25–32.

Belkacem, T., Teissedre, C. & Arens, M. (2020). Similarité Sémantique entre Phrases : Apprentissage par Transfert Interlingue. In *Actes de DEFT*, Nancy, France.

Belz, A., Agarwal, S., Shimorina, A. & Reiter, E. (2020a). Reprogen : Proposal for a shared task on reproducibility of human evaluations in nlg. In *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 232–236.

Belz, A., Mille, S. & Howcroft, D. M. (2020b). Disentangling the properties of human evaluation methods : A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 183–194, Dublin, Ireland : Association for Computational Linguistics.

Berkman, N. D., Sheridan, S. L., Donahue, K. E., Halpern, D. J. & Crotty, K. (2011). Low health literacy and health outcomes : An updated systematic review. *Annals of Internal Medicine*, **155**(2), 97–107.

Berkson, J. (1944). Application of the logistic function to bio-essay. *Journal of the American Statistical Association*, **39**, 357–365.

Billami, M. B., François, T. & Gala, N. (2018). Resyf : a French lexicon with ranked synonyms. In ACL, Ed., *27th International Conference on Computational Linguistics*, pp. 2570–2581, Santa Fe, New Mexico, USA.

Bingel, J. & Bjerva, J. (2018). Cross-lingual complex word identification with multitask learning. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 166–174, New Orleans, Louisiana : Association for Computational Linguistics.

Biran, O., Brody, S. & Elhadad, N. (2011). Putting it simply : a context-aware approach to lexical simplification. In *Annual Meeting of the Association for Computational Linguistics*.

Bird, S., Klein, E. & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Bjerva, J., Bos, J., van der Goot, R. & Nissim, M. (2014). The meaning factory : Formal semantics for recognizing textual entailment and determining semantic similarity. In *Workshop on Semantic Evaluation (SemEval 2014)*, pp. 642–646, Dublin, Ireland.

Braud, C. & Denis, P. (2016). Learning connective-based word representations for implicit discourse relation identification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 203–213, Austin, Texas : Association for Computational Linguistics.

Bray, J. & Curtis, J. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, **27**, 325–349.

Brin-Henry, F. (2014). *L'Education Thérapeutique du Patient en Orthophonie*, volume 259 of *Rééducation Orthophonique*. Ortho-Edition.

Brouwers, L., Bernhard, D., Ligozat, A.-L. & Francois, T. (2014). Syntactic sentence simplification for French. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pp. 47–56.

Brunato, D., Cimino, A., Dell'Orletta, F. & Venturi, G. (2016). PaCCSS-IT : A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 351–361, Austin, Texas : Association for Computational Linguistics.

Brunato, D., Dell'Orletta, F., Venturi, G. & Montemagni, S. (2014). Defining an annotation scheme with a view to automatic text simplification. In *CLICIT*, pp. 87–92.

Brunato, D., Dell'Orletta, F., Venturi, G. & Montemagni, S. (2015). Design and annotation of the first Italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pp. 31–41, Denver, Colorado, USA : Association for Computational Linguistics.

Buscaldi, D., Felhi, G., Ghoul, D., Le Roux, J., Lejeune, G. & Zhang, X. (2020). Calcul de similarité entre phrases : quelles mesures et quels descripteurs ? In *Actes de DEFT*, Nancy, France.

Cao, D., Benamar, A., Boumghar, M., Bothua, M., Ould-Ouali, L. & Suignard, P. (2020). Participation d'EDF R&D à DEFT 2020. In *Actes de DEFT*, Nancy, France.

Cardon, R. (2018). Approche lexicale de la simplification automatique de textes médicaux (lexical approach for the automatic simplification of medical texts). In *Actes de la Conférence TALN. Volume 2 - Démonstrations, articles des Rencontres Jeunes Chercheurs, ateliers DeFT*, pp. 159–174, Rennes, France : ATALA.

Cardon, R. & Grabar, N. (2018). Identification of parallel sentences in comparable monolingual corpora from different registers. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pp. 83–93, Brussels, Belgium : Association for Computational Linguistics.

Cardon, R. & Grabar, N. (2019a). Automatic detection of parallel sentences from comparable biomedical texts. In *20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, La Rochelle, France.

Cardon, R. & Grabar, N. (2019b). Détection automatique de phrases parallèles dans un corpus biomédical comparable technique / simplifié (automatic detection of parallel sentences in comparable biomedical corpora). In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, pp. 255–264, Toulouse, France : ATALA.

Cardon, R. & Grabar, N. (2019c). Parallel sentence retrieval from comparable corpora for biomedical text simplification. In *Proceedings of Recent Advances in Natural Language Processing*, pp. 168–177, Varna, Bulgaria.

Cardon, R. & Grabar, N. (2020a). French biomedical text simplification : When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, Barcelona, Spain (online) : Association for Computational Linguistics.

Cardon, R. & Grabar, N. (2020b). A French corpus for semantic similarity. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 6889–6894, Marseille, France : European Language Resources Association.

Cardon, R. & Grabar, N. (2020c). Parallel sentence alignment from biomedical comparable corpora. *Studies in health technology and informatics*, **270**, 362—366.

Cardon, R. & Grabar, N. (2020d). Reducing the search space for parallel sentences in comparable corpora. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pp. 44–48, Marseille, France : European Language Resources Association.

Cardon, R., Grabar, N., Grouin, C. & Hamon, T. (2020a). Actes de la 6e conférence conjointe journées d'études sur la parole (jep, 33e édition), traitement automatique des langues naturelles (taln, 27e édition), rencontre des étudiants

chercheurs en informatique pour le traitement automatique des langues (récital, 22e édition). atelier défi fouille de textes.

Cardon, R., Grabar, N., Grouin, C. & Hamon, T. (2020b). Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques (presentation of the DEFT 2020 challenge : open domain textual similarity and precise information extraction from clinical cases). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pp. 1–13, Nancy, France : ATALA et AFCP.

Carroll, J., Minnen, G., Canning, Y., Devlin, S. & Tait, J. (1998). Practical simplification of English newspaper text to assist aphasic readers. In *AAAI-98 Workshop on Integrating Artifficial Intelligence and Assistive Technology*, pp. 7–10.

Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S. & Tait, J. (1999). Simplifying Text for Language-Impaired Readers. In *EACL*, pp. 269–270.

Caseli, H. M., Pereira, T. F., Specia, L., Pardo, T. A. S., Gasperin, C. & Aluisio, S. M. (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *CICLING*, pp. 1–12.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I. & Specia, L. (2017). SemEval-2017 task 1 : Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 1–14, Vancouver, Canada : Association for Computational Linguistics.

Chandrasekar, R., Doran, C. & Srinivas, B. (1996). Motivations and methods for text simplification. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pp. 1041–1044, Stroudsburg, PA, USA : Association for Computational Linguistics.

Chandrasekar, R. & Srinivas, B. (1997). Automatic induction of rules for text simplification. *Knowledge Based Systems*, **10**(3), 183–190.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). Smote : Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, **16**(1), 321–357.

Chen, S. F. (1993). Aligning sentences in bilingual corpora using lexical information. In *Annual Meeting of the Association for Computational Linguistics*, pp. 9–16.

Chen, T. & Guestrin, C. (2016). Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA : Association for Computing Machinery.

Clough, P., Gaizauskas, R., Piao, S. S. & Wilks, Y. (2002). METER : Measuring text reuse. In *ACL*, pp. 152–159.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

Cooper, M. & Shardlow, M. (2020). CombiNMT : An exploration into neural text simplification models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 5588–5594, Marseille, France : European Language Resources Association.

Coster, W. & Kauchak, D. (2011a). Simple English wikipedia : A new text simplification task. In *Annual Meeting of the Association for Computational Linguistics*, pp. 665–669.

Coster, W. & Kauchak, D. (2011b). Simple English Wikipedia : A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pp. 665–669, Portland, Oregon, USA : Association for Computational Linguistics.

Côté, R. (1996). *Répertoire d'anatomopathologie de la SNOMED internationale, v3.4*. Université de Sherbrooke, Sherbrooke, Québec.

Côté, R. A., Rothwell, D. J., Palotay, J. L., Beckett, R. S. & Brochu, L. (1993). *The Systematised Nomenclature of Human and Veterinary Medicine : SNOMED International*. Northfield : College of American Pathologists.

Cover, T. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, **14**(3), 326–334.

De Belder, J. & Moens, M.-F. (2010). Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*, pp. 19–26 : ACM ; New York.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics.

Devlin, S. & Tait, J. (1998). The use of psycholinguistic database in the simplification of text for aphasic readers. In *Linguistic Database*, pp. 161–173.

Devlin, S. & Unthank, G. (2006). Helping aphasic people process online information. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '06, pp. 225–226, New York, NY, USA : ACM.

Drame, K., Sambe, G., Diop, I. & Faty, L. (2020). Approche supervisée de calcul de similarité sémantique entre paires de phrases. In *Actes de DEFT*, Nancy, France.

Duran, K., Rodriguez, J. & Bravo, M. (2014). Similarity of sentences through comparison of syntactic trees with pairs of similar words. In *11th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, pp. 1–6, Campeche.

Evans, R. & Orasan, C. (2019). Sentence simplification for semantic role labelling and information extraction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pp. 285–294, Varna, Bulgaria : INCOMA Ltd.

Evans, R. & Orăsan, C. (2019). Identifying signs of syntactic complexity for rule-based sentence simplification. *Natural Language Engineering*, **25**(1), 69–119.

