Extraction automatique de paraphrases grand public pour les termes médicaux

Natalia Grabar[1] Thierry Hamon[2]
(1) CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France
natalia.grabar@univ-lille3.fr
(2) LIMSI-CNRS, BP133, Orsay; Université Paris 13, Sorbonne Paris Cité, France
hamon@limsi.fr

Résumé. Nous sommes tous concernés par notre état de santé et restons sensibles aux informations de santé disponibles dans la société moderne à travers par exemple les résultats des recherches scientifiques, les médias sociaux de santé, les documents cliniques, les émissions de télé et de radio ou les nouvelles. Cependant, il est commun de rencontrer dans le domaine médical des termes très spécifiques (*e.g.*, *blépharospasme, alexitymie, appendicectomie*), qui restent difficiles à comprendre par les non spécialistes. Nous proposons une méthode automatique qui vise l'acquisition de paraphrases pour les termes médicaux, qui soient plus faciles à comprendre que les termes originaux. La méthode est basée 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 trouvées dans les documents non spécialisés et présentent une compréhension plus facile. En fonction des paramètres de la méthode, la précision varie entre 86 et 55 %. Ce type de ressources est utile pour plusieurs applications de TAL (*e.g.*, recherche d'information grand public, lisibilité et simplification de textes, systèmes de question-réponses).

Abstract.
Automatic extraction of layman paraphrases for medical terms.
We all have health concerns and sensibility to health information available in the modern society through modern media, such as scientific research, health social media, clinical documents, TV and radio broadcast, or novels. However, medical area conveys very specific notions (*e.g.*, *blepharospasm, alexitymia, appendicectomy*), which are difficult to understand by people without medical training. We propose an automatic method for the acquisition of paraphrases for technical medical terms. We expect that such paraphrases are 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 in non specialized documents and show easier understanding level. According to the setting of the method, precision of the extractions ranges between 86 and 55%. This kind of resources is useful for several Natural Language Processing applications (*e.g.*, information retrieval for lay people, text readability and simplification, question and answering systems).

Mots-clés : Domaines de spécialité, terminologie médicale, composition, analyse morphologique, paraphrase, compréhension.

Keywords: Specialized Area, Medical Terminology, Compounds, Morphological Analysis, Paraphrasis, Understanding.

1 Introduction

Nous sommes tous concernés par notre état de santé et restons sensibles aux informations de santé disponibles dans la société moderne à travers par exemple les résultats des recherches scientifiques, les médias sociaux de santé, les documents cliniques, les émissions de télé et de radio ou les nouvelles. Cependant, il est commun de rencontrer dans le domaine médical des termes très spécifiques, comme ceux présentés en exemple (1). Si la compréhension de ces termes est aisée pour certaines catégories du personnel médical (*e.g.*, médecins, étudiants en médecine, infirmiers, pharmaciens), les citoyens ordinaires non spécialistes du domaine médical peuvent avoir des difficultés de compréhension et d'utilisation de tels termes.

(1) *blépharospasme, alexitymie, apendicectomie, desmorrhexie, lombalgie*

La compréhension de tels termes est importante pour les patients et il a été montré qu'elle joue un rôle crucial pour un processus de santé réussi (AMA, 1999; McCray, 2005; Eysenbach, 2007). Toutefois, il a été également montré que ces 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 dans ces sources (Williams *et al.*, 1995) ;
– compréhension d'informations de santé disponibles sur les sites web à destination des patients (Berland *et al.*, 2001; Hargrave *et al.*, 2003; Kusec, 2004), et ceci en différentes langues (anglais, espagnol, français).
Ces constats peuvent avoir un impact négatif sur la communication entre les patients et les médecins, et le soins offerts aux patients (Tran *et al.*, 2009). Le contexte présenté correspond à la motivation principale 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; Iacobini, 1997; Amiot & Dal, 2005), comme exemplifié en (1). Une des particularités de ces termes est qu'ils impliquent souvent les bases venant du latin ou du grec (voir les exemples en (2) et (3)), ce qui les rends sémantiquement opaques et plus difficiles à comprendre que les mots formés avec les bases existant dans la langue française ({*anatomie*; *anatomique*}, {*livre*; *livresque*}). 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.

