Exploitation de la morphologie pour l'extraction automatique de paraphrases grand public des termes médicaux

Natalia Grabar[*] — **Thierry Hamon**[**]

[*] *CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France*
[**] *LIMSI-CNRS, BP133, Orsay; Université Paris 13, Sorbonne Paris Cité, France*
natalia.grabar@univ-lille3.fr, hamon@limsi.fr

RÉSUMÉ. *Le domaine médical manipule des termes très spécifiques comme* blépharospasme, appendicectomie*, qui sont difficiles à comprendre par les non-spécialistes. Nous proposons une méthode automatique pour l'acquisition de paraphrases, qui soient plus faciles à comprendre que les termes originaux. La méthode est fondée sur l'analyse morphologique des termes, l'analyse syntaxique et la fouille de textes non spécialisés. L'analyse et l'évaluation des résultats indiquent que de telles paraphrases peuvent être extraites et présentent une compréhension plus facile. En fonction de paramètres, la précision varie entre 90 et 7,4 %. Ce type de ressources est utile pour plusieurs applications de TAL (recherche d'information, simplification de textes...).*

ABSTRACT. *The medical area conveys very specific terms (*p. ex.*,* blepharospasm, appendicectomy*), which are difficult to understand by people without medical training. We propose an automatic method for the acquisition of paraphrases, which we expect to be easier to understand than the original terms. The method is based on the morphological analysis of terms, syntactic analysis of texts, and text mining of non specialized texts. An analysis of the results and their evaluation indicate that such paraphrases can indeed be found and show easier understanding level. According to the setting of the method, precision of the extractions ranges between 90 and 7.4%. Such resources are useful for several NLP applications (*p. ex.*, information retrieval, text simplification...).*

MOTS-CLÉS : *domaines de spécialité, terminologie médicale, composition, analyse morphologique, paraphrase, compréhension.*

KEYWORDS: *specialized area, medical terminology, compounds, morphological analysis, paraphrasis, understanding.*

TAL. Volume 57 – n° 1/2016, pages 85 à 109

1. Introduction

Comme tout domaine de spécialité, le domaine médical utilise des termes techniques qui véhiculent un sens très spécifique, comme *blépharospasme, alexithymie, appendicectomie, desmorrhexie, lombalgie*. Si la compréhension de ces termes est aisée pour certaines catégories du personnel médical (médecins, étudiants en médecine, infirmiers, pharmaciens...), les personnes ordinaires non spécialistes du domaine médical peuvent avoir des difficultés de compréhension et d'utilisation de tels termes. Et pourtant, la compréhension de ces termes est cruciale pour les patients. Il a été ainsi montré que la compréhension de termes médicaux joue un rôle important dans un processus de santé réussi (McCray, 2005 ; Eysenbach, 2007). Toutefois, il a été également montré que de telles notions ne peuvent pas être correctement maîtrisées par les patients dans plusieurs situations réelles :

– compréhension des étapes nécessaires à la préparation et la prise de médicaments (Patel *et al.*, 2002) ;

– compréhension des notices de médicaments et des informations fournies aux patients dans les brochures et les consensus informés. Par exemple, parmi 2 600 patients recrutés dans deux hôpitaux, 26 à 60 % ne peuvent pas comprendre les informations de santé fournies (Williams *et al.*, 1995) ;

– compréhension d'informations de santé disponibles sur les sites Web à destination des patients (Hargrave *et al.*, 2003 ; Kusec, 2004), et ceci en différentes langues (anglais, espagnol, français).

Ces différentes observations peuvent avoir un impact négatif sur la communication entre les patients et les médecins, et les soins offerts aux patients (Tran *et al.*, 2009).

Ce contexte correspond à la motivation de notre travail : proposer une méthode pour l'acquisition automatique de paraphrases pour expliquer les termes médicaux techniques. Plus particulièrement, nous proposons de nous concentrer sur les termes formés par la composition néoclassique (Booij, 2010). Une des particularités de ces termes est qu'ils impliquent souvent des bases venant du latin ou du grec (comme *myocardiaque* formé avec une base latine *myo (muscle)* et une base grecque *cardia (cœur)*, ou *cholécystectomie* formé avec deux bases grecques *chole (bile)* et *ectomy (ablation chirurgicale)*, et une base latine *cystis (vessie)*), ce qui les rend sémantiquement opaques et plus difficiles à comprendre que les mots formés avec les bases disponibles dans la langue française contemporaine, comme dans {*anatomie* ; *anatomique*} ou {*livre* ; *livresque*}. Les composés médicaux se trouvent parmi les termes qui sont difficiles à comprendre par les locuteurs (Grabar *et al.*, 2014). En effet, avant que le terme puisse être compris, il est d'abord nécessaire de le décomposer et de faire le lien avec la langue générale.

Nous présentons d'abord des travaux de l'état de l'art (section 2) et précisons nos objectifs (section 3). Nous présentons ensuite le matériel utilisé (section 4), et les étapes de la méthode (section 5). Nous décrivons et discutons les résultats obtenus (sections 6 et 7), et concluons avec des orientations pour les travaux futurs (section 8).

2. État de l'art

Notre travail est lié à plusieurs domaines en TAL : lisibilité (section 2.1), simplification lexicale (section 2.2), construction de ressources dédiées (section 2.3) et décomposition de composés néoclassiques (section 2.4). Ces travaux ont un lien entre eux et, vus tous ensemble, présentent un problème de recherche assez complexe.

2.1. *Lisibilité*

Les travaux en lisibilité étudient la facilité avec laquelle un texte peut être compris. Deux types de mesures de lisibilité sont distingués : classiques et computationnelles. Les mesures classiques sont essentiellement fondées sur le calcul du nombre de caractères et/ou de syllabes dans les mots, phrases ou documents, et sur les modèles de régression linéaire (Flesch, 1948 ; Gunning, 1973). La combinaison de ces critères et de leur pondération par des coefficients permet d'obtenir un score global. L'interprétation de ces scores est effectuée en suivant des grilles correspondantes. Les mesures computationnelles peuvent impliquer les modèles vectoriels et une grande variété de descripteurs et de leurs combinaisons (Wang, 2006 ; Chmielik et Grabar, 2011 ; François et Fairon, 2013). De telles mesures peuvent être appliquées à différents types de documents, en fonction des objectifs des études (textes spécialisés, textes du domaine général...). Elles sont essentiellement exploitées en mode supervisé et nécessitent de disposer de documents déjà catégorisés (par exemple, les documents pour les experts en médecine et pour les non-experts, ou bien les documents pour les personnes n'ayant pas fait d'études supérieures, avec deux ans d'études supérieures, avec cinq ans d'études supérieures, etc.). Les descripteurs sont alors calculés au sein de chaque catégorie et permettent de produire les modèles spécifiques à ces catégories.

2.2. *Simplification lexicale*

La simplification lexicale aide à rendre un texte plus facile à comprendre. Les stratégies de simplification lexicale dépendent du public visé.

Une des approches proposées exploite la ressource WordNet pour la substitution lexicale (avec des difficultés qui subsistent face à l'ambiguïté des mots et au choix de bons synonymes étant donné un contexte), ainsi que des données issues de la Toile afin de fournir des informations pour exemplifier les entités nommées (personnes et lieux) (Lal et Ruger, 2002). Une explication du vocabulaire est aussi proposée pour les apprenants de l'anglais (Burstein *et al.*, 2007). Par ailleurs, d'autres indicateurs peuvent être exploités, comme la présence d'un mot donné dans une base de données psycholinguistiques (Quinlan, 1992), la probabilité de rencontrer ce mot dans un corpus de textes faciles et le nombre de syllabes que ce mot contient (De Belder *et al.*, 2010). Nous notons aussi qu'un travail (Carroll *et al.*, 1998) traite de manière combinée les deux aspects de la simplification (syntaxique et lexicale) : la chaîne de traitement proposée est composée de plusieurs étapes (étiquetage morphosyntaxique,

analyse morphologique, analyseur syntaxique et résolution d'anaphores), tandis que les modules de simplification syntaxique et lexicale exploitent les travaux existants, respectivement (Chandrasekar et Srinivas, 1997 ; Devlin, 1999).

En 2012, la compétition *SemEval*[1] proposait une tâche de simplification de textes de la langue générale anglaise. Pour un texte court et un mot cible, plusieurs substitutions possibles et satisfaisant le contexte, l'objectif était de trier ces substitutions selon leur degré de simplicité (Specia *et al.*, 2012). Plusieurs critères ont été exploités par les participants : lexique d'un corpus oral et de Wikipédia, n-grammes de Google, WordNet (Sinha, 2012) ; longueur de mots, nombre des syllabes, information mutuelle, fréquence (Jauhar et Specia, 2012) ; fréquence dans Wikipédia, longueur de mots, n-grammes, complexité syntaxique des documents (Johannsen *et al.*, 2012) ; n-grammes, fréquence dans Wikipédia, n-grammes de Google (Ligozat *et al.*, 2012) ; WordNet, fréquence (Amoia et Romanelli, 2012). Les critères liés à la fréquence de mots sont parmi les plus efficaces pour la tâche. Notons cependant qu'une étape préalable à la simplification concerne la détection de mots ou de passages difficiles (Grabar *et al.*, 2014) qui devraient être simplifiés avec les méthodes proposées plus haut.

