Approche lexicale de la simplification automatique de textes médicaux

Rémi Cardon
UMR 8163 – STL – Savoirs Textes Langage, F-59000 Lille, France
remi.cardon@etu.univ-lille3.fr

RÉSUMÉ

Notre travail traite de la simplification automatique de textes. Ce type d'application vise à rendre des contenus difficiles à comprendre plus lisibles. À partir de trois corpus comparables du domaine médical, d'un lexique existant et d'une terminologie du domaine, nous procédons à des analyses et à des modifications en vue de la simplification lexicale de textes médicaux. L'alignement manuel des phrases provenant de ces corpus comparables fournit des données de référence et permet d'analyser les procédés de simplification mis en place. La substitution lexicale avec la ressource existante permet d'effectuer de premiers tests de simplification lexicale et indique que des ressources plus spécifiques sont nécessaires pour traiter les textes médicaux. L'évaluation des substitutions est effectuée avec trois critères : grammaticalité, simplification et sémantique. Elle indique que la grammaticalité est plutôt bien sauvegardée, alors que la sémantique et la simplicité sont plus difficiles à gérer lors des substitutions avec ce type de méthodes.

ABSTRACT

Lexical approach for the automatic simplification of medical texts

Our work addresses the automatic text simplification. This kind of application aims at improving the readability of texts that are difficult to read. Using three different corpora – which contain biomedical texts – an existing lexicon and a domain terminology, we perform analysis and modification of texts in order to achieve their lexical simplification. Manual alignment of sentences from comparable corpora provides reference data and permits to analyze the simplification procedures involved. Lexical substitution using existing resources permits to perform first tests of lexical simplification and indicates that specific resources are necessary when working with medical contents. The evaluation of substitutions is performed through three criteria : grammaticality, simplicity and semantics. It indicates that grammaticality is rather well preserved, while semantics and simplicity are more difficult to handle during the substitutions with this kind of methods.

MOTS-CLÉS : Simplification automatique de textes, analyse lexicale, domaine médical, simplification lexicale, substitution lexicale.

KEYWORDS: Automatic text simplification, lexical analysis, medical area, lexical simplification, lexical substitution.

1 Introduction

La simplification automatique de textes est un domaine du TAL, dans lequel il s'agit d'appliquer des transformations sur les phrases d'un texte pour les rendre plus lisibles, tout en conservant leur sens

intact. Cela est pratiqué aussi bien à destination des humains que pour faciliter les tâches nécessitant l'analyse automatique de textes (Chandrasekar *et al.*, 1996). La simplification, dont l'objectif est de faciliter des traitements d'analyse automatique, peut faire partie de différentes applications. Ainsi, la première application à l'avoir exploitée cherchait à simplifier les structures de phrases avant de procéder à leur analyse syntaxique automatique (Chandrasekar *et al.*, 1996). Dans d'autres contextes, la simplification peut être utilisée pour adapter certains genres 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). Pour la simplification à destination des humains, ces méthodes sont explorées pour différents objectifs et différents publics. Notons par exemple que la simplification est effectuée pour des personnes avec une faible compétence de lecture (Williams & Reiter, 2005), pour des personnes sourdes qui montrent des difficultés de lecture et d'écriture (Inui *et al.*, 2003), pour des lecteurs dyslexiques (Rello *et al.*, 2013) ou encore pour des personnes autistes (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). En effet, des études ont montré qu'une meilleure compréhension des informations de santé par les patients et leurs familles mène à une meilleure adhésion au traitement et à un processus de soins plus réussi (Davis & Wolf, 2004; Berkman *et al.*, 2011).

L'objectif de notre travail consiste à contribuer au domaine de la simplification de textes de spécialité, sur l'exemple de textes médicaux. D'une part, nous proposons de constituer des corpus comparables différenciés par leur spécialisation, d'effectuer un alignement de phrases à partir de ces corpus afin de faire une analyse des procédés de simplification mis en place. D'autre part, nous proposons d'effectuer de premiers tests de simplification lexicale en utilisant la substitution. Si la majorité de travaux de simplification traitent les données en langue anglaise, nous travaillons avec les données en français.

Dans la suite de ce travail, nous présentons d'abord l'état de l'art (section 2), ensuite les données sur lesquelles nous travaillons (section 3). La méthode est présentée dans la section 4 et les résultats dans la section 5. Nous proposons une conclusion et les perspectives de ce travail dans la section 6.

2 État de l'art

Les travaux en simplification automatique se positionnent essentiellement à deux niveaux : lexical et syntaxique. La *simplification lexicale* opère au niveau des unités lexicales. Nous allons illustrer la simplification lexicale avec la tâche de simplification proposée lors de la compétition *SemEval 2012* [1] pour la langue anglaise. Pour un texte court et un mot cible, plusieurs substitutions possibles et satisfaisant le contexte ont été proposées par les organisateurs. L'objectif consistait à trier ces substitutions selon leur degré de simplicité (Specia *et al.*, 2012) et donc de les positionner les unes par rapport aux autres, en fonction de leur difficulté. Par exemple, pour la phrase *Hitler committed terrible atrocities during the second World War*, le mot à substituer est *atrocities*. Les candidats synonymes proposés par les organisateurs sont *abomination, cruelty, enormity, violation*. Le choix de référence est *cruelty*, ce qui doit produire en sortie : *Hitler committed terrible cruelties during the second World War.* De manière générale, lors de la simplification lexicale, plusieurs étapes peuvent être distinguées :