(2) *myocardiaque* est formé avec une base latine *myo (muscle)* et une base grecque *cardia (cœur)*

(3) *cholecystectomie* est formé avec deux bases grecques *chole (bile)* et *ectomy (ablation chirurgicale)*, et une base latine *cystis (vessie)*

Nous présentons d'abord des travaux liés de l'état de l'art (section 2) et précisons les objectifs de notre travail (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 champs de recherche 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 (François, 2011). Les mesures classiques sont essentiellement basées sur le calcul du nombre de caractères et/ou syllabes dans les mots, phrases ou documents, et sur les modèles de régression linéaire (Flesch, 1948; Gunning, 1973; Dubay, 2004). Les mesures computationnelles peuvent impliquer les modèles vectoriels et une grande variété de descripteurs et de leurs combinaisons (Wang, 2006; Zeng-Treiler *et al.*, 2007; Leroy *et al.*, 2008; François & Fairon, 2013).

2.2 Simplification lexicale

La simplification lexicale aide à rendre un texte plus facile à comprendre. Par exemple, 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, et plusieurs substitutions possibles pour ce mot 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 extrait d'un

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

corpus oral et de la Wikipédia, n-grammes de Google, WordNet (Sinha, 2012) ; longueur de mots, nombre des syllabes, information mutuelle et fréquence de mots (Jauhar & Specia, 2012) ; fréquence dans la Wikipédia, longueur de mots, n-grammes, complexité syntaxique des documents (Johannsen *et al.*, 2012) ; n-grammes, fréquence dans la Wikipédia, n-grammes de Google (Ligozat *et al.*, 2012) ; WordNet et fréquences de mots (Amoia & 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 passages difficiles (Grabar *et al.*, 2014) qui devraient être simplifiés avec les méthodes proposées plus haut par exemple.

2.3 Ressources dédiées

Des ressources spécifiques sont nécessaires pour effectuer la simplification des textes. Dans les domaines de spécialité, comme dans le domaine médical, ces ressources se présentent souvent sous forme de lexiques où les termes sont mis en correspondance avec les expressions non spécialisées correspondantes, comme dans les exemples (4) à (7). La première initiative de ce type est apparue avec le travail collaboratif Consumer Health Vocabulary (CHV) (Zeng & Tse, 2006) (exemples (4)). 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 & Sutaria, 2007), les deux ensembles de termes étant fournis par l'UMLS et donc possiblement dérivés du Consumer Health Vocabulary (exemples (7)). Des travaux similaires dans d'autres langues ont suivi. En français, l'acquisition de variations morpho-syntaxiques à partir d'un corpus comparable spécialisé et non spécialisé (Deléger & Zweigenbaum, 2008; Cartoni & Deléger, 2011) a fourni des équivalences verbe/nom (exemples (5)) et un ensemble de variations syntaxiques plus large (exemples (6)). Dans ces deux travaux, la correspondance avec les terminologies médicales n'est pas établie. Notons aussi que les travaux en acquisition de variantes terminologiques (Hahn *et al.*, 2001), de synonymes (Fernández-Silva *et al.*, 2011) et de paraphrases (Max *et al.*, 2012) sont aussi pertinents pour cette thématique de recherche.

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

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

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

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

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 recherche d'information (Lovis *et al.*, 1995; Schulz *et al.*, 1999; Hahn *et al.*, 2001) ou en traduction automatique (Loginova-Clouet & 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 & Hamon, 2006). Par exemple, dans le composé *iridochoroïdite* nous pouvons établir la relation de *localisation*, car une *inflammation* est localisée dans l'*iris* et le *choroïde*. La décomposition automatique des termes exploite souvent des méthodes à base de règles ou des approches probabilistes en corpus (McCray *et al.*, 1988; Namer, 2003; Loginova-Clouet & Daille, 2013; Claveau & Kijak, 2014).