À côté de la simplification lexicale, il existe plusieurs travaux en simplification syntaxique, qui visent à rendre la structure syntaxique des phrases plus légère (Chandrasekar et Srinivas, 1997 ; Siddharthan, 2006 ; Max, 2008 ; Brouwers *et al.*, 2012).

Nous pouvons indiquer deux initiatives institutionnelles et politiques pour la simplification de documents. La première concerne la fondation Cochrane[2] dont l'objectif est de fournir des informations médicales avec un niveau de preuve élevé (Sackett *et al.*, 1996). Dans le souci de rendre les informations médicales accessibles et compréhensibles pour différentes catégories de personnes, les collaborateurs de Cochrane proposent, à côté des résumés à destination des experts, des résumé simplifiés (ou *Plain language summary*). Une deuxième initiative concerne la mouvance FALC (*facile à lire et à comprendre*) apparue suite à la législation européenne entrée en vigueur en janvier 2015. La motivation de telles lois est similaire aux motivations de notre travail : en effet, différents types d'informations institutionnelles restent difficilement compréhensibles par le grand public. Ainsi, la mouvance FALC vise à rendre l'information accessible à tous, et principalement aux personnes en situation de handicap intellectuel, mais aussi aux personnes ayant des difficultés d'accès à la lecture, aux personnes étrangères, etc. Ce processus repose sur la simplification du langage utilisé, une rédaction et une mise en page spécifiques pour faciliter la compréhension des informations. Mentionnons que ceci devient une obligation européenne et que tout établissement en contact avec le public doit avoir au moins un document rédigé selon les principes FALC.

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

2.3. *Ressources dédiées*

Des ressources spécifiques sont nécessaires pour effectuer la simplification des textes. Dans le domaine médical, comme dans tout domaine de spécialité, les ressources se présentent souvent sous forme de lexiques où les termes techniques sont mis en correspondance avec les expressions non spécialisées correspondantes, comme dans les exemples (1) à (4). La première initiative de ce type est apparue avec le travail collaboratif *Consumer Health Vocabulary* (CHV) (Zeng et Tse, 2006) (exemples (1)). Une des méthodes proposées consiste à utiliser les requêtes médicales les plus fréquentes et à les aligner avec les termes d'UMLS (*Unified Medical Language System*) (Lindberg *et al.*, 1993). Ensuite, les alignements sont validés manuellement. Un autre travail a exploité un petit corpus et plusieurs mesures d'association statistique pour construire un lexique de termes techniques alignés avec leurs équivalents non techniques (Elhadad et Sutaria, 2007), comme présentés dans les exemples (4). Ici, les deux ensembles de termes, techniques et non techniques, étant fournis par l'UMLS sont possiblement dérivés du *Consumer Health Vocabulary* et représentent un sous-ensemble de celui-là. Des travaux similaires dans d'autres langues ont suivi. En français, l'acquisition de variations morphosyntaxiques à partir d'un corpus comparable spécialisé et non spécialisé (Deléger et Zweigenbaum, 2008 ; Cartoni et Deléger, 2011) a fourni des équivalences verbe/nom (exemples (2)) et un ensemble de variations syntaxiques plus large (exemples (3)). Dans ces deux travaux, la correspondance avec les terminologies médicales n'est pas établie.

(1) {*myocardial infarction* ; *heart attack*}, {*abortion* ; *termination of pregnancy*}, {*acrodynia* ; *pink disease*}

(2) {*consommation régulière* ; *consommer de façon régulière*}, {*gêne à la lecture* ; *empêche de lire*}, {*évolution de l'affection* ; *la maladie évolue*}

(3) {*retard de cicatrisation* ; *retarder la cicatrisation*}, {*apports caloriques* ; *apport en calories*}, {*calculer les doses* ; *doses sont calculées*}, {*efficacité est renforcée* ; *renforcer son efficacité*}

(4) {*myocardial infarction* ; *heart attack*}, {*SBP* ; *systolic blood pressure*}, {*atrial fibrillation* ; *arrhythmia*}, {*hypercholesterolemia* ; *high cholesterol*}, {*mental stress* ; *stress*}

Notons aussi que les travaux en acquisition de variantes terminologiques (Grabar et Zweigenbaum, 2000 ; Hahn *et al.*, 2001), de synonymes (Fernández-Silva *et al.*, 2011) et de paraphrases (Max *et al.*, 2012 ; Fujita et Isabelle, 2015) sont aussi pertinents pour cette thématique de recherche.

2.4. *Décomposition de composés néoclassiques*

La décomposition de composés néoclassiques consiste à détecter leurs composants morphologiques. Dans les travaux de TAL, la décomposition est exploitée pour améliorer les résultats en indexation et en recherche d'information (Lovis *et al.*, 1995 ; Hahn *et al.*, 2001) ou en traduction automatique (Loginova-Clouet et Daille, 2013). En effet, il peut être intéressant de décomposer un terme comme *iridochoroïdite* en ses composants (*inflammation, iris*, et *choroïde*) pour trouver plus de documents ou de traductions pertinents. D'autres travaux s'intéressent de plus à l'établissement de relations sémantiques entre les composants de termes de manière manuelle (Pacak *et al.*, 1980 ; Dujols *et al.*, 1991 ; Wolff, 1987) ou automatique (Daille, 2003 ; Grabar et Hamon, 2006). Par exemple, au sein du composé *iridochoroïdite* nous pouvons établir deux relations : (1) la relation de *localisation*, car une *inflammation* est localisée dans l'*iris* et le *choroïde* ; (2) et la relation hiérarchique *est-un*, car *iridochoroïdite* est un type d'*inflammation*. La décomposition automatique des termes exploite souvent des méthodes à base de règles ou des approches probabilistes en corpus (Namer, 2009 ; Loginova-Clouet et Daille, 2013 ; Claveau et Kijak, 2014).

3. Objectifs

Le travail que nous proposons est lié à la décomposition de composés néoclassiques (section 2.4) et à la construction de ressources spécifiques (section 2.3). Notre objectif est de développer une méthode qui permet d'acquérir des paraphrases non spécialisées pour des termes techniques composés du domaine médical. Nous supposons entre autres que l'explicitation de bases grecques et latines rend la sémantique des termes plus transparente. De tels objectifs sont rarement poursuivis dans les travaux existants : seuls les exemples en (1) et (4) provenant de CHV contiennent ce type de paraphrases en anglais. Nous travaillons avec le matériel en français. Contrairement aux travaux existants, nous ne travaillons pas avec des corpus comparables spécialisés et non spécialisés, mais exploitons les termes fournis par des terminologies médicales existantes et les articles de Wikipédia. Nous supposons que Wikipédia peut contenir les paraphrases recherchées, comme dans les exemples en (5).

(5) {*myocardiaque* ; *muscle du cœur*}
{*cholécystectomie* ; *ablation de la vésicule biliaire*}

Par rapport à nos premiers travaux sur cette thématique (Grabar et Hamon, 2014a ; Grabar et Hamon, 2014b), nous nous concentrons sur l'exploitation de Wikipédia qui fournit des paraphrases plus riches (par comparaison avec les forums de discussion, où les paraphrases extraites sont très redondantes et offrent donc moins de couverture) et exploitons l'analyse syntaxique des textes et non pas des fenêtres de mots, ce qui permet d'extraire des paraphrases mieux fondées linguistiquement et de faire des comparaisons et des évaluations plus précises des données acquises. Par rapport à un autre travail publié (Grabar et Hamon, 2015), nous exploitons et analysons un

ensemble plus important de paraphrases extraites, ce qui permet de couvrir au final un nombre supérieur de termes techniques munis de paraphrases.

4. Données linguistiques

Trois types de données sont utilisés : les termes médicaux à paraphraser (section 4.1), le corpus duquel les paraphrases sont extraites (section 4.2), et les ressources linguistiques qui aident à établir le lien entre les termes et le corpus (section 4.3).

4.1. *Termes médicaux*

Les termes médicaux proviennent de la Snomed International (Côté *et al.*, 1997)[3] et de la partie française d'UMLS (Lindberg *et al.*, 1993). Ces terminologies contiennent des termes syntaxiquement simples (*acrodynie*) et complexes (*infarctus du myocarde*). Nous utilisons l'ensemble des termes disponibles, les termes syntaxiquement complexes étant segmentés en mots. Le seul filtre appliqué consiste en l'élimination de mots contenant des nombres car ceux-ci correspondent le plus souvent à des composés chimiques et sont gérés par un autre type de compositionnalité (Klinger *et al.*, 2008). Dans ce qui suit, *mot* et *terme* sont échangeables et peuvent signifier soit l'unité graphique obtenue suite à la segmentation des termes syntaxiquement complexes, soit la notion médicale.