Fabre, C., Hathout, N., Ho-Dac, L.-M., Morlane-Hondère, F., Muller, P., Sajous, F., Tanguy, L. & Van de Cruys, T. (2014). Presentation of the SemDis 2014 workshop : distributional semantics for two tasks - lexical substitution and exploration of specialized corpora (présentation de l'atelier SemDis 2014 : sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés) [in French]. In *TALN-RECITAL 2014 Workshop SemDis 2014 : Enjeux actuels de la sémantique distributionnelle (SemDis 2014 : Current Challenges in Distributional Semantics)*, pp. 196–205, Marseille, France : Association pour le Traitement Automatique des Langues.

Feitosa, D. & Pinheiro, V. (2017). Análise de medidas de similaridade semântica na tarefa de reconhecimento de implicação textual (analysis of semantic similarity measures in the recognition of textual entailment task)[in Portuguese]. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pp. 161–170, Uberlândia, Brazil : Sociedade Brasileira de Computação.

Ferguson, T. (1982). An inconsistent maximum likelihood estimate. *Journal of the American Statistical Association*, **77**(380), 831–834.

Fernando, S. & Stevenson, M. (2008). A semantic similarity approach to paraphrase detection. In *Comp Ling UK*, pp. 1–7.

Finnimore, P., Fritzsch, E., King, D., Sneyd, A., Ur Rehman, A., Alva-Manchego, F. & Vlachos, A. (2019). Strong baselines for complex word identification across multiple languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 970–977, Minneapolis, Minnesota : Association for Computational Linguistics.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**(2), 179–188.

Flesch, R. (1948). A new readability yardstick. *Jounal of Applied Psychology*, **23**, 221–233.

Franco-Salvador, M., Gupta, P., Rosso, P. & Banchs, R. E. (2016). Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language. *Knowledge-Based Systems*, **111**, 87 – 99.

François, T., Billami, M. B., Gala, N. & Bernhard, D. (2016). Automatic ranking of synonyms according to their reading and comprehension difficulty. In *JEP-TALN-RECITAL 2016*, volume 2 of *TALN*, pp. 15–28, Paris, France.

Fung, P. & Cheung, P. (2004). Mining very non-parallel corpora : Parallel sentence and lexicon extraction vie bootstrapping and em. In *Conference on Empirical Methods in Natural Language Processing*, pp. 57–63.

Gala, N., Tack, A., Javourey-Drevet, L., François, T. & Ziegler, J. C. (2020). Alector : A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. In *Language Resources and Evaluation for Language Technologies (LREC)*, Marseille, France.

Gale, W. A. & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Comp Linguistics*, **19**(1), 75–102.

Germann, U. (2008). Yawat : Yet Another Word Alignment Tool. In *Proceedings of the ACL-08 : HLT Demo Session*, pp. 20–23, Columbus, Ohio : Association for Computational Linguistics.

Gooding, S. & Kochmar, E. (2019). Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1148–1153, Florence, Italy : Association for Computational Linguistics.

Grabar, N. & Cardon, R. (2018a). CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pp. 3–9, Tilburg, the Netherlands : Association for Computational Linguistics.

Grabar, N. & Cardon, R. (2018b). CLEAR – Simple Corpus for Medical French. In *Workshop on Automatic Text Adaption (ATA)*, pp. 1–11, Tilburg, Netherlands.

Grabar, N., Farce, E. & Sparrow, L. (2018). Étude de la lisibilité des documents de santé avec des méthodes d'oculométrie (study of readability of health documents with eye-tracking methods). In *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, pp. 3–18, Rennes, France : ATALA.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A. & Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

Grégoire, F. & Langlais, P. (2018). Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1442–1453, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

Guo, W. & Diab, M. (2012). Modeling sentences in the latent space. In *ACL*, pp. 864–872.

He, H., Gimpel, K. & Lin, J. (2015). Multi-perspective sentence similarity modeling with convolutional neural networks. In *EMNLP*, pp. 1576–1586, Lisbon, Portugal.

Hewavitharana, S. & Vogel, S. (2011). Extracting parallel phrases from comparable data. In *4th Workshop on Building and Using Comparable Corpora*, pp. 61–68.

Ho, T. K. (1995). Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95, pp. 278, USA : IEEE Computer Society.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, **9**(8), 1735–1780.

Inui, K., Fujita, A., Takahashi, T., Iida, R. & Iwakura, T. (2003). Text simplification for reading assistance : A project note. In *Proceedings of the Second International Workshop on Paraphrasing*, pp. 9–16, Sapporo, Japan : Association for Computational Linguistics.

Jauhar, S. & Specia, L. (2012). UOW-SHEF : SimpLex – lexical simplicity ranking based on contextual and psycholinguistic features. In **SEM 2012*, pp. 477–481, Montréal, Canada.

Johannsen, A., Martínez, H., Klerke, S. & Søgaard, A. (2012). EMNLP@CPH : Is frequency all there is to simplicity ? In **SEM 2012*, pp. 408–412, Montréal, Canada.

Jonnalagadda, S., Tari, L., Hakenberg, J., Baral, C. & Gonzalez, G. (2009). Towards effective sentence simplification for automatic processing of biomedical text. In *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume : Short Papers*, pp. 177–180 : Association for Computational Linguistics.

Kajiwara, T. & Komachi, M. (2016). Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, pp. 1147–1158, Osaka, Japan : The COLING 2016 Organizing Committee.

Kandel, L. & Moles, A. (1958). Application de l'indice de Flesch à la langue française. *Cahiers Études de Radio-Télévision*, **19**, 253–274.

Kauchak, D. (2013). Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pp. 1537–1546, Sofia, Bulgaria : Association for Computational Linguistics.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R. & Fidler, S. (2015). Skip-thought vectors. In *Neural Information Processing Systems (NIPS)*, pp. 3294–3302.

Kitaev, N. & Klein, D. (2018). Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Melbourne, Australia : Association for Computational Linguistics.

Klein, G., Kim, Y., Deng, Y., Senellart, J. & Rush, A. M. (2017). OpenNMT : Open-source toolkit for neural machine translation. In *Proc. ACL*.

Klerke, S. & Søgaard, A. (2012). DSim, a Danish parallel corpus for text simplification. In *LREC*, pp. 4015–4018.

Koptient, A., Cardon, R. & Grabar, N. (2019). Simplification-induced transformations : typology and some characteristics. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 309–318, Florence, Italy : Association for Computational Linguistics.

Koptient, A. & Grabar, N. (2020). Rated Lexicon for the Simplification of Medical Texts. In *The Fifth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing HEALTHINFO 2020*, Porto, Portugal.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, **30**(1), 61–70.

Lai, A. & Hockenmaier, J. (2014). Illinois-LH : A denotational and distributional approach to semantics. In *Workshop on Semantic Evaluation (SemEval 2014)*, pp. 239–334, Dublin, Ireland.

Landis, J. & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Laurent, D., Nègre, S. & Séguéla, P. (2009). L'analyseur syntaxique Cordial dans Passage. In *Traitement Automatique des Langues Naturelles (TALN)*.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L. & Schwab, D. (2020). Flaubert : Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 2479–2490, Marseille, France : European Language Resources Association.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics. Doklady*, **707**(10).

Ligozat, A., Grouin, C., Garcia-Fernandez, A. & Bernhard, D. (2012). ANNLOR : A naïve notation-system for lexical outputs ranking. In **SEM 2012*, pp. 487–492.

Lindberg, D., Humphreys, B. & McCray, A. (1993). The Unified Medical Language System. *Methods Inf Med*, **32**(4), 281–291.

Martin, L., Humeau, S., Mazaré, P.-E., de La Clergerie, É., Bordes, A. & Sagot, B. (2018). Reference-less quality estimation of text simplification systems. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pp. 29–38, Tilburg, the Netherlands : Association for Computational Linguistics.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D. & Sagot, B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7203–7219, Online : Association for Computational Linguistics.

Mathur, N., Baldwin, T. & Cohn, T. (2020). Tangled up in BLEU : Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4984–4997, Online : Association for Computational Linguistics.

McCarthy, D. & Navigli, R. (2007). SemEval-2007 task 10 : English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 48–53, Prague, Czech Republic : Association for Computational Linguistics.

Mihalcea, R., Corley, C. & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, pp. 1–6.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 : 27th Annual Conference on Neural Information Processing Systems*, pp. 3111–3119, Lake Tahoe, Nevada, USA.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 : 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pp. 3111–3119.

Miller, G. A. (1995). Wordnet : A lexical database for english. *Commun. ACM*, **38**(11), 39–41.

Mueller, J. & Thyagarajan, A. (2016). Siamese recurrent architectures for learning sentence similarity. In *AAAI Conference on Artificial Intelligence*, pp. 2786–2792.

Munteanu, D. S. & Marcu, D. (2002). Processing comparable corpora with bilingual suffix trees. In *EMNLP*, pp. 289–295.

Nelken, R. & Shieber, S. M. (2006). Towards robust context-sensitive sentence alignment for monolingual corpora. In *EACL*, pp. 161–168.

Nisioi, S., Štajner, S., Ponzetto, S. P. & Dinu, L. P. (2017). Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pp. 85–91, Vancouver, Canada : Association for Computational Linguistics.

OCHIAI, A. (1957). Zoogeographical studies on the soleoid fishes found in japan and its neighhouring regions-ii. *NIPPON SUISAN GAKKAISHI*, **22**(9), 526–530.