1. L'identification de mots ou termes qui peuvent poser des difficultés de compréhension. Cette étape est le plus souvent accomplie à l'aide de ressources lexicales auxquelles sont associées

1. http://www.cs.york.ac.uk/semeval-2012/

des mesures de complexité des mots, même si ces mesures n'ont pas reçu le consensus de la communauté de recherche (Saggion, 2017; Shardlow, 2014). Comme indiqué plus bas, des mesures classiques et computationnelles, et donc plus récentes, sont distinguées ;

2. Le remplacement de ces unités par un équivalent jugé plus facile de compréhension. Cette étape repose sur la disponibilité d'un dictionnaire d'expressions synonymiques (Shardlow, 2014) ou même hyperonymiques ;
3. Lorsque plusieurs équivalents 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). C'était typiquement la tâche proposée lors de la compétition *SemEval 2012*. 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) ;
4. Il faut également s'assurer que les candidats à substitution sont acceptables dans le contexte de chaque phrase traitée. Dans *SemEval 2012*, cette condition était assurée par les organisateurs.

La *simplification syntaxique* opère au niveau de la syntaxe. Son objectif est de réorganiser la structure syntaxique des phrases. Quelques exemples d'opérations de réorganisation de cette structure sont : le découpage de phrases complexes en plusieurs phrases plus simples, l'ajout ou la suppression de propositions, la modification de temps verbaux (Brouwers *et al.*, 2014; Gasperin *et al.*, 2009; Seretan, 2012). Pour un exemple, la phrase *Elle compte, selon les autorités du pays, en 2009, 5 878 609 habitants pour l'agglomération, et 3 796 677 habitants pour la ville, soit 20 % de la population totale du pays* devient, après l'application d'une règle de suppression de propositions subordonnées et d'incises : *Elle compte, selon les autorités du pays, en 2009, 5 878 609 habitants pour l'agglomération, et 3 796 677 habitants pour la ville* (Brouwers *et al.*, 2014).

Pour ces deux types de simplification, des approches à base de règles et de probabilités ont été développées. Les approches à base de règles reposent sur l'expertise des concepteurs et leur connaissance des procédés de simplification. Notons que cette expertise peut également profiter d'un corpus de simplification déjà disponible. Les approches statistiques nécessitent de disposer de corpus de textes comparables, ou même parallèles et alignés, qui contiennent typiquement des textes complexes et leurs versions simplifiées (Zhu *et al.*, 2010; Specia, 2010; Woodsend & Lapata, 2011). Un des exemples typiques de corpus comparables, qui sont largement utilisés dans ce type de travaux, sont Wikipedia [2] et Simple Wikipedia [3] en langue anglaise. Si le premier propose des articles à destination de la population en général, Simple Wikipedia vise des populations spécifiques (comme par exemple les enfants, les apprenants d'anglais, les adultes en difficulté de lecture, etc.).

Une des tâches importantes de la simplification automatique consiste à pouvoir mesurer la lisibilité des unités linguistiques. Différents niveaux de mesure sont utilisés, que ce soit au niveau des termes ou des textes. Deux grands types de mesures de lisibilité peuvent être distingués : classiques et computationnelles (François, 2011). Les mesures classiques reposent sur le calcul de la complexité de surface des mots (nombre de syllabes) et de phrases pour évaluer leur lisibilité. Par exemple, si les mots et les phrases d'un texte sont longs, il est considéré être difficile à lire (Flesch, 1948; Gunning, 1973; Dubay, 2004). Les mesures computationnelles fournissent la possibilité d'associer les

2. https://en.wikipedia.org/wiki/Main_Page
3. https://simple.wikipedia.org/wiki/Simple_English_Wikipedia

unités, dont on souhaite mesurer la lisibilité (mots, phrases, textes...), à une variété de descripteurs : combinaisons de mesures classiques avec des informations terminologiques (Kokkinakis & Gronostaj, 2006), utilisation de *n-grammes* de caractères (Poprat *et al.*, 2006), descripteurs discursifs (Goeuriot *et al.*, 2007), descripteurs morphologiques (Chmielik & Grabar, 2011), etc. En général, ces mesures sont calculées de manière supervisée par rapport à une référence et fournissent des résultats fiables.

Par rapport aux travaux existants, nous nous positionnons au niveau de la simplification lexicale. Nous proposons d'effectuer la simplification en effectuant des substitutions lexicales grâce à l'exploitation d'un lexique existant. Notre tâche est assez proche de celle proposée lors de *SemEval 2012*.

3 Présentation des données

Les données exploitées sont de deux types : les corpus comparables provenant du domaine médical, et les ressources utilisées pour l'analyse et la substitution lexicale.

3.1 Corpus

3.1.1 Cochrane

Cochrane [4] est un organisme qui a pour objectif la diffusion de l'information médicale (Sackett *et al.*, 1996). Les textes publiés par Cochrane sont des synthèses de la littérature médicale sur une question spécifique (diagnostic, traitement). Ces synthèses sont créées à destination des professionnels de santé. Plus récemment, elles sont simplifiées par les collaborateurs de Cochrane pour les rendre accessibles au grand public également. De même, les synthèses écrites en anglais sont traduites en d'autres langues, y compris le français. Le corpus que nous utilisons est composé de 3 815 synthèses en français à destination des médecins et, pour chacune d'elle, sa version simplifiée (voir la figure 1).