3 Objectifs

Le travail que nous proposons est lié à plusieurs travaux de l'état de l'art : 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. De tels objectifs sont rarement poursuivis dans les travaux existants : seuls les exemples en (4) et (7) 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 la Wikipédia. Nous supposons que la Wikipédia peut contenir les paraphrases recherchées, comme dans {*myocardiaque*; *muscle du cœur*}, {*cholecystectomie*; *ablation de la vésicule biliaire*}. Par rapport à nos travaux précédents (Grabar & Hamon, 2014a,b), nous nous concentrons sur l'exploitation de la Wikipédia qui fournit des paraphrases plus riches (par rapport aux 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 évaluations plus précises des données acquises.

4 Données linguistiques

Trois types de données sont utilisés : les termes médicaux que nous voulons paraphraser (section 4.1), le corpus duquel les paraphrases sont extraites (section 4.2), et les ressources linguistiques (*i.e.* morphologie, synonymie, supplétion) 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) [2] et de la partie française d'UMLS (Lindberg *et al.*, 1993). Ces terminologies contiennent des termes syntaxiquement simples (*e.g. acrodynie*) et complexes (*e.g. infarctus du myocarde*). Nous utilisons l'ensemble des termes disponibles. Les termes syntaxiquement complexes sont segmentés en mots. Le seul filtre appliqué consiste à éliminer les 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; Jessop *et al.*, 2011). 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 la Wikipédia liés au Portail de la Médecine (version de janvier 2015). Ce corpus contient 18 434 articles (15 235 219 occurrences). Le corpus contient des informations encyclopédiques sur plusieurs notions médicales. Les contributeurs ont en général une bonne connaissance des sujets abordés. L'objectif est entre autre 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 accessibles 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 & Zweigenbaum, 2000). Ces ressources permettent de traiter l'aspect morphologique de la variation terminologique.

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.

2. Agence des Systèmes d'Information Partagés de Santé : `esante.gouv.fr/asip-sante`

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, 2003; Zweigenbaum & 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*}, {*peri*; *autour*}.

5 Méthode

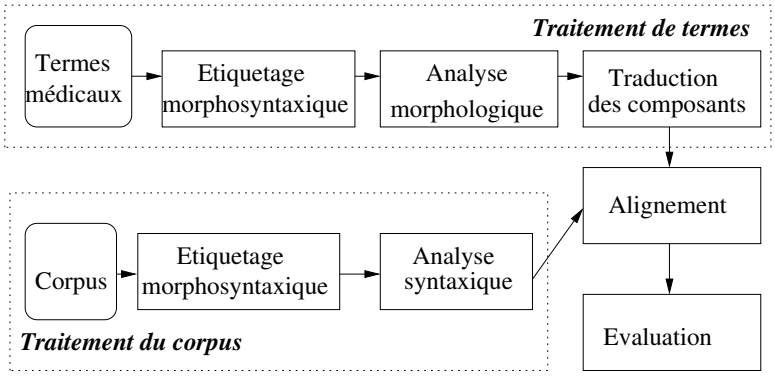

FIGURE 1 – Méthodologie générale de l'extraction de paraphrases grand public pour les termes composés.

La méthodologie est définie afin de pouvoir effectuer l'analyse des composés médicaux néoclassiques et de trouver ensuite les paraphrases correspondantes et non techniques dans les corpus. Dans certains cas, les paraphrases apparaissent dans les corpus dans des contextes définitoires (exemple (8)), auquel cas les paraphrases cooccurrent avec les termes techniques correspondant, ou bien de manière libre et sans être accompagnés de leur terme technique (exemple (9)). C'est ce deuxième type de contextes qui nous intéresse plus spécifiquement car il n'est pas contraint par l'occurrence du terme technique, et dans lequel nous pouvons en effet trouver la paraphrase *inflammation des cellules* qui correspond au terme *cellulite* dans l'exemple (9).

(8) *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.*

(9) *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 de termes (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

Pour accéder aux informations morphologiques des termes, nous effectuons trois traitements :

1. *Étiquetage morpho-syntaxique et lemmatisation des termes.* Les termes sont étiquetés morpho-syntaxiquement et lemmatisés avec `Cordial` (Laurent *et al.*, 2009). L'étiquetage morpho-syntaxique 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 termes avec leurs parties du discours (exemple (10)).