4.2. *Corpus*

Nous exploitons les articles de Wikipédia du *Portail de la médecine* (version de janvier 2015). Avec 18 434 articles (15 235 219 occurrences), ce corpus contient des informations encyclopédiques sur des notions médicales. Les contributeurs ont en général une bonne connaissance des sujets abordés. L'objectif de Wikipédia est entre autres de présenter les notions techniques et de les rendre accessibles au grand public. Nous nous attendons à ce que ces articles contiennent des paraphrases de termes techniques présentant un niveau de compréhension accessible pour les non-spécialistes.

4.3. *Ressources linguistiques*

Ressources morphologiques. Les ressources morphologiques comportent 155 468 paires de mots couvrant les dérivations {*aorte* ; *aortique*} et les flexions {*aortique* ; *aortiques*}. Elles sont issues des travaux précédents (Grabar et Zweigenbaum, 2000). Ces ressources permettent de traiter la variation morphologique des termes.

3. Agence des systèmes d'information partagés de santé : `esante.gouv.fr/asip-sante`

Ressources de synonymes. Les ressources de synonymes proviennent également des travaux précédents (Grabar *et al.*, 2009) et ont été complétées par les synonymes simples d'UMLS. Ces ressources sont adaptées à la langue médicale. Elles contiennent 14 914 paires de synonymes, comme {*embolie* ; *thrombose*}, {*tumeur* ; *fibrome*}. Ces ressources sont également utilisées pour traiter la variation des termes.

Ressources supplétives. Ces ressources contiennent des paires de mots au format {*base supplétive* ; *mot du français*}. Ce sont les ressources qui permettent de faire le lien entre les bases latines et grecques et les mots du français moderne. Ces ressources ont été construites lors des travaux précédents (Namer, 2009 ; Zweigenbaum et Grabar, 2003). Elles ne sont pas dédiées aux expériences présentées ici, mais elles restent néanmoins spécifiques au matériel traité que sont les termes médicaux. Ces ressources fournissent 1 022 paires, comme dans ces exemples : {*andr* ; *mâle*}, {*ectomie* ; *ablation*}, {*myo* ; *muscle*}, {*para* ; *contre*}, {*péri* ; *autour*}.

5. Méthode

La méthode est définie afin d'effectuer l'analyse des composés médicaux néoclassiques et de trouver leurs paraphrases non techniques dans les corpus. Dans certains cas, les paraphrases apparaissent dans des contextes définitoires (exemple (6)), auquel cas elles *cooccurrent* avec leurs termes techniques, ou bien de manière libre et sans être accompagnées de leur terme technique (exemple (7)). C'est ce deuxième type de contexte qui nous intéresse car il n'est pas contraint par l'occurrence du terme technique. Dans ce deuxième type de contexte (exemple (7)), nous pouvons en effet trouver la paraphrase *inflammation des cellules* qui correspond au terme *cellulite*.

(6) *La cellulite est une infection grave qui se propage sous la peau et s'attaque aux tissus mous comme la peau elle-même et les graisses sous-jacentes.*

(7) *L'infection virale cause une inflammation des cellules nerveuses, conduisant à la destruction partielle ou totale du ganglion des motoneurones.*

La méthode est composée de quatre grandes étapes présentées à la figure 1 : le traitement des termes médicaux (section 5.1), le traitement du corpus (section 5.2), l'alignement des termes et des segments du corpus pour l'extraction de paraphrases grand public (section 5.3), et l'évaluation des extractions (section 5.4).

5.1. *Traitement de termes médicaux*

Nous effectuons trois traitements sur les termes médicaux.

1) *Étiquetage morphosyntaxique et lemmatisation des termes :* les termes sont étiquetés morphosyntaxiquement et lemmatisés avec `Cordial` (Laurent *et al.*, 2009), qui a subi une adaptation au domaine médical en incorporant un lexique de ce domaine

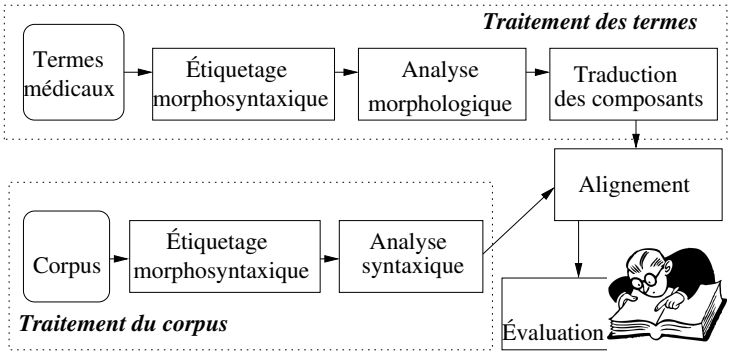

Figure 1. *Méthode générale d'extraction de paraphrases grand public des termes*

(Grabar et Zweigenbaum, 1999). L'étiquetage est effectué en contexte des termes. Si un mot donné reçoit plus d'une étiquette, c'est la plus fréquente qui est retenue. À cette étape, nous obtenons les lemmes des mots avec leurs parties du discours (exemple (8)). L'objectif est de sélectionner les catégories non grammaticales pour l'étape suivante.

(8) *myocardique/A*
cholécystectomie/N
polyneuropathie/N
acromégalie/N
galactosémie/N

2) *Analyse morphologique :* les lemmes des verbes, noms et adjectifs sont ensuite analysés morphologiquement par `DériF` (Namer, 2009). `DériF` exploite une méthode à base de règles, issue des travaux linguistiques (Corbin, 1987), une liste d'exceptions et une liste de références issues de TLFi (TLFi, 2001). Cet outil effectue une analyse des lemmes afin de calculer leur structure morphologique, de les décomposer en leurs composants (bases et affixes), et de les analyser sémantiquement. Nous présentons des exemples de l'analyse morphologique de quelques termes en (9).

(9) *myocardique/A* : *[[[myo N*] [carde N*] NOM] ique ADJ]*
cholécystectomie/N : *[[cholécysto N*] [ectomie N*] NOM]*
polyneuropathie/N : *[poly [[neur N*] [pathie N*] NOM] NOM]*
acromégalie/N : *[[acr N*] [mégal N*] ie NOM]*
galactosémie/N : *[[galactose NOM] [ém N*] ie NOM]*

Les bases et affixes calculés sont associés avec les catégories syntaxiques (*NOM, ADJ, V*). Lorsqu'une base est supplétive (elle est empruntée au latin ou au grec et n'existe pas en français moderne), `DériF` lui affecte la catégorie la plus probable (*N** pour les noms, *A** pour les adjectifs, *V** pour les verbes). Ainsi, l'analyse de *myocardique/A* indique que ce mot contient deux bases supplétives nominales *myo N* (muscle)* et *carde*

N (cœur)* et un affixe adjectival *-ique/ADJ*. À cette étape, les mots sont décomposés en leurs composants morphologiques. Nous observons que certaines bases (*galactose* et *cholécysto*) peuvent être décomposées encore plus finement, en *galact (lait)* et *ose (sucres)* pour *galactose*, et *chole (bile/biliaire)* et *cystis (vésicule)* pour *cholécysto*. Nous considérons que les mots qui contiennent plus d'une base sont des composés. Ils sont traités lors des étapes suivantes. Très souvent, ces bases sont empruntées au latin ou au grec, mais nous trouvons également des bases du français (*insulinodépendant, code-barres*). Comme présenté dans les exemples en (10), `Dérif` fournit également des gloses en langage semi-formel pour expliquer le sens des composés analysés.

(10) *myocardique/A* : *« (Partie de – Type particulier de) cœur en rapport avec le(s) muscle »*
cholécystectomie/N : *« ablation (de – vers) le(s) vésicule biliaire »*
polyneuropathie/N : *« neuropathies multiples, nombreux »*
acromégalie/N : *« Affection liée au(x) grandeur en rapport avec le(s) extrémité »*
galactosémie/N : *« Affection liée au(x) sang en rapport avec le(s) galactose »*

3) *Association des composants morphologiques avec les mots du français :* les bases latines et grecques obtenues suite à la décomposition sont associées avec les mots du français moderne grâce à la ressource supplétive présentée à la section 4.3. Les exemples en (11) présentent les données obtenues à cette étape.

(11) *myocardique/A* : *myo* = *muscle*, *carde* = *cœur*
cholécystectomie/N : *cholécysto* = *vésicule biliaire*, *ectomie* = *ablation*
polyneuropathie/N : *poly* = *nombreux*, *neuro* = *nerf*, *pathie* = *maladie*
acromégalie/N : *acr* = *extrémité*, *mégal* = *grandeur*
galactosémie/N : *galactose* = *galactose*, *ém* = *sang*

À cette étape, certains mots restent techniques (*galactose*, *vésicule biliaire*), alors que d'autres perdent complètement leur technicité (*mégal* = *grandeur*, *poly* = *nombreux*).