Version technique	*Version simplifiée*
L'hématome aigu de l'oreille est une affection qui se caractérise par la formation d'une collection sanguine sous le périchondre du pavillon. Il est souvent provoqué par un traumatisme contondant. En l'absence de traitement, il finit par entraîner une difformité couramment appelée oreille en chou-fleur ou oreille du boxeur.	L'hématome aigu de l'oreille est une affection qui se caractérise par la formation d'une collection sanguine dans le pavillon (oreille externe), souvent à la suite d'un traumatisme contondant. S'il n'est pas traité, il entraîne une difformité appelée oreille en chou-fleur ou oreille du boxeur.

FIGURE 1 – Exemple de textes comparables du corpus Cochrane.

3.1.2 Encyclopédie

Wikipedia [5] est une encyclopédie collaborative en ligne. Elle propose les informations à destination d'un large public et aborde un grand nombre de sujets. Vikidia [6] est également une encyclopédie

4. https://www.cochranelibrary.com
5. https://fr.wikipedia.org
6. https://fr.vikidia.org/

collaborative en ligne, sa spécificité est qu'elle est destinée aux enfants de 8 à 13 ans. Nous exploitons ces deux sources pour constituer un deuxième corpus de travail. Il est composé de 575 articles relatifs au portail de la Médecine de Wikipedia et des articles équivalents de Vikidia (voir la figure 2).

Version technique	*Version simplifiée*
La luette ou uvule est un appendice conique situé au fond de la cavité buccale et proche des tonsilles palatines. Le mot uvule vient du latin uva, qui signifie "grain de raisin". La luette est un organe de 10 à 15 millimètres de long, de forme tubulaire quand il est détendu, qui pend à la partie moyenne du bord inférieur du voile du palais. Elle est constituée d'un tissu membraneux et musculaire.	La luette ou uvule est un appendice conique située au fond de la bouche. C'est un organe fait de tissus membraneux et musculaires, d'environ 10 à 15mm de long, qui pend à la partie moyenne du voile du palais.

FIGURE 2 – Exemple de textes comparables du corpus Wikipedia/Vikidia.

3.1.3 Médicaments

Le troisième corpus contient les informations sur les médicaments issues de la base de données [7] du ministère de la santé. On peut accéder, pour un médicament donné, au résumé des caractéristiques du produit (RCP), créé à destination des professionnels, et à la notice, créée à destination du grand public. Ces dernières peuvent aussi être trouvées dans les boîtes de médicaments. Ce corpus contient 11 800 RCP techniques et leurs notices grand public (voir la figure 3).

Version technique	*Version simplifiée*
- hypersensibilité à l'huile de paraffine	- si vous avez une allergie à l'huile de paraffine
- colopathie obstructive, compte tenu de l'effet laxatif du médicament	- si vous êtes atteint de colopathie obstructive, compte tenu de l'effet laxatif du médicament
- syndrome douloureux abdominal de cause indéterminée et inflammatoire (rectocolite ulcéreuse, maladie de Crohn)	- si vous êtes atteint de syndrome douloureux abdominal de cause indéterminée et inflammatoire (rectocolite ulcéreuse, maladie de Crohn, ...)
- ne pas utiliser chez les personnes présentant des difficultés de déglutition en raison du risque d'inhalation bronchique et de pneumopathie lipoïde	- ne pas utiliser chez les personnes présentant des difficultés pour avaler en raison du risque d'inhalation de la paraffine liquide qui entraîne une pneumopathie lipoïde

FIGURE 3 – Exemple de textes comparables du corpus Médicament.

3.1.4 Bilan des corpus

Le tableau 1 fait le bilan des trois corpus et indique, pour chaque corpus : sa taille en nombre de documents, d'occurrences de mots et de lemmes uniques. Il s'agit de corpus comparables : les articles

7. https://base-donnees-publique.medicaments.gouv.fr/

concernent les mêmes sujets mais relèvent de différents discours. Comme le montrent les extraits, ils proposent rarement une réécriture de la version technique (originale) en langage simplifié. En effet, le plus souvent, la création de la version simplifiée semble être indépendante de la version technique. Nous pouvons voir que le corpus *Médicaments* est le plus gros des trois corpus étudiés, tandis que le corpus *Encyclopédie* est le plus petit. Par ailleurs, les versions techniques sont toujours plus volumineuses que les versions simplifiées.

Corpus	*nb doc.*	*nb occ.*	*nb lemmes uniques*
Cochrane technique	3 815	2 804 336	11 558
Cochrane simplifié	3 815	1 491 243	7 567
Médicaments technique	11 800	51 705 111	43 515
Médicaments simplifié	11 800	33 116 119	25 725
Wikipedia	575	2 186 891	19 287
Vikidia	575	183 051	3 117

TABLE 1 – Taille des corpus comparables.

3.2 Ressources

Nous utilisons deux ressources : une terminologie spécialisée, Snomed International, et un lexique généraliste issu du Wiktionary.