   (10) *myocardique/A*
   cholécystectomie/N
   polyneuropathie/N

acromégalie/N
galactosémie/N

2. *Analyse morphologique.* Les lemmes sont ensuite analysés morphologiquement par DériF (Namer, 2009). 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 (11).

   (11) *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 grec et n'existe pas en français moderne), DériF lui assigne la catégorie la plus probable (*e.g. N** pour les noms, *A** pour les adjectifs). Par exemple, 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 leur composants morphologiques. Nous pouvons observer que certaines bases (*e.g. galactose* et *cholécysto*) peuvent être décomposées encore plus finement, en *galact (lait)* et *ose (sucres)*, et *chole (bile/biliaire)* et *cystis (vésicule)*, respectivement. Nous considérons que les mots qui contiennent plus d'une base sont des composés. Ils sont traités lors des étapes suivantes de la méthode. Comme présenté dans les exemples en (12), Dérif fournit également des gloses pour expliquer le sens des composés analysés.

   (12) *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 obtenues suite à la décomposition sont associées avec (ou traduites en) mots du français moderne. Nous utilisons pour ceci la ressource de données supplétives présentées à la section 4.3. Les exemples en (13) présentent les données obtenues à cette étape.

   (13) *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*

   Nous pouvons voir que, suite à cette traduction, certains mots restent techniques (*e.g.*, *galactose*, *vésicule biliaire*), tandis que d'autres perdent tout leur sens technique (*e.g. mégal=grandeur*, *poly=nombreux*).

5.2 Traitement du corpus

Le corpus est traité par Cordial pour effectuer l'étiquetage morpho-syntaxique, la lemmatisation et l'analyse syntaxique. L'analyse syntaxique est utilisée pour définir les frontières des syntagmes.

5.3 Extraction de paraphrases grand public correspondant aux termes techniques

À cette étape, 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 bigrammes, les trigrammes et les quadrigrammes de syntagmes. Dans l'exemple (14), un des groupes nominaux contient les mots *muscle* et *cœur*, qui correspondent aux composants morphologiques de *myocardique* (exemple (13)). Ce groupe nominal est donc un bon candidat pour fournir une paraphrase de termes *myocarde* ou *myocardique*.

(14) *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 d'extraction de paraphrases en faisant varier quatre paramètres :
– 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,
– les ressources linguistiques pour gérer la variation terminologique. Nous avons alors trois possibilités : utilisation de formes brutes du corpus, utilisation de ressources morphologiques pour normaliser les flexions et les dérivations vers des lemmes, utilisation de la ressource de synonymes pour gérer les relations de synonymie au sein des paraphrases. Actuellement, nous n'effectuons pas la combinaison de ressources morphologiques et de synonymie,
– le taux d'alignement des termes techniques, ce qui permet de contrôler si tous les composants de ces termes sont alignés,
– le taux d'alignement des syntagmes syntaxiques, ce qui permet de contrôler si 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 (8)) sont extraites grâce aux patrons proposés dans la littérature (Péry-Woodley & 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

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

1. la paraphrase est correcte : *e.g. myocardique* paraphrasé en *muscle du cœur* ;
2. l'extraction de la paraphrase est basée sur une analyse morphologique incorrecte ({*sanglot*; *lot sang*}), la traduction vers le français n'est pas satisfaisante (*antisolaire* associé avec *sol* et *contre*), ou bien le terme traité n'est pas compositionnel et ses composants ne traduisent pas sa sémantique (*ostéodermie*, associé avec *peau* et *os* 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*) ;
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* ;
4. l'extraction est fausse et ne contient pas les informations utiles.

Ce type d'évaluation permet de calculer trois mesures :
– précision stricte $P_{stricte}$: seulement les paraphrases correctes sont considérées (cas 1) ;
– précision lâche P_{lache} : les paraphrases correctes et possiblement correctes sont considérées (cas 1 et 3) ;
– le taux d'erreurs évalue le taux d'extractions fausses (cas 4).