5.2. *Traitement du corpus*

Le corpus est traité par `Cordial` (Laurent *et al.*, 2009) pour effectuer l'étiquetage morphosyntaxique et la lemmatisation qui permettent de produire une analyse syntaxique des textes, utilisée pour définir les frontières des syntagmes.

5.3. *Extraction de paraphrases grand public correspondant aux termes techniques*

Les mots du français qui correspondent à la décomposition morphologique des termes sont projetés sur le corpus pour en extraire les syntagmes qui contiennent les paraphrases. Nous considérons tout syntagme syntaxique, de même que les bi-

grammes, les trigrammes et les quadrigrammes de syntagmes. Dans l'exemple (12), un des groupes nominaux contient les mots *muscle* et *cœur*, qui correspondent aux composants morphologiques de *myocardique* (exemple (11)). Ce groupe nominal est donc un bon candidat pour fournir une paraphrase des termes *myocarde* ou *myocardique*.

(12) *Les causes de tachycardie ventriculaire sont superposables à celles des extrasystoles ventriculaires : infarctus du myocarde, insuffisance cardiaque, hypertrophie du muscle du cœur et prolapsus de la valve mitrale.*

Nous effectuons plusieurs expériences en faisant varier quatre paramètres :

1) la taille de la fenêtre, qui varie d'un à quatre syntagmes syntaxiques, ce qui permet de récupérer les segments avec des paraphrases plus ou moins grandes, et donc de paraphraser des termes avec plus de composants ;

2) les ressources linguistiques pour gérer la variation terminologique. Nous avons trois possibilités : l'utilisation de formes brutes, l'utilisation de ressources morphologiques pour normaliser les flexions et les dérivations, l'utilisation de synonymes pour gérer les relations de synonymie au sein des paraphrases. Actuellement, nous n'effectuons pas la combinaison des ressources morphologiques et de synonymie ;

3) le taux d'alignement des termes techniques, ce qui permet de contrôler que tous les composants de ces termes sont alignés ;

4) le taux d'alignement des syntagmes syntaxiques, ce qui permet de contrôler que tous les mots des syntagmes sont alignés avec les composants.

Comme *baseline*, nous utilisons les contextes définitoires où les termes apparaissent. Les définitions (comme en (6)) sont extraites grâce aux patrons proposés dans la littérature (Péry-Woodley et Rebeyrolle, 1998), comme *est un*, *défini comme*. Avec cette approche, nous devons d'abord détecter le terme technique et ensuite le contexte définitoire correspondant. Si le test est positif, la phrase entière est extraite.

5.4. *Évaluation des extractions*

L'évaluation vise à vérifier que la méthode proposée permet d'acquérir les paraphrases de termes médicaux. Les extractions sont évaluées manuellement, ce qui permet de calculer la précision. Pendant l'évaluation, nous distinguons quatre situations :

1) la paraphrase est correcte : comme *myocardique* paraphrasé en *muscle du cœur* ;

2) la méthode exploite des informations incorrectes, non directement liées à la méthode, mais aux ressources ou bien à des prétraitements :

- l'extraction de la paraphrase est fondée sur une analyse morphologique incorrecte : {*sanglot* ; *lot sang*},

- la traduction vers le français n'est pas satisfaisante : *antisolaire* associé avec *sol* et *contre* au lieu d'être associé avec *soleil/solaire* et *contre*,

- le terme traité n'est pas compositionnel et ses composants ne traduisent pas sa sémantique : *ostéodermie* est associé avec *peau* et *os*, alors que le terme signifie *une structure d'écailles, de plaques osseuses ou d'autres compositions dans les couches dermiques de la peau, comme chez les lézards ou dinosaures*. Les deux éléments morphologiques (*peau* et *os*) sont présents dans la signification du terme mais sa sémantique complète va au-delà. Dans ce cas, les paraphrases recueillies ne sont pas suffisantes pour représenter correctement la sémantique du terme ;

3) la paraphrase contient les informations correctes au milieu d'autres informations ou bien des informations partielles. Par exemple *endophtalmie* est paraphrasé en *interne de l'œil*, alors que son explication complète est plus large *inflammation des tissus internes de l'œil*. Ici, bien que partielles, les paraphrases recueillies représentent la sémantique du terme assez correctement ;

4) l'extraction est fausse et ne contient pas les informations utiles.

Ce type d'évaluation permet de calculer quatre mesures : la précision stricte $P_{stricte}$ qui évalue les paraphrases utilisables directement, avec seulement les paraphrases correctes (cas 1) ; la précision lâche P_{lache} qui évalue les paraphrases utilisables moyennant un post-traitement, avec les paraphrases correctes et possiblement correctes (cas 1 et 3) ; le taux d'erreurs qui évalue le taux d'extractions fausses (cas 4) ; le rappel R qui évalue la couverture des paraphrases extraites. Nous pouvons calculer deux types de rappel : rappel strict R_{strict} (cas 1) et rappel lâche R_{lache} (cas 1 et 3). Le rappel est calculé par rapport aux 15 038 termes décomposés en deux bases au moins.

Les résultats sont présentés dans la section 6. Ils sont ensuite analysés du point de vue de la qualité des extractions (sections 7.1 à 7.3). Nous comparons aussi les résultats de la *baseline* avec les paraphrases extraites (section 7.4). Nous effectuons également une comparaison avec les travaux de l'état de l'art (section 7.5), et nous examinons finalement les termes qui ne reçoivent pas de paraphrases (section 7.6).

6. Résultats

Les 274 131 termes d'UMLS et de la Snomed International fournissent 76 536 mots sans nombres. 15 038 mots sont analysés par `Dérif` et décomposés en deux bases au moins. Ces 15 038 composés correspondent au matériel pour l'acquisition de paraphrases. Nous distinguons quatre ensembles quant à l'appariement entre les termes décomposés et les syntagmes, que nous exemplifions avec *myopathie* décomposé en *muscle* et *maladie* (les segments alignés sont soulignés) :

$E1$: les deux unités, le terme et le syntagme, sont complètes dans l'alignement : {*myo pathie* ; *maladie du muscle*} ;

$E2$: le terme est complet mais le syntagme est partiel dans l'alignement : {*myo pathie* ; *maladie du muscle cardiaque*} ;

$E3$: le terme est partiel mais le syntagme est complet dans l'alignement : {*myopathie* ; *la maladie*} ;

$E4$: le terme et le syntagme sont partiels dans l'alignement : {*myopathie* ; *l' origine de la maladie*}.

Nous pouvons gérer cet aspect grâce aux taux d'alignement calculés. Nous considérons qu'il est plus intéressant d'avoir un alignement complet du terme avec un alignement complet ou partiel du syntagme, ce qui correspond aux ensembles $E1$ et $E2$. L'ensemble $E1$ est le plus optimisé car il propose l'information recherchée plus exactement. Cependant, $E2$ est aussi à prendre en compte car il est possible de déduire, à partir du syntagme, la paraphrase requise. En plus du taux d'alignement, les informations sur les fréquences sont aussi disponibles : le nombre de fois où un syntagme donné est extrait du corpus. Cependant, ces informations ne sont pas indicatives car un mauvais candidat à la paraphrase peut être fréquent dans le corpus.

Nombre de	*unigrammes*			*bigrammes*		
	b	*l*	*s*	*b*	*l*	*s*
syntagmes	9 854	16 093	22 110	11 875	18 504	27 670
termes uniques	1 513	1 947	2 090	1 780	2 260	2 463
$syntagmes_{E1}$	2 681	4 163	5 370	1 109	1 611	2 521
$termes\ uniques_{E1}$	668	1 023	1 051	492	670	962
$syntagmes_{E2}$	3 893	6 486	8 876	3 937	6 290	9 590
$termes\ uniques_{E2}$	1 015	1 358	1 508	1 025	1 482	1 693
	trigrammes			*quadrigrammes*		
	b	*l*	*s*	*b*	*l*	*s*
syntagmes	7 936	12 284	19 984	4 701	7 542	12 804
termes uniques	1 523	1 966	2 231	1 079	1 515	1 922
$syntagmes_{E1}$	403	634	988	326	510	793
$termes\ uniques_{E1}$	239	358	472	204	297	419
$syntagmes_{E2}$	2 154	3 380	5 138	1 171	1 947	3 241
$termes\ uniques_{E2}$	752	1 038	1 401	517	768	1 047

Tableau 1. *Résultats d'extraction de paraphrases pour les termes techniques*

Le tableau 1 présente les résultats d'extraction de paraphrases grand public dans l'ensemble $E1$. Nous indiquons d'abord le nombre des syntagmes extraits (*Nombre de syntagmes*) et le nombre de types de termes paraphrasés (*Nombre de termes uniques*) pour l'ensemble des résultats. Nous distinguons plusieurs expériences en fonction de la taille de la fenêtre syntaxique (*unigrammes* et *bigrammes*) et des ressources utilisées (*b* sans les ressources, *l* ressources pour la normalisation morphologique et *s* ressources pour la normalisation de synonymes). Nous indiquons ensuite les informations correspondantes pour les ensembles $E1$ et $E2$.