Paetzold, G. & Specia, L. (2016). SemEval 2016 task 11 : Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 560–569, San Diego, California : Association for Computational Linguistics.

Paetzold, G. & Specia, L. (2017). Lexical simplification with neural ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, pp. 34–40, Valencia, Spain : Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Pylieva, H., Chernodub, A., Grabar, N. & Hamon, T. (2019). RNN embeddings for identifying difficult to understand medical words. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 97–104, Florence, Italy : Association for Computational Linguistics.

Qiu, L., Kan, M.-Y. & Chua, T.-S. (2006). Paraphrase recognition via dissimilarity significance classification. In *Empirical Methods in Natural Language Processing*, pp. 18–26, Sydney, Australia.

Quinlan, P. (1992). *The Oxford psycholinguistic database*. Oxford, UK : Oxford University Press.

Rello, L., Baeza-Yates, R., Bott, S. & Saggion, H. (2013). Simplify or help ? text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, W4A '13, New York, NY, USA : Association for Computing Machinery.

Rosenblatt, F. (1958). The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**(6), 386–408.

Rosenblatt, F. (1961). *Principles of Neurodynamics : Perceptrons and the Theory of Brain Mechanisms*. Washington DC : Spartan Books.

Rychalska, B., Pakulska, K., Chodorowska, K., WojciechWalczak & Andruszkiewicz, P. (2016). Samsung Poland NLP team at SemEval-2016 task 1 : Necessity for diversity ; combining recursive autoencoders, wordnet and ensemble methods to measure semantic similarity. In *SemEval-2016*, pp. 614–620.

Sackett, D., Rosenberg, W., Gray, J., Haynes, R. & Richardson, W. (1996). Evidence based medicine : what it is and what it isn't. *BMJ*, **312**(7023), 71–2.

Saggion, H. (2017). *Automatic Text Simplification*, volume 32 of *Synthesis Lectures on Human Language Technologies*. University of Toronto : Morgan & Claypool.

Saggion, H., Bott, S. & Rello, L. (2013). Comparing resources for Spanish lexical simplification. In A. Dediu, C. Martin-Vide, R. Mitkov & B. Truthe, Eds., *Statistical language and speech processing. SLSP 2013. Lecture notes in computer science*, volume 7978. Springer, Berlin, Heidelberg.

Saggion, H., Štajner, S., Bott, S., Mille, S., Rello, L. & Drndarevic, B. (2015). Making it simplext : Implementation and evaluation of a text simplification system for spanish. *ACM Trans. Access. Comput.*, **6**(4).

Scarton, C. & Specia, L. (2018). Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pp. 712–718, Melbourne, Australia : Association for Computational Linguistics.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Int Conf on New Methods in Language Processing*, pp. 44–49.

Seretan, V. (2012). Acquisition of syntactic simplification rules for french. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis, Eds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* : European Language Resources Association (ELRA). ID : unige :30961.

Severyn, A., Nicosia, M. & Moschitti, A. (2013). Learning semantic textual similarity with structural representations. In *Annual Meeting of the Association for Computational Linguistics*, pp. 714–718.

Shardlow, M. (2013). A comparison of techniques to automatically identify complex words. In *ACL Student Research Workshop*, pp. 103–109.

Shardlow, M. (2014). A survey of automated text simplification. *Int J Advanced Computer Science and Applications*, **1**, 1–13.

Shardlow, M. & Nawaz, R. (2019). Neural text simplification of clinical letters with a domain specific phrase table. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 380–389, Florence, Italy : Association for Computational Linguistics.

Siddharthan, A. (2014). A survey of research on text simplification. *ITL - International Journal of Applied Linguistics*, **165**, 259–298.

Sinha, R. (2012). UNT-SimpRank : Systems for lexical simplification ranking. In **SEM 2012*, pp. 493–496.

Smith, D. & Eisner, J. (2006). Quasi-synchronous grammars : Alignment by soft projection of syntactic dependencies. In *Proceedings on the Workshop on Statistical Machine Translation*, pp. 23–30, New York City, USA : Association for Computational Linguistics.

Soğancioğlu, G., Öztürk, H. & Özgür, A. (2017). BIOSSES : a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, **33**(14), i49–i58.

Sparck Jones, K. (1988). *A Statistical Interpretation of Term Specificity and Its Application in Retrieval*, In *Document Retrieval Systems*, pp. 132–142. Taylor Graham Publishing : GBR.

Specia, L., Jauhar, S. & Mihalcea, R. (2012). Semeval-2012 task 1 : English lexical simplification. In **SEM 2012*, pp. 347–355.

Štajner, S. & Popović, M. (2019). Automated text simplification as a preprocessing step for machine translation into an under-resourced language. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pp. 1141–1150, Varna, Bulgaria : INCOMA Ltd.

Sulem, E., Abend, O. & Rappoport, A. (2018a). BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 738–744, Brussels, Belgium : Association for Computational Linguistics.

Sulem, E., Abend, O. & Rappoport, A. (2018b). Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pp. 162–173, Melbourne, Australia : Association for Computational Linguistics.

Tai, K. S., Socher, R. & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *Annual Meeting of the Association for Computational Linguistics*, pp. 1556–1566, Beijing, China.

Tapi Nzali, M. (2020). DEFT 2020 : détection de similarité entre phrases et extraction d'information. In *Actes de DEFT*, Nancy, France.

Tillmann, C. & Xu, J.-M. (2009). A simple sentence-level extraction algorithm for comparable data. In *Companion Vol. of NAACL HLT*.

Tsubaki, M., Duh, K., Shimbo, M. & Matsumoto, Y. (2016). Non-linear similarity learning for compositionality. In *AAAI Conference on Artificial Intelligence*, pp. 2828–2834.

Ukkonen, E. (1992). Approximate string-matching with q-grams and maximal matches. *Theor. Comput. Sci.*, **92**(1), 191–211.

Utiyama, M. & Isahara, H. (2003). Reliable measures for aligning Japanese-English news articles and sentences. In *Annual Meeting of the Association for Computational Linguistics*, pp. 72–79.

van Noord, R., Abzianidze, L., Toral, A. & Bos, J. (2018). Exploring neural methods for parsing discourse representation structures. *Transactions of the Association for Computational Linguistics*, **6**, 619–633.

Vapnik, V. & Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, **24**, 709–715.

Vila, M., Antònia Mart, M. & Rodríguez, H. (2011). Paraphrase concept and typology. A linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, **46**, 83–90.

Štajner, S., Franco-Salvador, M., Ponzetto, S. P. & Rosso, P. (2018). Cats : A tool for customised alignment of text simplification corpora. In *Proceedings of the 11th Language Resources and Evaluation Conference, LREC 2018, Miyazaki, Japan, May 7-12*.

Vu, T., Hu, B., Munkhdalai, T. & Yu, H. (2018). Sentence simplification with memory-augmented neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, pp. 79–85, New Orleans, Louisiana : Association for Computational Linguistics.

Wan, S., Dras, M., Dale, R. & Paris, C. (2006). Using dependency-based features to take the "para-farce" out of paraphrase. In *Australasian Language Technology Workshop*, pp. 131–138.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, **8**(3–4), 229–256.

Williams, S. & Reiter, E. (2005). Generating readable texts for readers with low basic skills. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.

Woodsend, K. & Lapata, M. (2011). Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 409–420, Edinburgh, Scotland, UK. : Association for Computational Linguistics.

Wubben, S., van den Bosch, A. & Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pp. 1015–1024.

Xu, W., Callison-Burch, C. & Napoles, C. (2015). Problems in current text simplification research : New data can help. *Transactions of the Association for Computational Linguistics*, **3**, 283–297.

Xu, W., Napoles, C., Pavlick, E., Chen, Q. & Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, **4**, 401–415.

Yang, C. C. & Li, K. W. (2003). Automatic construction of English/Chinese parallel corpora. *J. Am. Soc. Inf. Sci. Technol.*, **54**(8), 730–742.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A. & Zampieri, M. (2018). A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 66–78, New Orleans, Louisiana : Association for Computational Linguistics.

Yimam, S. M., Štajner, S., Riedl, M. & Biemann, C. (2017). Multilingual and cross-lingual complex word identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pp. 813–822, Varna, Bulgaria : INCOMA Ltd.

Yu, Q., Max, A. & Yvon, F. (2012). Revisiting sentence alignment algorithms for alignment visualization and evaluation. In *The 5th Workshop on Building and Using Comparable Corpora*, pp.1̃0.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. (2020). Bertscore : Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhang, X. & Lapata, M. (2017). Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 584–594, Copenhagen, Denmark : Association for Computational Linguistics.

Zhang, Y. & Patrick, J. (2005). Paraphrase identification by text canonicalization. In *Australasian Language Technology Workshop*, pp. 160–166.

Zhang, Z. & Zweigenbaum, P. (2017). zNLP : Identifying parallel sentences in Chinese-English comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pp. 51–55, Vancouver, Canada : Association for Computational Linguistics.

Zhao, B. & Vogel, S. (2002). Adaptive parallel sentences mining from web bilingual news collection. In *IEEE Int Conf on Data Mining*, pp. 745–748.

Zhao, J., Zhu, T. T. & Lan, M. (2014). ECNU : One stone two birds : Ensemble of heterogenous measures for semantic relatedness and textual entailment. In *Workshop on Semantic Evaluation (SemEval 2014)*, pp. 271–277.