Les résultats sont présentés dans la section 6. Ils sont ensuite analysées 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 par la méthode proposée (section 7.4). Nous effectuons également une comparaison avec les travaux de l'état de l'art (section 7.5). 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 qui ne contiennent pas de nombres. De ces mots, 15 121 sont analysés par `Dérif` et décomposés en deux bases au moins. Ces 15 121 composés correspondent donc au matériel traité par notre méthode pour l'acquisition de paraphrases. Il est possible de distinguer 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 dans les exemples) :

$E1$: 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. Pour la tâche visée, 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.

Nombre de	*unigrammes*			*bigrammes*			*trigrammes*			*quadrigrammes*		
	b	*l*	*s*	*b*	*l*	*s*	*b*	*l*	*s*	*b*	*l*	*s*
syntagmes	9854	16093	22110	11875	18504	27670	7936	12284	19984	4701	7542	12804
termes uniques	1513	1947	2090	1780	2260	2463	1523	1966	2231	1079	1515	1922
syntagmes$_{E1}$	2681	4163	5370	1109	1611	2521	403	634	988	326	510	793
termes uniques$_{E1}$	668	1023	1051	492	670	962	239	358	472	204	297	419
syntagmes$_{E2}$	3893	6486	8876	3937	6290	9590	2154	3380	5138	1171	1947	3241
termes uniques$_{E2}$	1015	1358	1508	1025	1482	1693	752	1038	1401	517	768	1047

TABLE 1 – Résultats d'extraction de paraphrases pour les termes techniques.

Le tableau 1 présente les résultats d'extraction de paraphrases grand public. 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, bigrammes, trigrammes* et *quadrigrammes*) 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 $E2$.

Nombre de	*unigrammes*			*bigrammes*			*trigrammes*			*quadrigrammes*		
	b	*l*	*s*	*b*	*l*	*s*	*b*	*l*	*s*	*b*	*l*	*s*
syntagmes$_{E1}$	2681	4163	5370	1109	1611	2521	403	634	988	326	510	793
termes uniques$_{E1}$	668	1023	1051	492	670	962	239	358	472	204	297	419
correct	549	785	644	378	517	461	195	290	257	175	254	235
pos. correct	39	32	67	22	45	75	10	19	41	7	13	35
ttt termes	47	60	44	28	28	46	9	10	26	9	9	26
incorrect	33	146	296	64	80	380	25	39	148	13	21	123
$P_{stricte}$	82	77	61	77	77	48	82	81	55	86	86	56
P_{lache}	88	80	68	81	84	40	86	86	63	89	90	64
$\%_{incorrect}$	5	14	28	13	12	39	11	11	31	6	7	29

TABLE 2 – Évaluation de l'ensemble $E1$.

Dans le tableau 2, nous indiquons les résultats d'évaluation de l'ensemble $E1$: 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 lorsqu'aucune ressource n'est utilisée et la moins élevée avec les synonymes, où le risque de générer des alignements erronés est plus important. La précision stricte varie entre 86 et 55 %, et la précision lâche entre 90 et 40 %. Le taux d'erreurs varie entre 5 et 39 %. Au total, ces expériences fournissent 1 031 paraphrases correctes et 1 128 paraphrases correctes et possiblement correctes.

7 Discussion

7.1 Analyse morphologique des termes

La décomposition et 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 par exemple *sanglot* décomposé en *lot* et *sang* ou *exotique* décomposé en *externe* et *oreille*. Les décompositions erronées génèrent des paraphrases erronées à l'étape suivante. Ces erreurs ne sont pas comptabilisées dans les taux d'erreurs présentés dans le tableau 2.

7.2 Extraction de paraphrases et leur évaluation

Nous pouvons extraire plusieurs paraphrases correctes et intéressantes, comme celles présentées en (15).

(15) - *podalgie : douleur du pied* (termes bruts)
- *mastite : inflammation du sein* (termes bruts)
- *desmorrhexie* : *rupture des ligaments* (variation morphologique)
- *bronchite : inflammation des bronches, inflammation bronchique* (variation morphologique : bronche->bronches, bronche->bronchique)
- *dentalgie : douleurs dentaires* (variation morphologique : dents->dentaires)
- *cystoprostatectomie : ablation de la vessie et de la prostate* (termes bruts)
- *aclasie : absence de fracture* (variation de synonymie : cassure->fracture)
- *enterectomie : résection des intestins* (variation de synonymie : ablation->résection)

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 (exemples en (16)). Mais le plus souvent, les corpus fournissent les relations sémantiques correctes entre les composants. Ceci correspond à un grand avantage de la méthode basée sur l'analyse syntaxique. En effet, dans le travail précédent (Grabar & Hamon, 2014b), qui exploitait des corpus plus grands et dont la méthode d'extraction était basée 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).