Les résultats de l'évaluation de l'ensemble $E1$ sont indiqués dans les tableaux 2 et 3, tandis que les résultats de l'évaluation de l'ensemble $E2$ se trouvent dans les

Nombre de	*unigrammes*			*bigrammes*		
	b	*l*	*s*	*b*	*l*	*s*
syntagmes$_{E1}$	2 681	4 163	5 370	1 109	1 611	2 521
termes uniques$_{E1}$	668	1 023	1 051	492	670	962
correct	549	785	644	378	517	461
pos. correct	39	32	67	22	45	75
ttt termes	47	60	44	28	28	46
incorrect	33	146	296	64	80	380
$P_{stricte}$ *(%)*	82,0	77,0	61,0	77,0	77,0	48,0
P_{lache} *(%)*	88,0	80,0	68,0	81,0	84,0	40,0
$\%_{incorrect}$	5,0	14,0	28,0	13,0	12,0	39,0
R_{strict} *(%)*	3,7	5,2	4,3	2,5	3,4	3,0
R_{lache} *(%)*	3,9	5,4	4,7	2,7	3,7	3,6

Tableau 2. *Évaluation de l'ensemble* $E1$*, unigrammes et bigrammes*

Nombre de	*trigrammes*			*quadrigrammes*		
	b	*l*	*s*	*b*	*l*	*s*
syntagmes$_{E1}$	403	634	988	326	510	793
termes uniques$_{E1}$	239	358	472	204	297	419
correct	195	290	257	175	254	235
pos. correct	10	19	41	7	13	35
ttt termes	9	10	26	9	9	26
incorrect	25	39	148	13	21	123
$P_{stricte}$ *(%)*	82,0	81,0	55,0	86,0	86,0	56,0
P_{lache} *(%)*	86,0	86,0	63,0	89,0	90,0	64,0
$\%_{incorrect}$	11,0	11,0	31,0	6,0	7,0	29,0
R_{strict} *(%)*	1,3	1,9	1,7	1,2	1,7	1,6
R_{lache} *(%)*	1,4	2,1	2,0	1,2	1,8	1,8

Tableau 3. *Évaluation de l'ensemble* $E1$*, trigrammes et quadrigrammes*

tableaux 4 et 5. Nous indiquons plusieurs informations : le nombre de paraphrases correctes (*correct*), le nombre de paraphrases possiblement correctes (*pos. correct*), le nombre de paraphrases dont l'analyse morphologique ou la « traduction » doivent être améliorées (*ttt termes*), et le nombre de paraphrases incorrectes (*incorrect*).

La précision varie en fonction des ressources exploitées : elle est la plus élevée dans l'ensemble $E1$ et lorsque aucune ressource n'est utilisée. Elle est la moins élevée avec les synonymes, où le risque de produire des alignements erronés est plus important. Avec l'ensemble $E1$, la précision stricte varie entre 86 et 55 %, et la précision

Nombre de	*unigrammes*			*bigrammes*		
	b	*l*	*s*	*b*	*l*	*s*
syntagmes$_{E2}$	3893	6486	8876	3937	6290	9590
termes uniques$_{E2}$	1015	1358	1508	1025	1482	1693
correct	183	229	185	111	138	125
pos. correct	537	754	572	654	876	725
ttt termes	96	125	163	94	119	177
incorrect	199	250	588	166	349	666
$P_{stricte}$ *(%)*	18,0	16,9	12,3	10,8	9,3	7,4
P_{lache} *(%)*	70,9	72,4	50,2	74,6	68,4	50,2
$\%_{incorrect}$	19,6	18,4	39,0	16,2	23,5	39,3
R_{strict} *(%)*	1,2	1,5	1,2	0,7	0,9	0,8
R_{lache} *(%)*	4,8	6,5	5,0	5,1	6,7	5,7

Tableau 4. *Évaluation de l'ensemble* $E2$*, unigrammes et bigrammes.*

Nombre de	*trigrammes*			*quadrigrammes*		
	b	*l*	*s*	*b*	*l*	*s*
syntagmes$_{E2}$	2154	3380	5138	1171	1947	3241
termes uniques$_{E2}$	752	1038	1401	517	768	1047
correct	98	119	110	81	109	95
pos. correct	443	626	540	316	471	404
ttt termes	63	87	138	53	75	121
incorrect	148	206	613	67	113	427
$P_{stricte}$ *(%)*	13,0	11,5	7,9	15,7	14,2	9,1
P_{lache} *(%)*	71,9	71,8	46,4	76,8	75,5	47,7
$\%_{incorrect}$	19,7	19,8	43,8	13,0	14,7	40,8
R_{strict} *(%)*	0,7	0,8	0,7	0,5	0,7	0,6
R_{lache} *(%)*	3,6	5,0	4,3	2,6	3,9	3,3

Tableau 5. *Évaluation de l'ensemble* $E2$*, trigrammes et quadrigrammes.*

lâche entre 90 et 40 %. Le taux d'erreurs varie entre 5 et 39 %. Avec l'ensemble $E2$, la précision stricte varie entre 18 et 7,4 %, et la précision lâche entre 75,5 et 46,4 %. Le taux d'erreurs varie entre 13 et 43,8 %. Comme l'appariement est plus lâche dans l'ensemble $E2$, car il autorise des syntagmes plus longs du côté du corpus, le risque de produire des paraphrases incorrectes ou possiblement correctes est plus important. Nous voyons d'ailleurs que la précision stricte est beaucoup plus faible dans l'ensemble $E2$ par rapport à l'ensemble $E1$, mais aussi par rapport à la précision lâche de $E2$. En effet, les syntagmes extraits sont en majorité plus larges par rapport à ce qui est attendu, comme dans {*myo pathie* ; *maladie du muscle cardiaque*}. De tels syntagmes

demandent donc un post-traitement pour les réduire aux paraphrases nécessaires et suffisantes pour expliquer les termes techniques. Cependant, pour cette même raison, $E2$ permet d'extraire plus de paraphrases candidates et de paraphrases correctes ou possiblement correctes. En fonction des expériences et des ressources, $E1$ couvre entre 1,2 % et 5,4 % des 15 038 termes composés analysés morphologiquement. Dans $E2$, le rappel va de 0,5 % à 6,5 %. Le rappel lâche dans $E2$ est globalement supérieur à celui de $E1$. $E1$ fournit 1 031 paraphrases correctes et 1 128 paraphrases correctes et possiblement correctes, tandis que $E2$ fournit 336 paraphrases correctes et 1 485 paraphrases correctes et possiblement correctes. Au total, ces deux ensembles fournissent 1 179 paraphrases correctes et 1 665 paraphrases correctes et possiblement correctes.

7. Discussion

7.1. *Analyse morphologique des termes*

La décomposition et l'analyse morphologique de `Dérif` peuvent fournir quelques erreurs ou ambiguïtés. Nous avons par exemple des décompositions ambiguës, où il existe plus d'une décomposition possible mais dont une seule est correcte. Par exemple, *posturographie* est décomposé en : *[post [[uro N*] [graphie N*] NOM] NOM]*, ce qui peut être glosé *contrôle pendant la période qui suit la thérapie faite sur le système urinaire*. Cependant, la décomposition correcte est *[[posturo N*] [graphie N*] NOM]*, glosée *définition de la position optimale du corps en posture assise ou debout*.

Certaines décompositions sont erronées, comme en (13). Pour ces exemples, nous indiquons les termes, leur décomposition et un exemple de paraphrases extraites. Les décompositions erronées produisent des paraphrases erronées à l'étape suivante.

(13) - *sanglot* (*lot* et *sang*) : *des lots de sang*
- *exotique* (*externe* et *oreille*) : *l'oreille externe*
- *divin* (*deux* et *vin*) : *deux litres de vin*

7.2. *Extraction de paraphrases et leur évaluation*

Nous extrayons plusieurs paraphrases correctes et intéressantes, comme celles obtenues avec l'appariement brut en (14), l'appariement avec la normalisation morphologique en (15) et l'appariement avec la normalisation sémantique en (16).