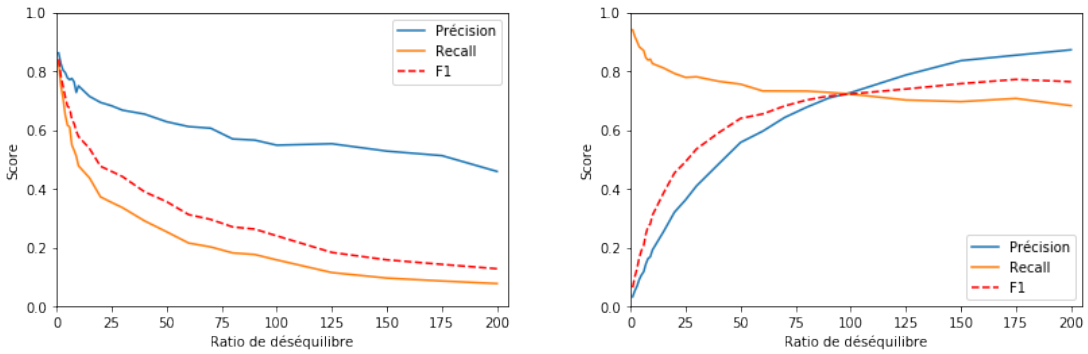
Zhou, W., Ge, T., Xu, K., Wei, F. & Zhou, M. (2019). BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3368–3373, Florence, Italy : Association for Computational Linguistics.

Zhu, Z., Bernhard, D. & Gurevych, I. (2010a). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 1353–1361, Beijing, China : Coling 2010 Organizing Committee.

Zhu, Z., Bernhard, D. & Gurevych, I. (2010b). A monolingual tree-based translation model for sentence simplification. In *COLING 2010*, pp. 1353–1361.

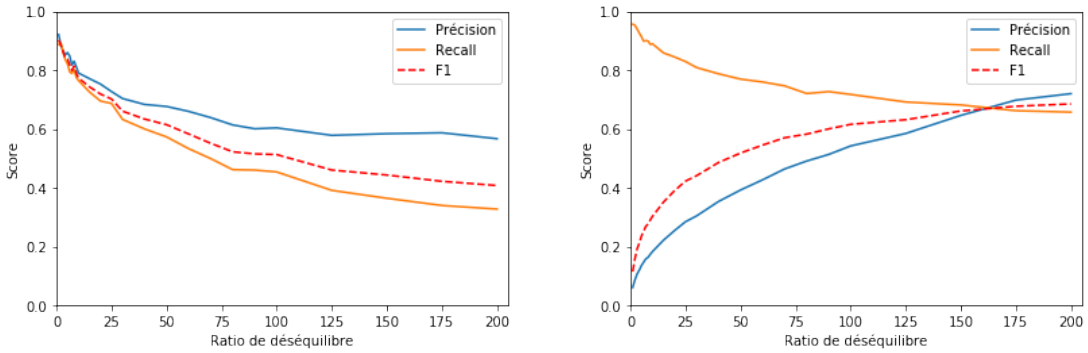
Ștefănescu, D., Ion, R. & Hunsicker, S. (2012). Hybrid parallel sentence mining from comparable corpora. In *16th EAMT Conference*, pp. 137–144.

A. Classification avec différents ratios de déséquilibre par ensembles de descripteurs



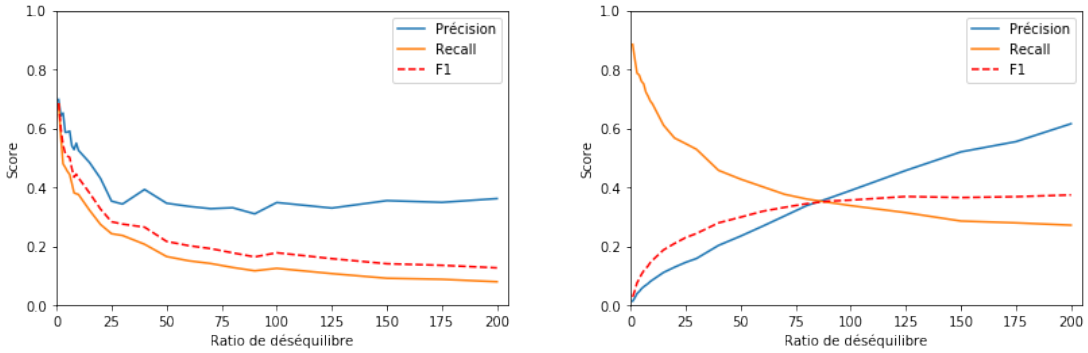
(a) *DDE, descripteurs BL* (b) *DRE, descripteurs BL*

FIGURE A.1. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (BL = *baseline*).



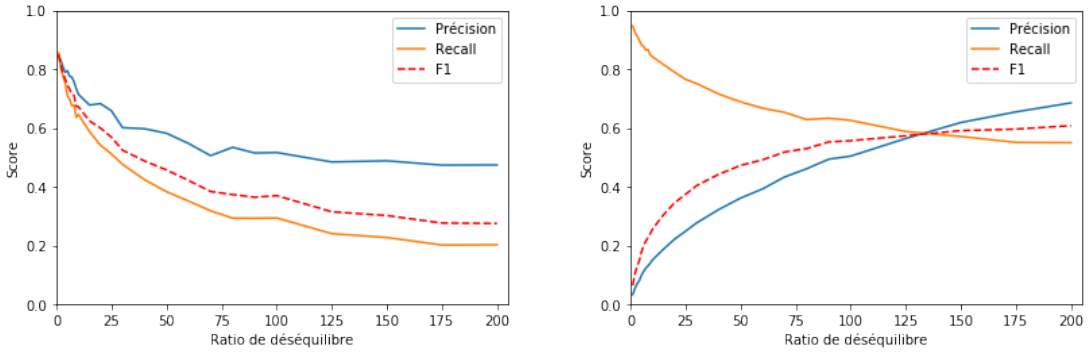
(a) *DDE, descripteurs S* (b) *DRE, descripteurs S*

FIGURE A.2. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (S = mesures de similarité).



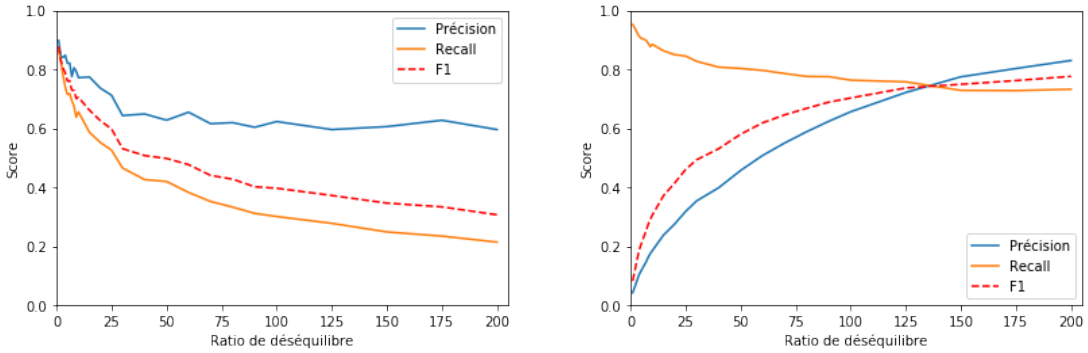
(a) *DDE, descripteurs L* (b) *DRE, descripteurs L*

FIGURE A.3. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (L = Levenshtein).



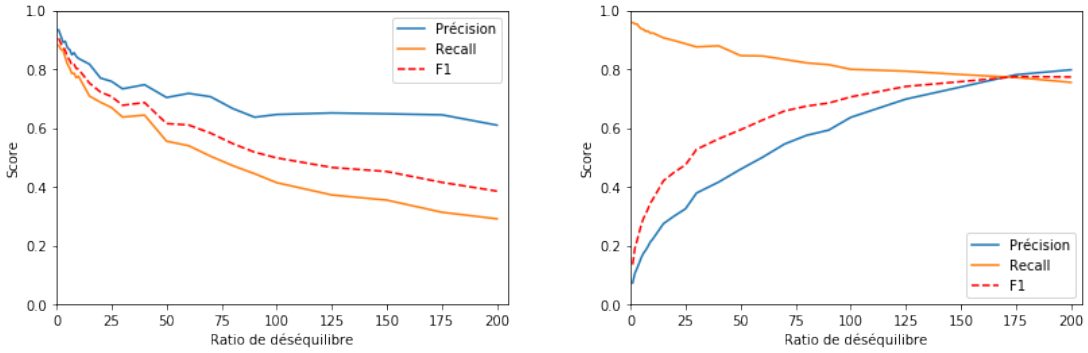
(a) *DDE, descripteurs N* (b) *DRE, descripteurs N*

FIGURE A.4. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (N = *ngrams*).