3.2.1 Terminologie médicale Snomed

Nous exploitons la terminologie médicale Snomed International (Côté, 1996) telle que diffusée par ASIP santé. [8] La vocation de cette terminologie est de décrire le domaine médical. La terminologie contient 151 104 termes médicaux structurés en onze axes sémantiques (maladies et anomalies, actes médicaux, produits chimiques, organismes vivants, anatomie). Selon notre hypothèse, le contenu de cette terminologie permet d'estimer la couverture en termes et mots médicaux d'un texte donné.

3.2.2 Lexique issu du Wiktionary

Nous utilisons un lexique obtenu à partir des articles du Wiktionary [9], dans sa version GLAWI (Sajous & Hathout, 2015). Nous retenons les entrées dont au moins une définition est associée à l'une des catégories suivantes : *anat*, *chirurgie*, *génétique*, *maladie(s)*, *médecine*, *médicaments*, *microbiologie*, *neurologie*, *pathologie*, *pédologie*, *pharmacologie*, *physiologie*, *squelette*, *virologie*. Cela fournit un lexique avec 8 012 entrées uniques qui peuvent être liées au domaine médical. Les catégories les plus importantes sont *médecine* avec 4 967 entrées et *anat* (pour *anatomie*) avec 1 925 entrées. Les autres catégories contiennent entre quelques dizaines et quelques centaines d'entrées.

Cette ressource fournit des séries de synonymes et des hyperonymes pour certains termes. Ainsi, 25 % des entrées de GLAWI ont au moins un synonyme, avec une moyenne de 2,42 synonymes par entrée.

8. http://esante.gouv.fr/services/referentiels/referentiels-d-interoperabilite/snomed-35vf

9. https://fr.wiktionary.org/

Seules 452 entrées ont des hyperonymes : 318 avec un hyperonyme et 134 avec 2 à 6 hyperonymes. Les séries de synonymes et d'hyperonymes comportent le plus souvent des mots de la même partie du discours. Cette ressource contient essentiellement les entrées simples, mais nous avons également 1 157 entrées polylexicales. Enfin, cette ressource offre le paradigme flexionnel (le lemme et ses formes fléchies) des entrées.

4 Méthode

4.1 Pré-traitements

Les corpus sont pré-traités : l'étiquetage morpho-syntaxique et la lemmatisation sont effectués avec le TreeTagger (Schmid, 1994). Cela permet de normaliser le contenu des corpus grâce à la lemmatisation. Les termes de la terminologie Snomed International sont projetés sur les textes, ce qui permet d'effectuer une analyse lexicale de ces corpus.

4.2 Alignement de phrases

Nos corpus sont des corpus comparables. Notre objectif est de constituer une base de phrases alignées, qu'elles soient parallèles ou comparables, à partir desquelles nous pourrions faire des observations sur les manières dont la simplification peut être effectuée.

Une sélection aléatoire d'articles a fourni : 2*13 documents du corpus *Cochrane*, 2*12 documents du corpus *Médicaments* et 2*14 documents du corpus *Encyclopédie*. Ces articles sont utilisés pour effectuer l'alignement manuel de phrases provenant des corpus techniques et simplifiés. Deux annotateurs ont effectué l'alignement de manière indépendante. Des séances de consensus ont permis ensuite de résoudre les désaccords. Les critères qui guidaient l'alignement sont les suivants :

1. Les deux phrases doivent avoir le même sens ou sinon des sens proches, comme dans ces phrases du corpus *Cochrane* :
   — *les sondes gastriques sont couramment utilisées pour administrer des médicaments ou une alimentation entérale aux personnes ne pouvant plus avaler*
   — *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*
2. Le sens d'une phrase peut se retrouver intégralement dans le sens de l'autre phrase. Dans l'idéal, il devrait s'agir de l'inclusion sémantique, comme dans cet exemple, où la phrase technique indique le nombre de participants et la mesure d'évaluation en plus :
   — *peu de données (43 participants) étaient disponibles concernant la détection d'un mauvais placement (la spécificité) en raison de la faible incidence des mauvais placements*
   — *cependant, peu de données étaient disponibles concernant les sondes placées incorrectement et les complications possibles d'une sonde mal placée*
3. Les cas d'intersection sémantique, où chaque phrase apporte des informations spécifiques, seraient à proscrire. L'exemple du corpus *Cochrane* qui suit illustre ce cas :
   — *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*

Les cas d'intersection sémantique sont plus difficiles à généraliser et à reproduire.

Dans ces exemples, nous voyons que le passage d'une phrase technique vers sa version simplifiée requiert des modifications syntaxiques et lexicales. Dans notre travail, nous nous concentrons sur la simplification lexicale effectuée au moyen de substitutions lexicales.