(14) - *podalgie : douleur du pied*
- *mastite : inflammation du sein*
- *cystoprostatectomie : ablation de la vessie et de la prostate*
- *thoracostomie : l'ouverture du thorax*
- *antipapillomavirus : contre le papillomavirus humain*

- *tachycardique : cœur trop rapide*
- *myoplégie : paralysie des muscle*
- *parodontopathie : maladie touchant le parodonte*
- *anophtalmie : absence d'œil*
- *glossalgie : douleur au niveau de la langue*
- *anurique : absence de production d'urine*

(15)
- *desmorrhexie* : *rupture des ligaments*
- *bronchite : inflammation des bronches, inflammation bronchique* (bronche → bronches, bronche → bronchique)
- *dentalgie : douleurs dentaires* (dents → dentaires)
- *périovulatoire : autour de la date d'ovulation* (ovulatoire → ovulation)
- *synovite : inflammation de la gaine synoviale* (synovie → synoviale)
- *microangiopathie : des maladies des petits vaisseaux sanguins* (maladie → maladies, vaisseau → vaisseaux, petit → petits)
- *vaso-dilatateur : dilater les vaisseaux sanguins* (vaisseau → vaisseaux)

(16)
- *aclasie : absence de fracture* (cassure → fracture)
- *entérectomie : résection des intestins* (ablation → résection)
- *tumorectomie : exérèse chirurgicale de la tumeur* (ablation → exérèse)
- *tensiomètre : mesurer la pression artérielle* (tension → pression)
- *myographie : enregistrement de l'activité d'un muscle* (mesure → enregistrement)
- *mycotoxique : toxine produite par un champignon* (toxique → toxine)

Les paraphrases extraites correspondent le plus souvent aux groupes nominaux, comme dans la plupart des exemples (14) à (16). Cependant, nous pouvons également extraire d'autres types de groupes syntaxiques, comme par exemple :

– des groupes prépositionnels : {*périovulatoire* ; *autour de la date d'ovulation*}, {*akinésie* ; *en l'absence de tout mouvement*} ;

– des groupes participiaux : {*malformatif* ; *mal formé*}, {*aortite* ; *se caractérisant par une inflammation de l'aorte*} ;

– des groupes verbaux : {*sinoscopie* ; *observer les sinus maxillaires*}, {*tensiometre* ; *mesurer la pression artérielle*} ;

– des subordonnées : {*agalactie* ; *qui se caractérise par l'absence de lait*}.

Certaines de ces paraphrases peuvent être utilisées en l'état lors de la tâche de simplification par exemple, d'autres devront être post-traitées pour devenir utilisables.

Parmi les paraphrases erronées, nous trouvons parfois des erreurs de relations sémantiques entre les composants. Il s'agit typiquement de proposer la coordination entre les composants qui sont en relation de subordination, comme dans *hématospermie* paraphrasée en *le sang ou le sperme* au lieu de *présence de sang dans le sperme*. Mais le plus souvent, les corpus fournissent les relations sémantiques correctes entre les composants. Ceci correspond à un grand avantage de la méthode fondée sur l'ana-

lyse syntaxique. En effet, dans le travail précédent (Grabar et Hamon, 2014b), qui exploitait des corpus plus grands et dont la méthode d'extraction était fondée sur la fenêtre graphique d'une largeur donnée, le taux d'erreurs était beaucoup plus élevé, pouvant atteindre 59 % pour un nombre de termes paraphrasés moindre (273 termes avec des paraphrases correctes et 343 termes avec des paraphrases correctes et possiblement correctes). D'autres paraphrases incorrectes concernent les termes qui ne sont pas compositionnels, comme *ostéodermie* ou *causalgie*, et dont le sens précis ne peut plus être dérivé de leurs composants.

Une importante partie de termes paraphrasés sont des termes à deux composants. Les termes à trois composants ou plus sont rares. L'augmentation de la fenêtre syntaxique permet d'augmenter la taille des termes paraphrasés. Actuellement, la longueur moyenne des termes paraphrasés varie entre 2,002 et 2,125 composants. Le rappel varie entre 0,5 % et 6,5 % de termes analysés morphologiquement. L'augmentation de la couverture est une perspective importante de notre travail.

7.3. *Impact des ressources linguistiques*

Comme nous pouvons le voir dans les tableaux 2 et 3, l'utilisation de ressources linguistiques permet d'augmenter la couverture car plus de propositions sont alors extraites, en revanche cela diminue la précision car les propositions risquent d'apporter du bruit. Nous pouvons aussi voir que l'utilisation de synonymes mène à l'extraction d'un plus grand pourcentage de propositions erronées. Comme dans d'autres tâches en recherche et extraction d'information, l'explication principale est que les synonymes correspondent souvent à des valeurs contextuelles : selon les contextes ils sont plus ou moins acceptables. En revanche, les ressources morphologiques contiennent des paires de mots dont la substituabilité contextuelle est plus évidente : la variation flexionnelle ou dérivationnelle n'apporte que peu de changements sémantiques. En (17), nous présentons quelques exemples d'erreurs dues à l'utilisation de synonymes. Pour un composé néoclassique donné (comme *cardialgie*), nous indiquons sa sémantique attendue (*douleur du cœur*) et parfois sa décomposition. Nous présentons ensuite la ou les paraphrases erronées extraites pour ce composé (*plaie du cœur*) et la raison de cette extraction. Dans l'exemple cité, il s'agit de l'utilisation de la paire de synonymes {*douleur* ; *plaie*}. Ces synonymes sont corrects et mutuellement substituables dans plusieurs contextes, mais pas dans le contexte de la paraphrase de *cardialgie*. Notons que dans la compétition de simplification proposée par *SemEval* (Specia *et al.*, 2012), les candidats à remplacement étaient présélectionnés et satisfaisaient le contexte. Tandis que dans notre travail, la présélection des ressources n'est pas effectuée.

(17) - *cardialgie* (*douleur de cœur*) : *plaie du cœur* – {*douleur* ; *plaie*}
- *chéiropathie* (*maladie des mains*) : *le syndrome main* – {*maladie* ; *syndrome*}
- *choroïde* (est décomposé en *forme* et *membrane*, et signifie *une des couches de la paroi du globe oculaire*) : *aspect de l'épithélium* – {*forme* ; *aspect*},

{*membrane* ; *épithélium*}
- *cinépathie* (est décomposé en *mouvement* et *maladie*, est aussi connu sous le terme de *mal des transports*) : *évolution du syndrome* – {*mouvement* ; *évolution*} et {*maladie* ; *syndrome*}
- *myasthénie* (*faiblesse du muscle*) : *le muscle de la tristesse* – {*faiblesse* ; *tristesse*}

Comme nous l'avons noté, nous n'effectuons pas actuellement la combinaison de ressources morphologiques et de synonymie pour deux raisons : le coût de calcul devient alors très élevé et de plus cela multiplie les erreurs dues à la synonymie. Lorsque nous pourrons gérer mieux les valeurs contextuelles des synonymes, la combinaison de ces deux types de ressources pourra apporter des solutions pour augmenter la couverture de termes médicaux paraphrasés.

7.4. *Comparaison avec les contextes définitoires*

Le nombre total de définitions extraites avec les patrons définitoires est de 2 037, portant sur 1 286 termes uniques. Le patron le plus fréquent `est un` est reconnu le plus souvent. D'autres patrons, comme `également appelé` et `peut être défini comme`, sont aussi trouvés mais avec une fréquence moindre. Nous distinguons les définitions correctes (exemple (18)) et les définitions incorrectes ou apportant des informations non suffisantes pour la compréhension (exemple (19)). Comme pour la méthode principale, le calcul de la précision stricte est fondé sur les définitions correctes, tandis que la précision lâche accepte aussi les définitions possiblement correctes. La précision stricte est de 52,5 %, et la précision lâche de 68 %.

(18) *L'angiographie est une technique d'imagerie médicale portant sur les vaisseaux sanguins qui ne sont pas visibles sur des radiographies standards.*
La néphrite est une inflammation du rein (du grec : nephro- , le rein, et -itis , inflammation).

(19) *L'angiographie est un examen invasif.*
Les deux principales causes de néphrite sont les infections ou les maladies auto-immunes.

Les contextes considérés comme corrects fournissent les définitions pour 849 termes, alors que nous obtenons des définitions correctes ou possiblement correctes pour 1 028 termes. Parmi les termes définis, nous trouvons des termes composés (*achillodynie*, *clinodactylie*), des termes affixés (*choroïde*, *surmoi*), et des termes morphologiquement simples (*acide*, *hypnose*). En relation avec la méthode d'acquisition de paraphrases, seules les définitions pour les termes composés sont comparables. Comme les définitions portant sur les composés néoclassiques correspondent à la majorité des termes définis, nous prenons en compte toutes les définitions extraites. La qualité des définitions est variable. Certains termes sont sous-définis (*L'adénomyose*

est un type d'endométriose interne), d'autres très techniques. Par exemple, les trois définitions de *péricarde* qui suivent ont des niveaux de lisibilité variables. À notre avis, la première définition est la plus appropriée pour les non-experts :

– La couche extérieure du cœur est appelée péricarde.

– Le péricarde est un sac à double paroi contenant le cœur et les racines des gros vaisseaux sanguins.

– Le péricarde est un organe de glissement, formé de deux feuillets limitant une cavité virtuelle, la cavité péricardique, qui permet les mouvements cardiaques.

En comparaison avec la méthode principale, nous observons que les paraphrases couvrent un nombre légèrement supérieur de termes. Nous nous sommes attendus à ce résultat car l'extraction de paraphrases ne requiert pas la présence du terme analysé mais seulement de ses composants. Concernant la précision, elle est ainsi également plus élevée avec la méthode de paraphrases. Quant à l'utilité de ces définitions, nous pensons qu'elles peuvent être utilisées telles quelles ou bien transformées en paraphrases. Dans les deux cas, elles sont supplémentaires aux paraphrases extraites. L'ensemble des extractions analysées (paraphrases issues de $E1$ et $E2$, et contextes définitoires) fournit 1 966 termes définis ou paraphrasés correctement et 2 596 termes définis ou paraphrasés correctement ou possiblement correctement.