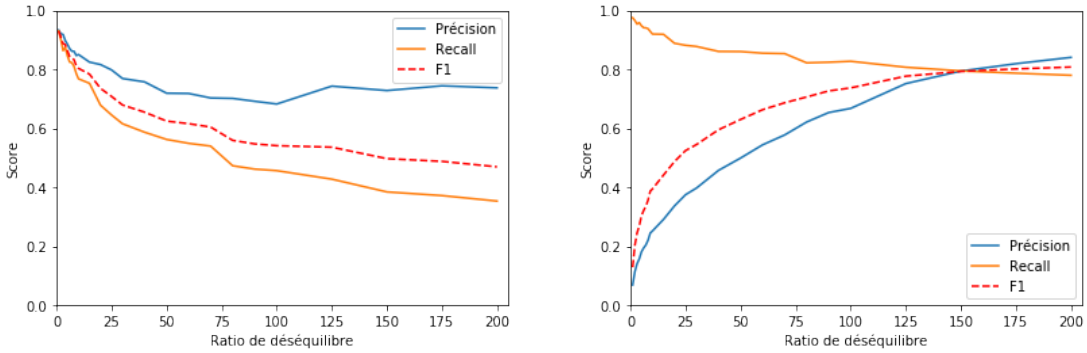
(a) *DDE, descripteurs PL* (b) *DRE, descripteurs PL*

FIGURE A.5. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (PL = plongements lexicaux).



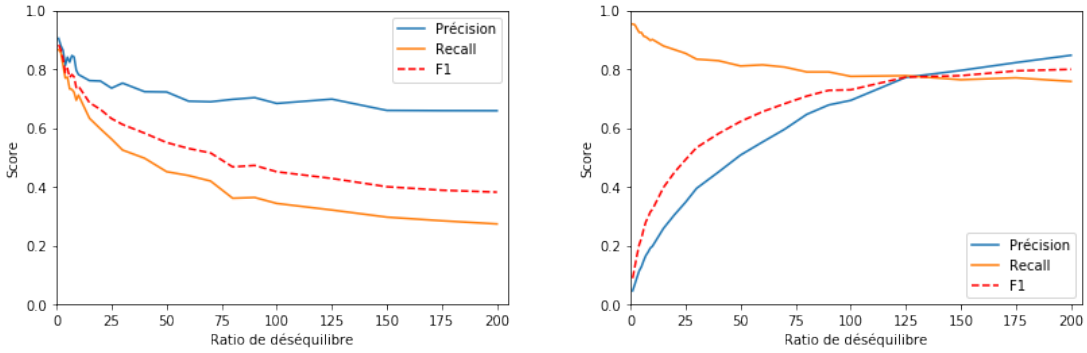
(a) *DDE, descripteurs L,S* (b) *DRE, descripteurs L,S*

FIGURE A.6. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (L = Levenshtein, S = mesures de similarité).



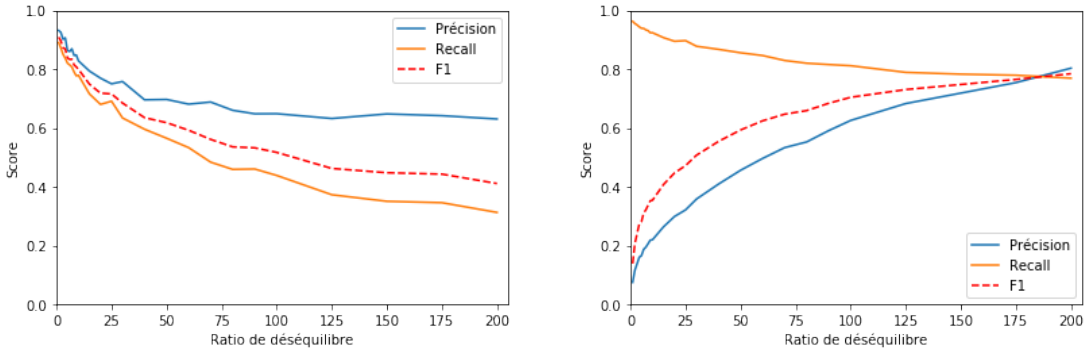
(a) *DDE, descripteurs L,N* (b) *DRE, descripteurs L,N*

FIGURE A.7. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (L = Levenshtein, N = *ngrams*).



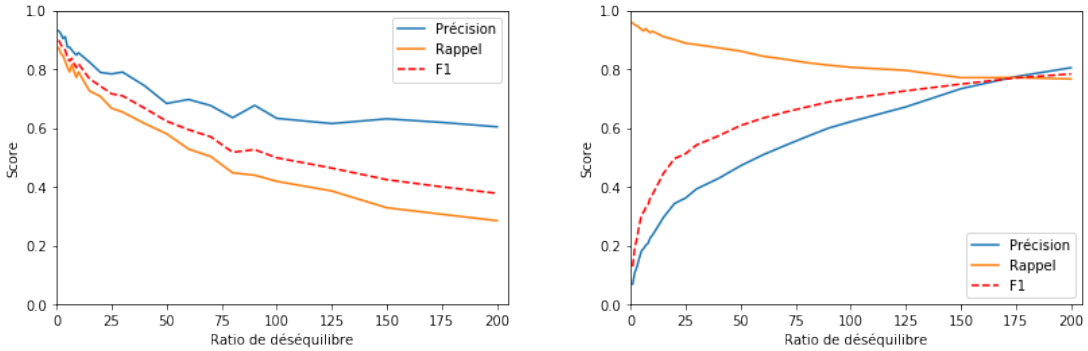
(a) *DDE, descripteurs L,PL* (b) *DRE, descripteurs L,PL*

FIGURE A.8. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (L = Levenshtein, PL = plongements lexicaux).



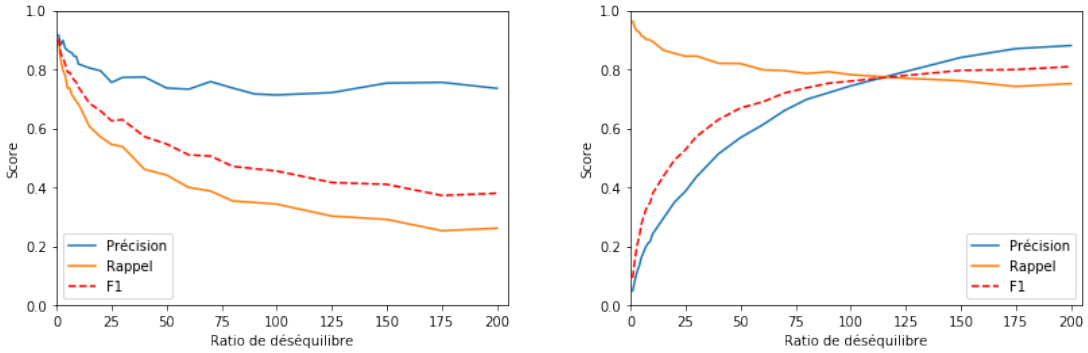
(a) *DDE, descripteurs S,N* (b) *DRE, descripteurs S,N*

FIGURE A.9. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (S = mesures de similarité, N = *ngrams*).



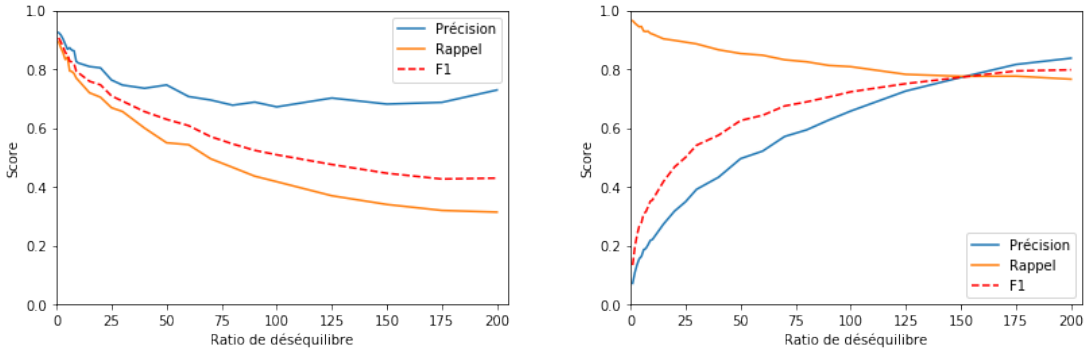
(a) *DDE, descripteurs S,PL* (b) *DRE, descripteurs S,PL*

FIGURE A.10. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (S = mesures de similarité, PL = plongements lexicaux).



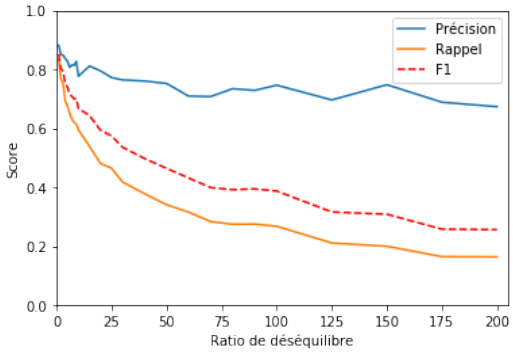
(a) *DDE, descripteurs BL,L* (b) *DRE, descripteurs BL,L*

FIGURE A.11. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (BL = *baseline*, L = Levenshtein).

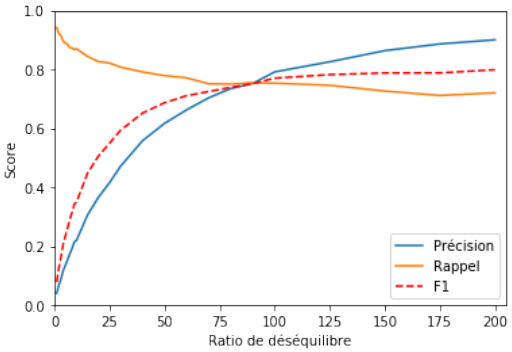


(a) *DDE, descripteurs BL,S* (b) *DRE, descripteurs BL,S*

FIGURE A.12. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (BL = *baseline*, S = mesures de similarité).