4.3 Substitution lexicale

Nous proposons d'aborder la simplification au niveau lexical grâce aux substitutions de mots par leurs équivalents supposés être plus simples et faciles à comprendre. Notre approche est fondée sur des règles. Elle suit les étapes suivantes :

— Les phrases des corpus techniques sont exploitées pour effectuer les tests de substitution ;
— L'ensemble médical du lexique Wiktionary est exploité pour fournir les candidats à substitution. Ce lexique propose en effet des ensembles de synonymes et d'hyperonymes ;
— Les entrées de ce lexique sont filtrées. Comme nous l'avons vu, le corpus *Vikidia* propose le contenu le moins spécialisé par rapport au reste des corpus. Nous projetons donc les entrées du lexique et leurs synonymes et hyperonymes sur le corpus *Vikidia*. Si une entrée, ses synonymes ou hyperonymes, sont reconnues dans ce corpus, nous supposons qu'il s'agit des entrées plus faciles à comprendre, et les retenons pour effectuer les substitutions lexicales ;
— Si une phrase contient une entrée du lexique qui possède des synonymes ou hyperonymes, si cette entrée ne se trouve pas dans le corpus *Vikidia*, et si un de ses synonymes ou hyperonymes se trouve dans le corpus *Vikidia*, ce synonyme ou hyperonyme est utilisé pour la substitution.

Il s'agit d'une méthode souvent exploitée dans les travaux de l'état de l'art, qui visent à effectuer la simplification lexicale des textes (Biran *et al.*, 2011; Wubben *et al.*, 2012; Horn *et al.*, 2014; Glavas & Stajner, 2015; Abualhaija *et al.*, 2017).

4.4 Évaluation

Les résultats de la substitution lexicale sont évalués avec plusieurs critères exploités dans les travaux existants en simplification lexicale (Biran *et al.*, 2011; Wubben *et al.*, 2012) :

— *Grammaticalité*. Le jugement sur la grammaticalité doit répondre à la question de savoir si la phrase reste grammaticale après les modifications effectuées. Pour assurer le respect de ce critère, lors de la simplification lexicale par exemple, la plupart des travaux effectuent la substitution avec des mots de la même catégorie syntaxique que le mot substitué (Biran *et al.*, 2011; Horn *et al.*, 2014; Glavas & Stajner, 2015; Abualhaija *et al.*, 2017) ;
— *Sémantique*. Le jugement sur la sémantique doit répondre à la question de savoir si la transformation effectuée préserve la sémantique originale de la phrase. En effet, quelle que soit la simplification effectuée, la sémantique des textes doit rester préservée ;
— *Simplicité*. Le jugement sur la simplicité doit répondre à la question de savoir si la simplification effectuée sur une phrase la rend plus simple à comprendre.

Ces critères sont évalués manuellement par l'auteur.

5 Résultats et leur discussion

Nous présentons les résultats relatifs à l'analyse lexicale des corpus (section 5.1), à l'alignement de phrases (5.2) et à la substitution lexicale en vue de simplification (section 5.3).

5.1 Analyse lexicale

L'analyse lexicale, basée sur la terminologie Snomed International, permet de calculer : (1) le nombre de ses mots et termes qui apparaissent dans chacun des corpus, et (2) le ratio d'occurrences des termes de la Snomed entre les textes en version technique et simplifiée. Notons que la lemmatisation avec TreeTagger peut empêcher la reconnaissance d'expressions polylexicales présentes dans la terminologie, comme {*bilirubine totale* ; *bilirubine total*} ou {*amnésie passagère* ; *amnésie passager*}. L'objectif de la projection ici est d'obtenir une estimation du degré de spécialisation de chaque sous-corpus, et plus précisément de pouvoir comparer ces estimations entre elles. Nous n'appliquons pas de traitement spécifique pour le repérage des expressions polylexicales.

Corpus	*Termes*
Cochrane simplifié	2 316
Cochrane technique	2 505
Médicaments simplifié	2 700
Médicaments technique	3 332
Encyclopédie simplifié	1 635
Encyclopédie technique	3 999

TABLE 2 – Nombre de termes uniques de la Snomed International dans les corpus.

Ratio	*Valeur*
Cochrane : simplifié / technique	0.60
Cochrane : technique / simplifié	1.66
Médicaments : simplifié / technique	0.62
Médicaments : technique / simplifié	1.62
Encyclopédie : simplifié / technique	0.10
Encyclopédie : technique / simplifié	9.67

TABLE 3 – Ratio des occurrences de termes de la Snomed International dans les corpus.

Les tableaux 2 et 3 présentent les résultats. Selon le tableau 2, nous trouvons plus de termes de la Snomed dans les versions techniques. Par ailleurs, le corpus Vikidia est le plus pauvre en termes spécialisés. Il s'agit certainement du corpus dont le niveau de lisibilité est le plus élevé. Ces observations sont corroborées par les indications du tableau 3 qui indique les ratios de termes entre les corpus techniques et spécialisés. Ces ratios sont comparables pour les corpus *Cochrane* et *Médicaments*. En revanche, nous observons une différence beaucoup plus importante entre les versions technique et simplifiée du corpus encyclopédique, où les versions simplifiées d'articles contiennent beaucoup moins de termes spécialisés.

5.2 Alignement de phrases

Corpus	*nb doc.*	*nb phrases total*	*nb phrases alignées*	*ratio*
Cochrane	26	653	240	36,75%
Médicaments	24	7101	258	3,63%
Encyclopédie	28	2651	163	6,15%

TABLE 4 – Nombre de phrases alignées pour chaque corpus.