7.5. *Comparaison avec les travaux existants*

	Types de termes	*Nb. para.*	*Précision*
(Elhadad et Sutaria, 2007)	tous	152	58
(Deléger et Zweigenbaum, 2008)	m-synt.	65, 82	67, 60
(Cartoni et Deléger, 2011)	m-synt.	109	66
notre travail	composés	1 665	47-90

Tableau 6. *Bilan de comparaison avec les travaux existants*

Le tableau 6 synthétise la comparaison avec trois travaux existants (Deléger et Zweigenbaum, 2008 ; Cartoni et Deléger, 2011 ; Elhadad et Sutaria, 2007) :

– types de termes. Nous travaillons avec les composés néoclassiques contenant au moins deux bases. De tels termes sont en effet assez difficiles à comprendre par les locuteurs : des paraphrases grand public sont donc nécessaires pour apporter des informations qui facilitent la compréhension. Dans les travaux existants (exemples (1) à (3)), seul le travail sur l'anglais (Elhadad et Sutaria, 2007), qui est un sous-ensemble de CHV (Zeng et Tse, 2006), fournit des paraphrases des termes composés, tandis que les deux autres travaux (Deléger et Zweigenbaum, 2008 ; Cartoni et Deléger, 2011) se concentrent sur la variation morphosyntaxique des termes (exemples (2) et (3)) ;

– nombre de paraphrases extraites. Nous extrayons 1 179 paraphrases correctes et 1 665 paraphrases correctes et possiblement correctes. Comme indiqué dans la section

7.4, les définitions améliorent la couverture jusqu'à 2 596 termes définis ou paraphrasés correctement ou possiblement correctement. Les travaux existants proposent 65 et 82 paraphrases (Deléger et Zweigenbaum, 2008), 109 paraphrases (Cartoni et Deléger, 2011), et 152 paraphrases (Elhadad et Sutaria, 2007) ;

– *précision*. Nos valeurs de la précision lâche varient entre 90 et 40 %, en fonction des ressources et des fenêtres syntaxiques exploitées, avec une moyenne de 76 % sur l'ensemble des expériences et de 86 % sans l'utilisation de synonymes. Dans les travaux existants, la précision est de 67 % et 60 % (Deléger et Zweigenbaum, 2008), 66 % (Cartoni et Deléger, 2011), et 58 % (Elhadad et Sutaria, 2007).

Notons aussi qu'un seul (Elhadad et Sutaria, 2007) exploite des termes venant d'une terminologie. Les autres travaux exploitent le contenu des corpus et n'établissent pas de lien avec les terminologies existantes. De manière générale, notre travail va au-devant des travaux de l'état de l'art pour les paramètres discutés ici. Il est difficile de comparer nos résultats avec les travaux autour de la construction du CHV, car il s'agit d'une série de plusieurs travaux souvent faits de manière manuelle et collaborative.

Dans le tableau 7, nous proposons une comparaison des paraphrases avec les gloses de DériF. Pour un terme donné, nous indiquons la glose de DériF D et la paraphrase P. Il apparaît que les paraphrases extraites offrent des informations exprimées plus naturellement, et sont plus faciles à comprendre et à exploiter lors de la simplification de textes. Notons cependant que, grâce au langage formel, DériF propose une glose pour tous les termes analysés morphologiquement, alors que la couverture de paraphrases extraites dépend du contenu des corpus et des ressources linguistiques.

myocarde	D : *« (Partie de - Type particulier de) cœur en rapport avec le muscle »*
	P : *muscle du coeur*
desmorrhexie	D : *« rupture (du - liée au) ligament»*
	P : *rupture des ligaments*
glycémie	D : *« Affection liée au(x) sang en rapport avec le(s) sucre »*
	P : *le taux de sucre dans le sang*
hémobilie	D : *« Affection liée au(x) bile en rapport avec le(s) sang »*
	P : *fuite de la bile dans le sang*

Tableau 7. *Comparaison entre les gloses de* DériF *et les paraphrases*

7.6. *Termes non paraphrasés*

Plusieurs termes restent non paraphrasés, comme *leptoméningé* (*affaibli*, *méningé*, *hémidesmosome* (*corpuscule*, *demi*, *ligament*) ou *hémohistioblaste* (*cellule embryonnaire*, *tissu*, *sang*). Une des raisons est que certains termes contiennent plus de deux composants, ce qui rend la détection de leurs paraphrases plus difficile car les syntagmes avec l'ensemble de ces composants n'apparaissent pas dans le corpus. Nous avons vu cependant qu'avec l'augmentation des fenêtres syntaxiques la

taille des termes paraphrasés augmente également. D'autres termes non analysés contiennent des préfixes ou des composants qui apparaissent moins fréquemment dans les textes. Nous pensons que l'utilisation de corpus complémentaires permettra d'acquérir d'autres paraphrases. Un autre fait qui peut réduire le taux d'extractions de paraphrases concerne l'association de composants supplétifs avec les mots du français. En effet, plusieurs traductions sont parfois possibles mais ne peuvent pas être captées avec la méthode de traduction actuelle. D'autres méthodes, comme celle proposée dans (Claveau et Kijak, 2014), devraient être exploitées pour améliorer cet aspect.

8. Conclusion et travaux futurs

Nous avons proposé d'exploiter les articles de Wikipédia pour détecter les paraphrases pour les termes techniques du domaine médical. Nous nous sommes concentrés sur les composés néoclassiques (*myocardiaque, cholécystectomie, galactose, acromégalie*). Les données traitées sont en français. La méthode s'appuie sur l'analyse morphologique de termes, la traduction des composants de termes vers le français moderne ({*card* ; *cœur*}), et leur projection sur les syntagmes syntaxiques. La méthode permet d'extraire les paraphrases correctes et possiblement correctes pour 1 665 termes composés, tandis que les définitions fournissent des explications pour 1 028 termes. Mis ensemble, cela correspond à 2 596 termes uniques. Un des avantages de la méthode est que les relations sémantiques entre les composants sont aussi extraites à partir des textes. Nous pensons que cette méthode peut en effet être utilisée pour la création d'un lexique nécessaire pour la simplification de termes médicaux. Notons aussi que la méthode proposée traite les composés néoclassiques qui en général ne sont pas traités par les méthodes existantes, car ils ne présentent pas de similarité formelle avec leurs paraphrases.

Une des limitations actuelles est liée à la couverture des termes paraphrasés ou définis. Dans les travaux futurs, nous prévoyons d'utiliser d'autres méthodes, comme par exemple les méthodes distributionnelles (Claveau et Kijak, 2014), pour la segmentation de termes et leur association aux mots du français. Il est en effet possible qu'actuellement cette étape soit trop restrictive. Des corpus plus grands doivent aussi être exploités pour couvrir plus de matériel linguistique. De même, d'autres méthodes peuvent être utilisées pour augmenter la couverture du vocabulaire paraphrasé. Par exemple, les paraphrases déjà extraites peuvent être exploitées, en contexte ou hors contexte, pour détecter d'autres expressions qui apparaissent dans des contextes similaires et peut-être élargir ainsi la couverture de termes paraphrasés.

Nous voulons aussi traiter les termes complexes syntaxiquement (*vaporisateur hypodermique, fistule trachéo-œsophagienne, cardiopathie artérioscléreuse*), car ils peuvent aussi être difficiles à comprendre par les patients. La méthode proposée peut être appliquée à d'autres langues lorsque l'analyse morphologique et l'association aux mots de la langue peuvent être effectuées. L'objectif final de notre travail est d'exploiter la ressource, qui met en relation les termes spécialisés et leurs paraphrases grand public, pour la simplification de textes de spécialité.

9. Bibliographie

Amoia M., Romanelli M., « SB : mmSystem - Using Decompositional Semantics for Lexical Simplification », **SEM 2012*, Montréal, Canada, p. 482-486, 2012.

Booij G., *Construction Morphology*, Oxford University Press, Oxford, 2010.

Brouwers L., Bernhard D., Ligozat A.-L., François T., « Simplification syntaxique de phrases pour le français », *TALN*, p. 211-224, 2012.

Burstein J., Shore J., Sabatini J., Lee Y., Ventura M., « The automated text adaptation tool », *NAACL-HLT, Demo session*, p. 3-4, 2007.

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

Cartoni B., Deléger L., « Découverte de patrons paraphrastiques en corpus comparable : une approche basée sur les n-grammes », *TALN*, 2011.

Chandrasekar R., Srinivas B., « Automatic induction of rules for text simplification », *Knowledge Based Systems*, vol. 10, n° 3, p. 183-190, 1997.

Chmielik J., Grabar N., « Détection de la spécialisation scientifique et technique des documents biomédicaux grâce aux informations morphologiques », *TAL*, vol. 51, n° 2, p. 151-179, 2011.