(16) *hematospermie* : *le sang ou le sperme* au lieu de *le sperme dans le sang*

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 restent rares. L'augmentation de la fenêtre syntaxique permet justement d'augmenter la taille des termes paraphrasés. Il nous reste cependant à analyser l'ensemble $E2$ pour apprécier mieux l'effet de la fenêtre syntaxique. Actuellement, la longueur moyenne des termes paraphrasés varie entre 2,002 et 2,125 composants : la plupart des termes paraphrasés contiennent deux composants. De manière générale, il est difficile de calculer le rappel des résultats obtenus. Nous pensons que la manière la plus appropriée de l'évaluer est de prendre en compte le nombre de termes analysés morphologiquement. Dans ce cas, les termes paraphrasés (1 128) couvrent 7,5 % des 15 121 termes analysés morphologiquement.

L'augmentation de la couverture est une perspective importante de notre travail.

7.3 Utilisation de ressources linguistiques

Comme nous pouvons le voir dans le tableau 2, l'utilisation de ressources linguistiques complémentaires permet d'augmenter la couverture car plus de propositions sont alors extraites, par contre cela diminue la précision car les propositions risquent alors d'apporter du bruit. Nous pouvons aussi voir que l'utilisation de ressources de synonymie 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 (*i.e. cardialgie*), nous indiquons sa sémantique attendue (*douleur de 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 pré-sélectionnés et satisfaisaient le contexte. Tandis que dans notre travail, la pré-sélection des ressources n'est pas effectuée.

(17) - *cardialgie* (*douleur de cœur*) : *plaie du cœur* – {*douleur*; *plaie*}
- *cheiropathie* (*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 connue sous le terme de *mal des transports*) : *évolution du syndrome* – {*mouvement*; *évolution*} et {*maladie*; *syndrome*}

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 2 037, portant sur 1 286 termes uniques. Le patron le plus fréquent `est un` est reconnu le plus fréquemment. 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 basé 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 radiographie s 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 ces termes, nous trouvons :

1. termes composés : *achillodynie*, *clinodactylie*, *dyslexie*, *bronchodilatateur*,
2. termes affixés : *choroïde*, *amaigrissement*, *surmoi*,
3. termes morphologiquement simples : *acide*, *hypnose*, *deni*.

En relation avec la méthode d'acquisition de paraphrases de composés néoclassiques, seules les définitions pour les termes composés sont comparables directement. 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*) ou bien gardent des définitions 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 pouvons observer que les paraphrases couvrent un nombre légèrement plus important de termes. Nous nous sommes attendus à ce résultat car la méthode d'extraction de paraphrases ne requiert

pas la présence du terme analysé dans le texte, mais seulement de ses composants. Concernant la précision, elle est ainsi également plus élevée avec la méthode de paraphrase. 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. Les deux ensembles (paraphrases issues de l'ensemble $E1$ et contextes définitoires) fournissent 1 827 termes définis ou paraphrasés correctement et 2 089 termes définis ou paraphrasés correctement ou possiblement correctement.

7.5 Comparaison avec les travaux existants

Nous pouvons comparer les résultats obtenus avec ceux présentés dans trois travaux existants (Deléger & Zweigenbaum, 2008; Cartoni & Deléger, 2011; Elhadad & Sutaria, 2007) :