Le tableau 4 indique les résultats consensuels de l'alignement manuel des phrases. Nous pouvons observer que les phrases alignées, qui font la correspondance entre le contenu technique et simplifié, sont relativement plus rares pour les corpus *Médicaments* et *Encylopédie*, alors que le corpus *Cochrane* en offre plus par rapport à sa taille. Les raisons de cet état de chose peuvent être les suivantes :

— La ligne directrice de rédaction des versions simplifiées des résumés de la fondation Cochrane affiche explicitement une volonté de simplifier le contenu de ses résumés d'origine pour le grand public. Les rédacteurs et traducteurs prennent donc comme point de départ les résumés originaux et techniques et les simplifient au fur et à mesure de l'avancement ;

— Pour les deux autres corpus, les principes ne sont pas aussi stricts. Ainsi, l'objectif de Vikidia est de traiter des sujets présents dans Wikipedia mais pour un public d'enfants. La création d'articles de Vikidia est rarement basée sur les articles de Wikipedia : le plus souvent, il s'agit d'une écriture indépendante. Quant au corpus *Médicaments*, les mêmes médicaments sont décrits et spécifiés dans les versions technique et simplifiée. Cependant, certaines informations sur les médicaments sont propres aux RCP (composition plus détaillée, action sur l'organisme, molécules, détail sur les effets indésirables...), alors que d'autres informations sont propres aux notices destinées au grand public (précautions d'emploi, mises en garde...).

Une analyse de ces phrases alignées nous indique aussi que les procédés de simplification (lexicale, syntaxique et stylistique) ne sont pas les mêmes selon les corpus :

— *Simplification lexicale.* Dans Vikidia, les notions complexes sont plutôt explicitées, alors que dans les corpus *Médicaments* et *Cochrane*, les notions complexes sont souvent suivies par leurs équivalents entre parenthèses :

  — *en revanche, les ultrasons associés à d'autres tests (par exemple, la visualisation de l'irrigation saline (injecter une solution saline à travers la sonde et l'observer à l'intérieur de l'estomac par ultrasons)) pourraient être utiles pour confirmer le placement des tubes utilisés pour le drainage gastrique*

  — *l'alimentation offerte au travers d'un tube placé par erreur dans la trachée (un conduit où passe l'air respiré) peut entraîner une pneumonie grave (une infection des poumons)*

  Quel que soit le corpus, les notions complexes peuvent aussi être remplacées par leurs équivalents plus simples. En voilà quelques exemples au format {*technique* ; *simplifié*} :

  {*alimentation* ; *nourriture*}, {*entérale* ; *directement dans le tractus gastro-intestinal*}, {*fournir* ; *être*}, {*dans des contextes* ; *lorsque*}, {*mauvais placement* ; *placé incorrectement*}, {*incidence* ; *complications possibles*}...

  Dans plusieurs cas, ces différents procédés de simplification lexicale (définitions, équivalents, substitutions) sont employés en même temps dans une même phrase ;

— *Simplification syntaxique.* Le fait le plus marquant de la simplification syntaxique concerne les énumérations et les exemples virgulés. Ainsi, une phrase coordonnée peut être segmentée en une liste avec des items. Cependant, il n'y a pas de règles sur ce qui est approprié à faire

pour effectuer la simplification car parfois les énumérations virgulées se trouvent dans les documents techniques et dans d'autres cas dans les documents simplifiés ;

— *Style.* Dans le corpus *Médicaments*, certains énoncés deviennent personnels et s'adressent directement à la personne grâce à l'emploi de pronoms personnels (*vous, votre, vos...*), comme dans les exemples de la figure 3.

5.3 Substitution lexicale

Pour l'évaluation de la substitution lexicale, nous avons sélectionné aléatoirement 10 documents dans chacun des trois corpus techniques pour y appliquer la méthode. Cela représente 7 892 phrases (2 456 pour le corpus *Médicaments*, 5 057 pour le corpus *Encyclopédie*, 379 pour le corpus *Cochrane*). La substitution lexicale, effectuée avec la ressource issue de Wiktionary, a permis de traiter 86 phrases. Cette faible couverture suggère que des ressources plus spécifiques sont nécessaires.

Critère	*% Méthode*	*Devlin*	*Biran*
Grammaticalité	70%	70.23%	77.91%
Simplicité	14.46%	46.43%	75.58%
Sémantique	18.51%	55.95%	46.43%

TABLE 5 – Évaluation manuelle des substitutions.

Dans le tableau 5, colonne *% Méthode*, nous indiquons les résultats d'évaluation des substitutions effectuées. Globalement, les substitutions fournissent des résultats qui restent grammaticaux : la ressource utilisée contient des séries de synonymes et d'hyperonymes qui appartiennent le plus souvent à la même catégorie grammaticale, comme {*absorption* ; *ingestion*} ou {*traiter* ; *soigner*}. Concernant la simplicité, les substitutions n'apportent pas toujours la simplification des phrases d'origine : un filtrage supplémentaire ou différent de la ressource est nécessaire. Finalement, la substitution peut aussi introduire des nuances sémantiques dans les phrases traitées. Nous comparons nos résultats avec deux travaux en substitution lexicales effectués en anglais (Devlin & Unthank, 2006; Biran *et al.*, 2011) : la ressource WordNet est exploitée pour traiter des textes de la langue générale. Nos résultats sont comparables quant à la grammaticalité, en revanche nous obtenons des résultats de simplicité et de sémantique plus faibles. Nous pensons que la raison principale de ces faibles résultats vient de la ressource utilisée, qui n'est pas adaptée à la simplification de textes médicaux techniques ou spécialisés. Des ressources plus spécifiques sont donc nécessaires.