Claveau V., Kijak E., « Generating and using Probabilistic Morphological Resources for the Biomedical Domain », *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 3348-3354, 2014.

Corbin D., *Morphologie dérivationnelle et structuration du lexique*, vol. 1, Presse universitaire de Lille, Lille, 1987.

Côté R. A., Brochu L., Cabana L., *SNOMED Internationale – Répertoire d'anatomie pathologique*, Secrétariat francophone international de nomenclature médicale, Sherbrooke, Québec. 1997.

Daille B., « Conceptual structuring through term variations », *Proceedings of the ACL Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, p. 9-16, 2003.

De Belder J., Deschacht K., Moens M.-F., « Lexical Simplification », *ITEC*, 2010.

Deléger L., Zweigenbaum P., « Paraphrase acquisition from comparable medical corpora of specialized and lay texts », *AMIA 2008*, p. 146-50, 2008.

Devlin S., Simplifying natural language for aphasic readers, Ph.d. thesis, University of Sunderland, Sunderland, UK, 1999.

Dujols P., Aubas P., Baylon C., Grémy F., « Morphosemantic Analysis and Translation of Medical Compound Terms », *Meth Inform Med*, vol. 30, p. 30-35, 1991.

Elhadad N., Sutaria K., « Mining a lexicon of technical terms and lay equivalents », *BioNLP*, p. 49-56, 2007.

Eysenbach G., « Poverty, Human Development, and the Role of eHealth », *J Med Internet Res*, vol. 9, n° 4, p. e34, 2007.

Fernández-Silva S., Freixa J., Cabré M., « A proposed method for analysing the dynamics of cognition through term variation », *Terminology*, vol. 17, n° 1, p. 49-73, 2011.

Flesch R., « A new readability yardstick », *Journ Appl Psychol*, vol. 23, p. 221-233, 1948.

François T., Fairon C., « Les apports du TAL à la lisibilité du français langue étrangère », *TAL*, vol. 54, nº 1, p. 171-202, 2013.

Fujita A., Isabelle P., « Expanding paraphrase lexicons by exploiting lexical variants », *NAACL-HLT*, p. 630-640, 2015.

Grabar N., Hamon T., « Terminology structuring through the derivational morphology », *FinTAL 2006*, nº 4139 in *LNAI*, Springer, p. 652-663, 2006.

Grabar N., Hamon T., « Automatic extraction of layman names for technical medical terms », *ICHI 2014*, Pavia, Italy, 2014a.

Grabar N., Hamon T., « Unsupervised method for the acquisition of general language paraphrases for medical compounds », *Computerm 2014*, Dublin, Ireland, 2014b.

Grabar N., Hamon T., « Extraction automatique de paraphrases grand public pour les termes médicaux », *TALN 2015*, Caen, France, 2015. 14 p.

Grabar N., Hamon T., Amiot D., « Automatic diagnosis of understanding of medical words », *EACL PITR Workshop*, p. 11-20, 2014.

Grabar N., Varoutas P., Rizand P., Livartowski A., Hamon T., « Automatic acquisition of Synonym Ressources and Assessment of their Impact on the Enhanced Search in EHRs », *Meth Inform Med*, vol. 48, nº 2, p. 149-154, 2009.

Grabar N., Zweigenbaum P., « Acquisition automatique de connaissances morphologiques sur le vocabulaire médical », *TALN*, p. 175-184, 1999.

Grabar N., Zweigenbaum P., « A General Method for Sifting Linguistic Knowledge from Structured Terminologies », *JAMIASUP*p. 310-314, 2000.

Gunning R., *The art of clear writing*, McGraw Hill, New York, NY, 1973.

Hahn U., Honeck M., Piotrowsky M., Schulz S., « Subword segmentation - leveling out morphological variations for medical document retrieval », *AMIA*, 229-233, 2001.

Hargrave D., Bartels U., Lau L., Esquembre C., Bouffet E., « Évaluation de la qualité de l'information médicale francophone accessible au public sur internet : application aux tumeurs cérébrales de l'enfant », *Bulletin du Cancer*, vol. 90, nº 7, p. 650-5, 2003.

Jauhar S., Specia L., « UOW-SHEF : SimpLex – Lexical Simplicity Ranking based on Contextual and Psycholinguistic Features », **SEM 2012*, Montréal, Canada, p. 477-481, 2012.

Johannsen A., Martínez H., Klerke S., Søgaard A., « EMNLP@CPH : Is frequency all there is to simplicity ? », **SEM 2012*, Montréal, Canada, p. 408-412, 2012.

Klinger R., Kolárik C., Fluck J., Hofmann-Apitius M., Friedrich C., « Detection of IUPAC and IUPAC-like chemical names », *ISMB 2008*, p. 268-276, 2008.

Kusec S., « Les sites web relatifs au diabète, sont-ils lisibles ? », *Dibète et société*, vol. 49, nº 3, p. 46-48, 2004.

Lal P., Ruger S., « Extract-based summarization with simplification », *DUC 2002 : Workshop on Text Summarization*, 2002.

Laurent D., Nègre S., Séguéla P., « Apport des cooccurrences à la correction et à l'analyse syntaxique », *TALN*, 2009.

Ligozat A., Grouin C., Garcia-Fernandez A., Bernhard D., « ANNLOR : A Naïve Notation-system for Lexical Outputs Ranking », **SEM 2012*, p. 487-492, 2012.

Lindberg D., Humphreys B., McCray A., « The Unified Medical Language System », *Methods Inf Med*, vol. 32, nº 4, p. 281-291, 1993.

Loginova-Clouet E., Daille B., « Segmentation multilingue des mots composés », *TALN 2013*, p. 564-571, 2013.

Lovis C., Michel P.-A., Baud R., Scherrer J.-R., « Word segmentation processing : a way to exponentially extend medical dictionaries », *MIE*, p. 28-32, 1995.

Max A., « Local rephrasing suggestions for supporting the work of writers », *GOTAL*, p. 324-335, 2008.

Max A., Bouamor H., Vilnat A., « Generalizing Sub-sentential Paraphrase Acquisition across Original Signal Type of Text Pairs », *EMNLP*, p. 721-31, 2012.

McCray A., « Promoting Health Literacy », *J of Am Med Infor Ass*, vol. 12, p. 152-163, 2005.

Namer F., *Morphologie, Lexique et TAL : l'analyseur DériF. TIC et Sciences cognitives*, Hermes Sciences Publishing, London, 2009.

Pacak M. G., Norton L. M., Dunham G. S., « Morphosemantic analysis of -itis forms in medical language », *Methods in Medical Informatics (MIM)*, vol. 19, n° 2, p. 99-105, 1980.

Patel V., Branch T., Arocha J., « Errors in interpreting quantities as procedures : The case of pharmaceutical labels », *Int journ med inform*, vol. 65, n° 3, p. 193-211, 2002.

Péry-Woodley M., Rebeyrolle J., « Domain and genre in sublanguage text : definitional micro-texts in three corpora », *LREC*, p. 987-992, 1998.

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

Sackett D., Rosenberg W., Gray J., Haynes R., Richardson W., « Evidence based medicine : what it is and what it isn't », *BMJ*, vol. 312, n° 7023, p. 71-2, 1996.

Siddharthan A., « Syntactic Simplification and Text Cohesion », *Research on Language & Computation*, vol. 4, n° 1, p. 77-109, 2006.

Sinha R., « UNT-SimpRank : Systems for Lexical Simplification Ranking », **SEM 2012*, p. 493-496, 2012.

Specia L., Jauhar S., Mihalcea R., « SemEval-2012 Task 1 : English Lexical Simplification », **SEM 2012*, p. 347-355, 2012.

TLFi, *Trésor de la Langue Française - I*, INaLF/ATILF, 2001. Disponible à l'adresse `www.tlfi.fr`.

Tran T., Chekroud H., Thiery P., Julienne A., « Internet et soins : un tiers invisible dans la relation médecine/patient ? », *Ethica Clinica*, vol. 53, p. 34-43, 2009.

Wang Y., « Automatic recognition of text difficulty from consumers health information », *in* IEEE (ed.), *Computer-Based Medical Systems*, p. 131-136, 2006.

Williams M., Parker R., Baker D., Parikh N., Pitkin K., Coates W., Nurss J., « Inadequate functional health literacy among patients at two public hospitals », *JAMA*, vol. 274, n° 21, p. 1677-1682, 1995.

Wolff S., « Automatic Coding of Medical Vocabulary », *in* N. Sager, C. Friedman, M. S. Lyman (eds), *Medical Language Processing. Computer Management of Narrative Data*, Addison-Wesley, New-York, chapter 7, p. 145-162, 1987.

Zeng Q., Tse T., « Exploring and developing Consumer Health Vocabularies », *JAMIA*, vol. 13, p. 24-29, 2006.

Zweigenbaum P., Grabar N., « Corpus-based associations provide additional morphological variants to medical terminologies », *AMIA*, 2003.