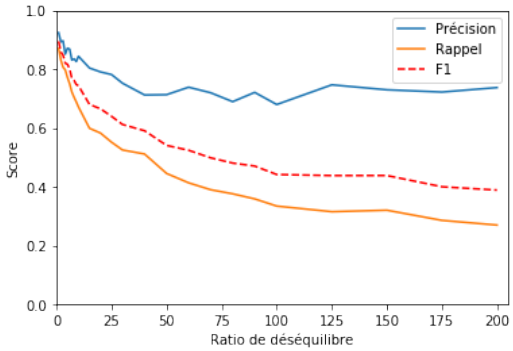


(a) *DDE, descripteurs BL,N*

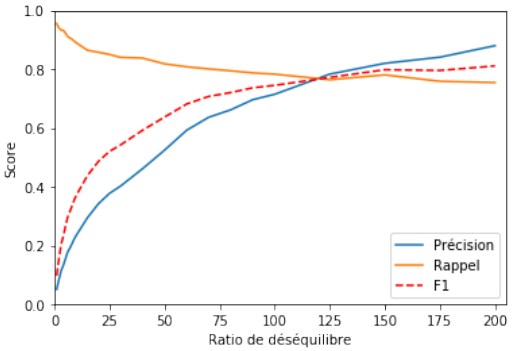


(b) *DRE, descripteurs BL,N*

FIGURE A.13. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (BL = *baseline*, N = *ngrams*).

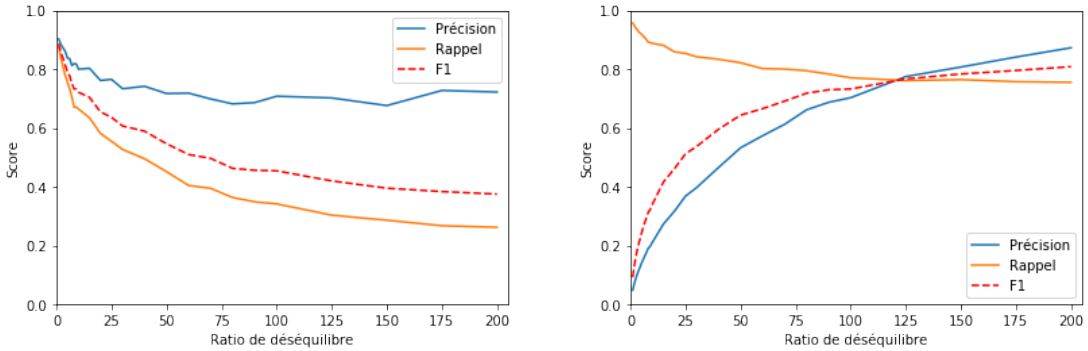


(a) *DDE, descripteurs BL,PL*



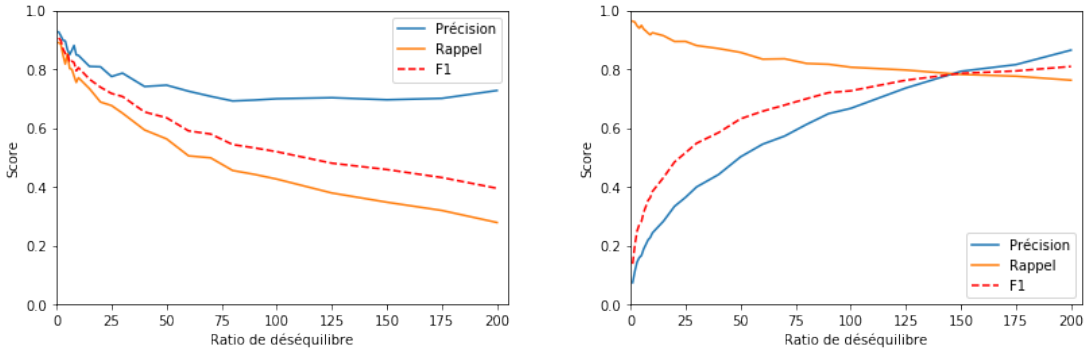
(b) *DRE, descripteurs BL,PL*

FIGURE A.14. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (BL = *baseline*, PL = plongements lexicaux).



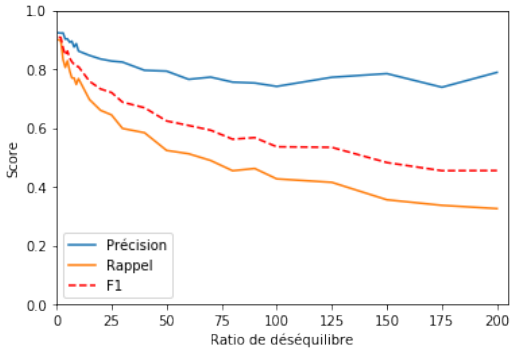
(a) *DDE, descripteurs N,PL* (b) *DRE, descripteurs N,PL*

FIGURE A.15. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (N = *ngrams*, PL = plongements lexicaux).

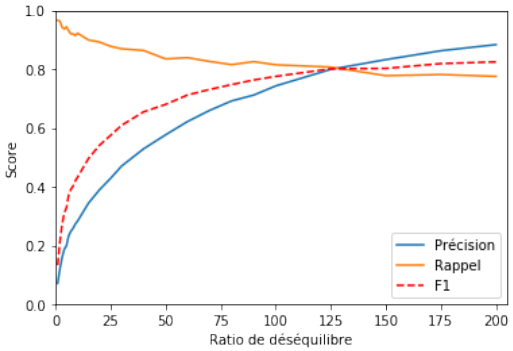


(a) *DDE, descripteurs BL,L,S* (b) *DRE, descripteurs BL,L,S*

FIGURE A.16. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (BL = *baseline*, L = Levenshtein, S = mesures de similarité).

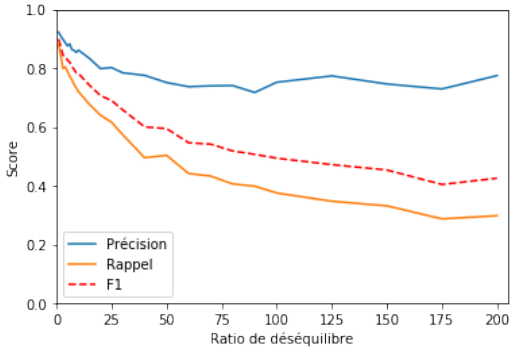


(a) *DDE, descripteurs BL,L,N*

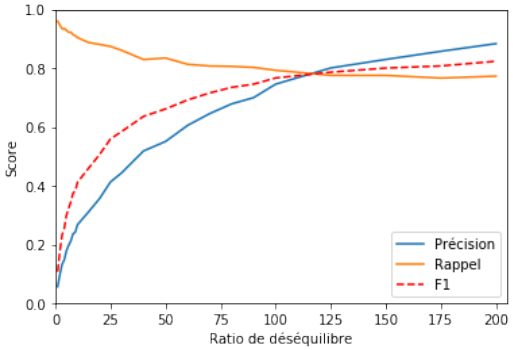


(b) *DRE, descripteurs BL,L,N*

FIGURE A.17. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (BL = *baseline*, L= Levenshtein, N = *ngrams*).

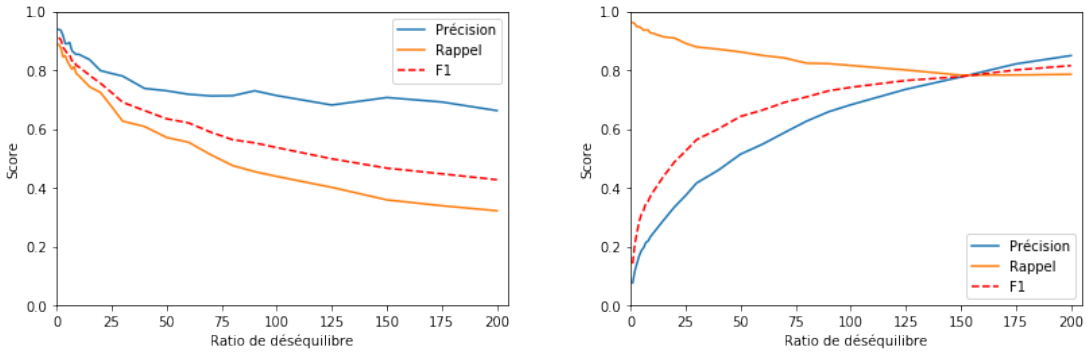


(a) *DDE, descripteurs BL,L,PL*



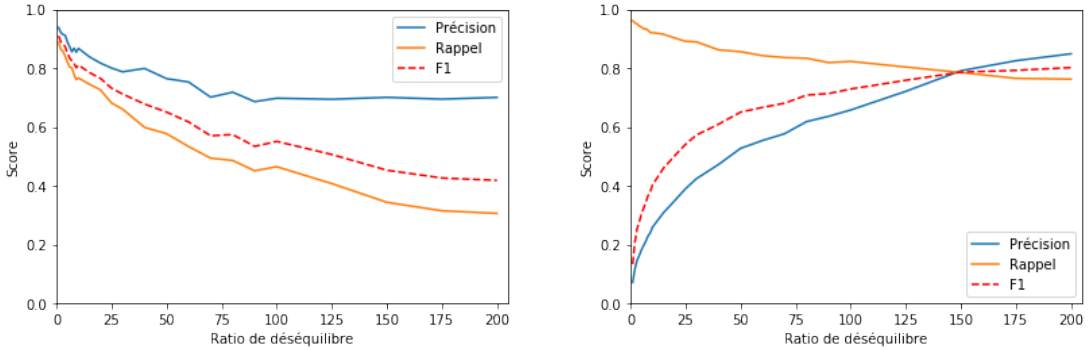
(b) *DRE, descripteurs BL,L,PL*

FIGURE A.18. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (BL = *baseline*, L = Levenshtein, PL = plongements lexicaux).