Les figures 4 et 5 proposent quelques exemples de substitutions effectuées avec les ressources disponibles. Ainsi, la figure 4 propose des substitutions réussies, où la sémantique des phrases reste fidèle aux phrases d'origine, grâce aux synonymes comme {*absorption* ; *ingestion*}, {*traitement* ; *prescription*} ou {*traiter* ; *soigner*}. Alors que la figure 5 propose des substitutions non réussies, où la sémantique des phrases n'est pas sauvegardée. Par exemple, la sémantique change dans le cas des synonymes {*corps* ; *mort*}, alors que dans l'exemple avec les synonymes comme {*main* ; *pince*}, {*dents* ; *chicots*} ou {*tête* ; *citron*}, il s'agit de synonymes qui appartiennent à différents niveaux de la langue {*normé* ; *jargon*}. Même si cela ne modifie pas beaucoup la sémantique des phrases, la formulation devient plus familière, ce qui n'etait pas l'effet recherché.

Avant substitution	*Après substitution*
La nourriture n'a pas d'effet sur l'absorption d'anastrozole.	La nourriture n'a pas d'effet sur l'ingestion d'anastrozole.
Vous devez discuter avec votre médecin sur les risques et les options de traitement.	Vous devez discuter avec votre médecin sur les risques et les options de prescription.
Votre médecin peut vous prescrire un médicament visant à prévenir ou traiter cette perte osseuse.	Votre médecin peut vous prescrire un médicament visant à prévenir ou soigner cette perte osseuse.

FIGURE 4 – Exemples de substitutions réussies.

Avant substitution	*Après substitution*
Un abcès est une accumulation de pus sous la peau ou à l'intérieur du corps.	Un abcès est une accumulation de pus sous la peau ou à l'intérieur du mort.
Syndrome du canal carpien (fourmillement, douleur, sensation de froid, faiblesse dans certaines parties de la main).	Syndrome du canal carpien (fourmillement, douleur, sensation de froid, faiblesse dans certaines parties de la pince).

FIGURE 5 – Exemples de substitutions non réussies.

6 Conclusion et perspectives

Dans ce travail, nous avons proposé d'effectuer la simplification automatique de textes médicaux en français. Notre travail propose plusieurs contributions : (1) création de corpus comparables avec des textes médicaux techniques et simplifiés ; (2) alignement manuel de phrases ; (3) observations des procédés de simplification présents dans les corpus ; (4) premiers tests de substitution lexicale ; (5) évaluation des résultats avec trois critères de jugement (grammaticalité, simplification et sémantique).

Nous avons plusieurs perspectives à ce travail : (1) préparer et exploiter un lexique plus approprié pour la substitution lexicale dans les textes médicaux, comme ceux proposés dans les travaux existants (Grabar & Hamon, 2016), ce qui devrait permettre d'augmenter la couverture des substitutions ; (2) mieux gérer l'ambiguïté contextuelle des synonymes, ce qui devrait permettre d'augmenter l'acceptabilité sémantique des substitutions ; (3) augmenter le volume de phrases alignées, ce qui devrait permettre de tester d'autres approches pour la substitution, y compris les approches probabilistes ; (4) combiner différents types de modifications lexicales (substitutions, ajouts de paraphrases et de définitions) ; (5) combiner la simplification lexicale avec la simplification syntaxique pour fournir des résultats plus complets.

Remerciements

La présente publication s'inscrit dans le projet *CLEAR (Communication, Literacy, Education, Accessibility, Readability)* financé par l'ANR sous la référence ANR-17-CE19-0016-01.

Je remercie les relecteurs pour leurs remarques constructives. Je remercie également Natalia Grabar, pour son aide dans la réalisation des travaux décrits ici, ainsi que dans la rédaction de cette publication.

Références

ABUALHAIJA S., MILLER T., ECKLE-KOHLER J., GUREVYCH I. & ZIMMERMANN K.-H. (2017). Metaheuristic approaches to lexical substitution and simplification. In *EACL 2017*, p. 1–11.

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

BARBU E., MARTIN-VALDIVIA M., ALFONSO L. & LOPEZ U. (2013). Open book : a tool for helping ASD users' semantic comprehension. In *Proceedings of the 2nd workshop of natural language processing for improving textual accessibility NLP4ITA*, p. 11–19, Atlanta, United States.

BERKMAN N., SHERIDAN S., DONAHUE K., HALPERN D. & CROTTY K. (2011). Low health literacy and health outcomes : An updated systematic review. *Annals of Internal Medicine*, **155**(2), 97–107.

BIRAN O., BRODY S. & ELHADAD N. (2011). Putting it simply : a context-aware approach to lexical simplification. In ACL, Ed., *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers-Volume 2*, p. 496–501.

BRIN-HENRY F. (2014). Éducation thérapeutique du patient et orthophonie. In *Communiquer malgré l'aphasie*. S. Médical.

BROUWERS L., BERNHARD D., LIGOZAT A.-L. & FRANÇOIS T. (2014). Syntactic Sentence Simplification for French. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL 2014*, p. 47–56, Gothenburg, Sweden.

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, p. 1041–1044, Stroudsburg, PA, USA : Association for Computational Linguistics.

CHMIELIK J. & GRABAR N. (2011). *Détection de la spécialisation scientifique et technique des documents biomédicaux grâce aux informations morphologiques*, In *TAL*, volume 51(2), p. 151–179.

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

DAVIS T. & WOLF M. (2004). Health literacy : implications for family medicine. *Fam Med*, **36**, 595–598.

DE BELDER J. & MOENS M. (2010). Text simplification for children. In *Workshop on accessible search systems of SIGIR*, p. 1–8.

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, p. 225–226, New York, NY, USA : ACM.

DUBAY W. (2004). *The principles of readability*, In *Impact Information*.

FLESCH R. (1948). *A new readability yardstick*, In *Jounal of Applied Psychology*, volume 23, p. 221–233.

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*, p. 15–28, Paris, France.

FRANÇOIS T. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère*. PhD thesis, Université Catholique de Louvain, Louvain.

GASPERIN C., MAZIERO E., SPECIA L., PARDO T. & ALUISIO R. M. (2009). *Natural language processing for social inclusion : a text simplification architecture for different literacy levels*, In *SEMISH-XXXXVI*, p. 397–404.

GLAVAS G. & STAJNER S. (2015). Simplifying lexical simplification : Do we need simplified corpora ? In *ACL-COLING*, p. 63–68.

GOEURIOT L., GRABAR N. & DAILLE B. (2007). *Caractérisation des discours scientifique et vulgarisé en français, japonais et russe*, In *TALN*, p. 93–102.

GRABAR N. & HAMON T. (2016). Exploitation de la morphologie pour l'extraction automatique de paraphrases grand public des termes médicaux. *TAL*, **57**(1), 85–109.

GUNNING R. (1973). *The art of clear writing*. New York, NY : McGraw Hill.

HORN C., MANDUCA C. & KAUCHAK D. (2014). Learning a lexical simplifier using Wikipedia. In *ACL Annual Meeting*, p. 458–463.

INUI K., FUJITA A., TAKAHASHI T., IIDA R. & IWAKURA T. (2003). Text simplification for reading assistance : a project note. In *Proc. of the 2nd international workshop on paraphrasing : paraphrase acquisition and applications*, p. 9–16.

JAUHAR S. & SPECIA L. (2012). UOW-SHEF : SimpLex – lexical simplicity ranking based on contextual and psycholinguistic features. In **SEM 2012*, p. 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*, p. 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*, p. 177–180 : Association for Computational Linguistics.

KOKKINAKIS D. & GRONOSTAJ M. T. (2006). *Comparing lay and professional language in cardiovascual disordes corpora*, In *WSEAS transactions on biology and biomedicine*, p. 429–437.

LIGOZAT A., GROUIN C., GARCIA-FERNANDEZ A. & BERNHARD D. (2012). Annlor : A naïve notation-system for lexical outputs ranking. In **SEM 2012*, p. 487–492.

POPRAT M., MARKÓ K. & HAHN U. (2006). *A language classifier that automatically divides medical documents for experts and health care consumers*, In *MIE 2006 – Proceedings of the XX international congress of the european federation for medical informatics*, p. 503–508. Maastricht.

RELLO L., BAEZA-YATES R. A., BOTT S. & SAGGION H. (2013). Simplify or help ? : text simplification strategies for people with dyslexia. In *W4A*.

SACKETT D. L., ROSENBERG W. M. C., GRAY J. A. M., HAYNES R. B. & RICHARDSON W. S. (1996). Evidence based medicine : what it is and what it isn't. *BMJ*, **312**(7023), 71–72.

SAGGION H. (2017). *Automatic Text Simplification*. Morgan & Claypool Publishers.

SAJOUS F. & HATHOUT N. (2015). GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the eLex 2015 conference*, p. 405–426, Herstmonceux, England.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees.

SERETAN V. (2012). *Acquisition of Syntactic Simplification Rules for French*. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA). ID : unige :30961.

SHARDLOW M. (2014). Out in the open : Finding and categorising errors in the lexical simplification pipeline. In N. CALZOLARI, K. COUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proc. of the 9th International Conference on Language Resources and Evaluation*, LREC, Reykjavik, Iceland : European Language Resources Association (ELRA).

SINHA R. (2012). Unt-simprank : Systems for lexical simplification ranking. In **SEM 2012*, p. 493–496.

SPECIA L. (2010). *Translating from complex to simplified sentences*, In *International conference on computational processing of the portuguese language (Propor-2010)*, p. 30–39.

SPECIA L., JAUHAR S. & MIHALCEA R. (2012). Semeval-2012 task 1 : English lexical simplification. In **SEM 2012*, p. 347–355.

WILLIAMS S. & REITER E. (2005). Generating readable texts for readers with low basic skills. In *ENLG*.

WOODSEND K. & LAPATA M. (2011). *Learning to simplify sentences with quasi-synchronous grammar and integer programming*, In *EMNLP*, p. 409–420.

WUBBEN S., VAN DEN BOSCH A. & KRAHMER E. (2012). Sentence simplification by monolingual machine translation. In *ACL*, p. 1015–1024.

ZHU Z., BERNHARD D. & GUREVYCH I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, p. 1353–1361, Stroudsburg, PA, USA : Association for Computational Linguistics.