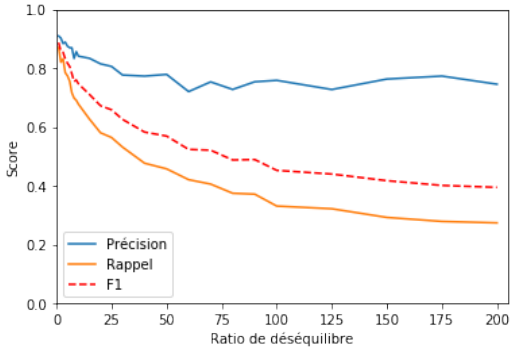
(a) *DDE, descripteurs BL,S,N* (b) *DRE, descripteurs BL,S,N*

FIGURE A.19. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (BL = *baseline*, S = mesures de similarité, N = *ngrams*).

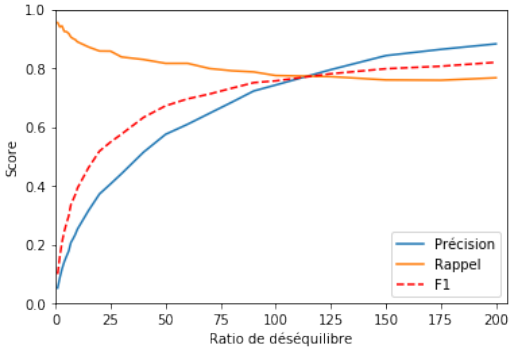


(a) *DDE, descripteurs BL,S,PL* (b) *DRE, descripteurs BL,S,PL*

FIGURE A.20. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (BL = *baseline*, S = mesures de similarité, PL = plongements lexicaux).

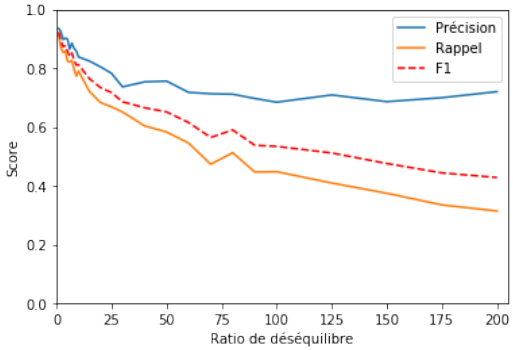


(a) *DDE, descripteurs BL,N,PL*

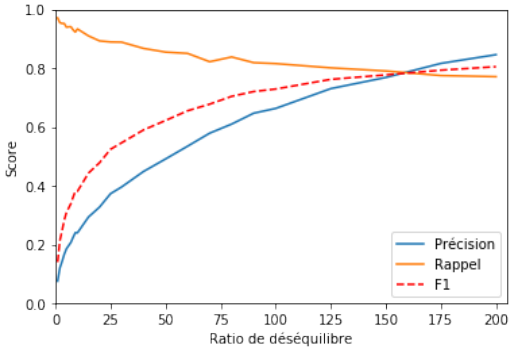


(b) *DRE, descripteurs BL,N,PL*

FIGURE A.21. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (BL = *baseline*, N = *ngrams*, PL = plongements lexicaux).

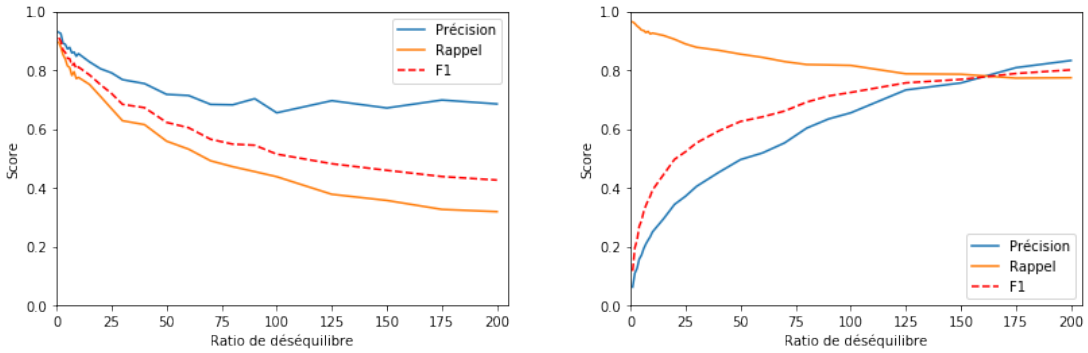


(a) *DDE, descripteurs L,S,N*



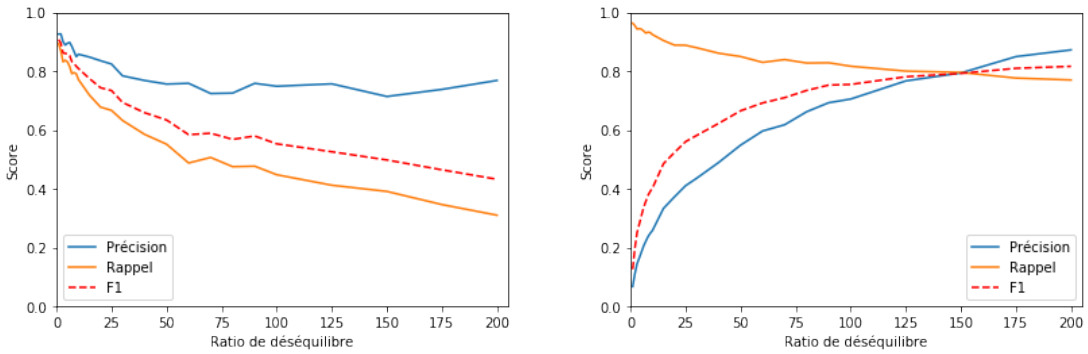
(b) *DRE, descripteurs L,S,N*

FIGURE A.22. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (L = Levenshtein, S = mesures de similarité, N = *ngrams*).



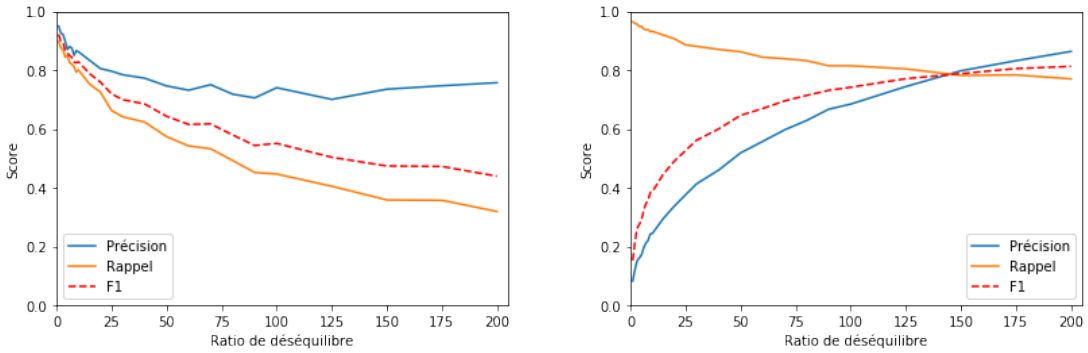
(a) *DDE, descripteurs L,S,PL* (b) *DRE, descripteurs L,S,PL*

FIGURE A.23. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (L = Levenshtein, S = mesures de similarité, PL = plongements lexicaux).



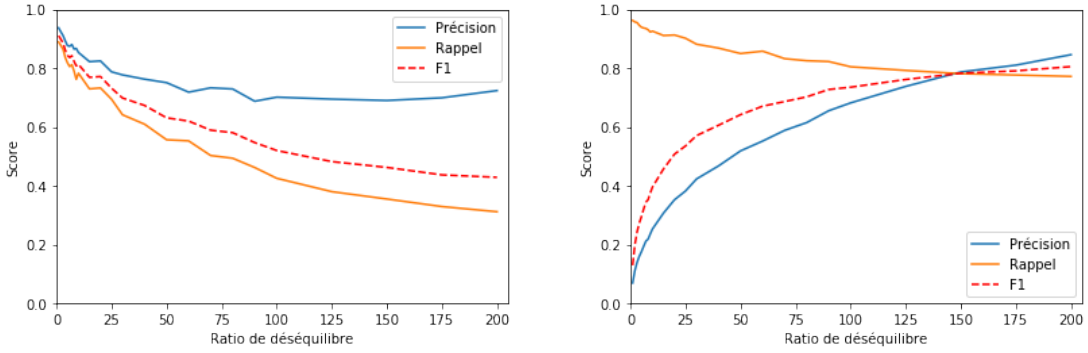
(a) *DDE, descripteurs L,N,PL* (b) *DRE, descripteurs L,N,PL*

FIGURE A.24. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (L = Levenshtein, N = *ngrams*, PL = plongements lexicaux).



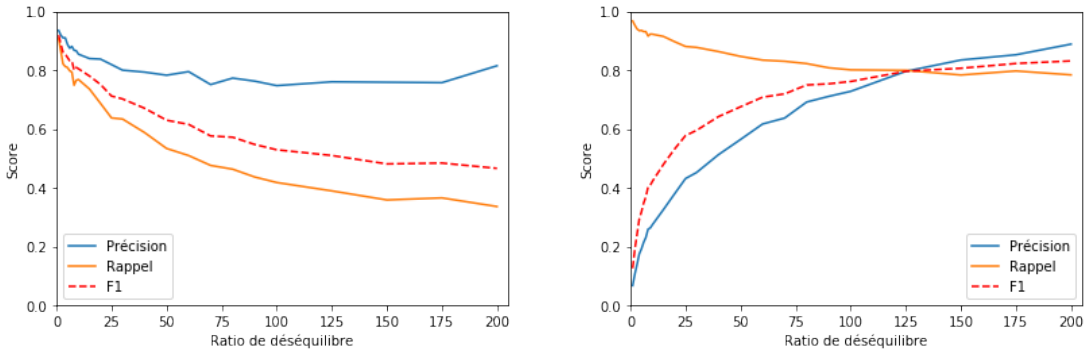
(a) *DDE, descripteurs BL,L,S,N* (b) *DRE, descripteurs BL,L,S,N*

FIGURE A.25. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (BL = *baseline*, L = Levenshtein, S = mesures de similarité, N = *ngrams*).



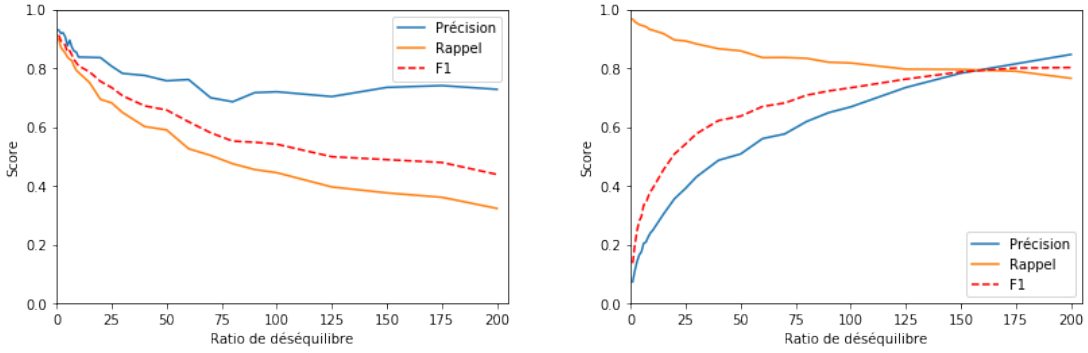
(a) *DDE, descripteurs BL,L,S,PL* (b) *DRE, descripteurs BL,L,S,PL*

FIGURE A.26. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (BL = *baseline*, L = Levenshtein, S = mesures de similarité, PL = plongements lexicaux).



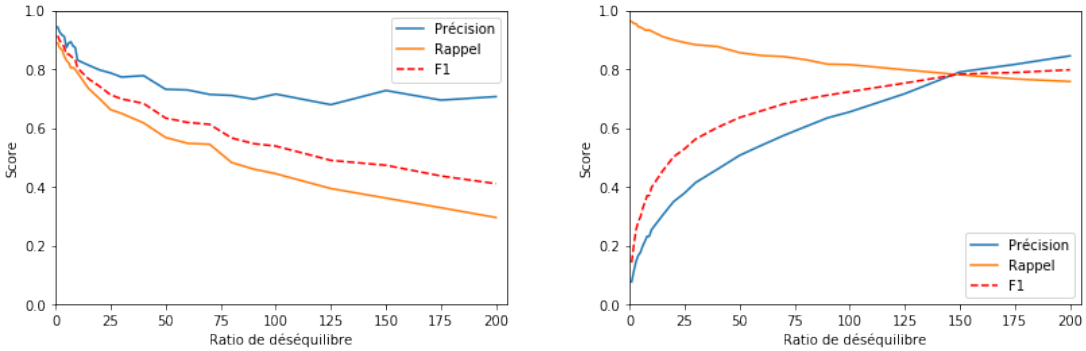
(a) *DDE, descripteurs BL,L,N,PL* (b) *DRE, descripteurs BL,L,N,PL*

FIGURE A.27. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (BL = *baseline*, L = Levenshtein, N = *ngrams*, PL = plongements lexicaux).



(a) *DDE, descripteurs BL,S,N,PL* (b) *DRE, descripteurs BL,S,N,PL*

FIGURE A.28. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (BL = *baseline*, S = mesures de similarité, N = *ngrams*, PL = plongements lexicaux).



(a) *DDE, descripteurs L,S,N,PL* (b) *DRE, descripteurs L,S,N,PL*

FIGURE A.29. – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*) sur les données déséquilibrées, par ensembles de descripteurs (L = Levenshtein, S = mesures de similarité, N = *ngrams*, PL = plongements lexicaux).

Simplification automatique de textes spécialisés et techniques

Résumé

La simplification automatique de textes est un domaine du traitement automatique des langues (TAL) qui vise à traiter des textes difficiles à lire pour un public donné de façon à les rendre plus accessibles. Notre objectif consiste à simplifier automatiquement les textes médicaux et de santé. Nous présentons l'ensemble de notre travail sur cette question, qui va de la collecte et analyse de corpus jusqu'aux expériences en simplification automatique.

Nous commençons par la collecte d'un corpus comparable de textes médicaux. Ce corpus est constitué de couples de documents qui traitent du même sujet : l'un s'adressant à un public spécialiste et l'autre à un public néophyte. Le corpus contient trois types de textes : des informations sur les médicaments, des revues systématiques de littérature médicale et des articles encyclopédiques. Une fois les documents collectés, nous annotons un sous-ensemble de ces documents et analysons les transformations linguistiques qui y sont mises en œuvre lors de la simplification.

À partir du corpus comparable, nous mettons en place une méthode pour en extraire un corpus parallèle, c'est-à-dire un corpus comprenant des couples de phrases qui ont le même sens mais diffèrent par leur degré de difficulté. Ce type de corpus représente le matériau principal pour les méthodes de simplification automatique. Notre méthode d'extraction de phrases parallèles comporte deux étapes : (1) le préfiltrage de paires de phrases candidates à l'alignement selon des heuristiques syntaxiques et (2) la classification binaire permettant de distinguer les phrases en relation de simplification. Nous évaluons différents classifieurs ainsi que l'influence du déséquilibre des données sur les performances. Afin de valoriser ce corpus parallèle, nous créons également un corpus de paires de phrases annotées selon leur similarité sémantique, avec des scores allant de 0 (sémantique indépendante) à 5 (même sémantique). Les deux corpus sont disponibles pour la recherche.

Enfin, nous présentons une série d'expériences en simplification automatique de textes médicaux en français. Ainsi, nous mettons à l'œuvre une méthode neuronale issue de la traduction automatique. Nous utilisons plusieurs ressources : le corpus parallèle médical construit par nous, le corpus parallèle de langue générale automatiquement traduit par nous de l'anglais vers le français ainsi qu'un lexique qui apparie des termes médicaux avec des termes ou paraphrases accessibles au grand public. Nous décrivons le protocole expérimental et menons une évaluation en deux volets, quantitatif et qualitatif. Les résultats sont comparables à l'état de l'art de la simplification en langue générale et montrent que les simplifications produites peuvent être exploitées dans le cadre d'une tâche de simplification assistée par ordinateur.

Automatic text simplification of specialized and technical texts

Abstract

Automatic text simplification is a subdomain of natural language processing (NLP). It aims at processing texts that are difficult to read for a given audience in order to make them more accessible. Our goal consists in automatically simplifying medical texts. We present our whole work on that question, that goes from data collection and analysis to automatic simplification experiments.

We begin with the process of collecting a comparable corpus of biomedical texts. The corpus is made of document pairs that deal with the same subject : one is written for a specialist audience and the other is written for non specialists. The corpus contains three types of texts : drug information, medical literature reviews and encyclopedia articles. Once the documents are collected, we annotate a subset of the corpus and analyze the linguistic transformations that occur during simplification.

From the comparable corpus, we build a method to extract a parallel corpus, a corpus that contains sentence pairs where the sentences have the same meaning but differ by their degree of difficulty. This type of corpus represents the basic material for automatic simplification methods. Our parallel sentences extraction method is made of two steps : (1) prefiltering the pairs that are candidate for alignment using syntactic heuristics and (2) using a binary classifier to distinguish sentences that have the same meaning. We evaluate various classifiers as well as the impact of the data imbalance on the results. In order to promote the parallel corpus, also create a corpus of sentence pairs that are annotated according to their degree of semantic similarity, with scores ranging from 0 (no similarity) to 5 (same meaning). Both corpora are available for research.

Finally, we present a series of experiments for the automatic simplification of biomedical french texts. Indeed, we use a neural method that comes from automatic translation. We use several resources : the parallel medical corpus that we built, the parallel general language corpus that we automatically translated from English to French and a lexicon that matches medical terms with terms or paraphrases that are more accessible. We describe the experimental protocol and evaluate the results in two manners, quantitatively and qualitatively. The results are similar to the state of the art in general language simplification and show that the resulting simplifications can be exploited as part of a computer aided simplification task.