– *Types de termes*. Dans notre étude, nous travaillons surtout avec les termes composés, qui sont assez difficiles à comprendre par les locuteurs, et pour lesquels les paraphrases grand public apportent des informations nécessaires à leur compréhension. Dans les travaux existants (exemples (4) à (6)), seul le travail sur l'anglais (Elhadad & Sutaria, 2007) fournit des paraphrases des termes composés, tandis que les deux autres travaux (Deléger & Zweigenbaum, 2008; Cartoni & Deléger, 2011) se concentrent sur la variation morpho-syntaxique des termes ;
– *Nombre de paraphrases extraites*. Dans notre étude, nous extrayons 1 031 paraphrases correctes et 1 128 paraphrases correctes et possiblement correctes. Comme nous l'indiquons dans la section 7.4, les définitions améliorent la couverture. Dans les travaux existants, nous pouvons noter l'extraction de 65 et 82 paraphrases (Deléger & Zweigenbaum, 2008), de 109 paraphrases (Cartoni & Deléger, 2011), et de 152 paraphrases (Elhadad & Sutaria, 2007) ;
– *Précision des résultats*. Dans notre étude, les valeurs de la précision lâche varient en fonction des ressources et des fenêtres syntaxiques exploitées entre 90 et 40 %, avec une moyenne de 76 % sur l'ensemble des expériences et de 86 % pour les expériences sans l'utilisation de ressources de synonymes. Dans les travaux existants, la précision est de 67% et 60% (Deléger & Zweigenbaum, 2008), 66 % (Cartoni & Deléger, 2011), et 58 % (Elhadad & Sutaria, 2007).

Notons aussi dans les trois travaux cités, un seul (Elhadad & Sutaria, 2007) est basé sur l'exploitation de termes venant d'une terminologie existante. 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 collaborative.

Par rapport aux gloses proposées par `DériF` (Namer, 2009), en (15), nous présentons les paraphrases pour quelques termes, qui sont à comparer avec les termes glosés en (12). Nous pensons que les paraphrases extraites offrent des informations exprimées plus naturellement et sont plus faciles à comprendre. Notons cependant que, grâce au langage formel de `DériF`, tous les termes décomposés et analysés morphologiquement reçoivent une glose, alors que la couverture de paraphrases que nous extrayons dépend du contenu des corpus et des ressources linguistiques exploitées.

7.6 Termes non paraphrasés

Plusieurs termes restent non paraphrasés, comme ceux présentés en (20). Une des raisons est que certains termes, comme *hémidesmosome* ou *hémohistioblaste*, contiennent plus de deux composants, ce qui rend la détection de leurs paraphrases plus difficile. 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'extraction de paraphrases concerne l'association de composants supplétifs avec les mots du français. En effet, plusieurs traductions sont parfois possibles mais ne pouvent pas être captées avec la méthode de traduction actuelle. D'autres méthodes, comme par exemple celle proposée dans (Claveau & Kijak, 2014), devraient être exploitées pour améliorer cet aspect.

(20) *leptoméningé* : *affaibli*, *méningé*
hémipénis : *pénis*, *demi*
hémidesmosome : *corpuscule*, *demi*, *ligament*
hémohistioblaste : *cellule embryonnaire*, *tissu*, *sang*

8 Conclusion et travaux futurs

Nous avons proposé d'exploiter les articles de la Wikipédia pour détecter les paraphrases pour les termes techniques du domaine médical. Nous nous sommes concentrés sur les composés (*e.g.*, *myocardiaque, cholecystectomie, 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 (*e.g.* {*card*; *cœur*}), et leur projection sur les syntagmes syntaxiques. La méthode permet d'extraire les paraphrases correctes et possiblement correctes pour 1 128 termes composés, tandis que les définitions fournissent des explications pour 1 028 termes. Mis ensemble, cela correspond à 2 089 termes. 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 difficultés 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 & 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.

Nous voulons aussi traiter les termes complexes syntaxiquement (*e.g. 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.

Références

AMA (1999). Health literacy : report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association. *JAMA*, **281**(6), 552–7.

AMIOT D. & DAL G. (2005). Integrating combining forms into a lexeme-based morphology. In *Mediterranean Morphology Meeting (MMM5)*, p. 323–336.

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

BERLAND G., ELLIOTT M., MORALES L., ALGAZY J., KRAVITZ R., BRODER M., KANOUSE D., MUNOZ J., PUYOL J. & ET AL M. L. (2001). Health information on the internet. accessibility, quality, and readability in english ans spanish. *JAMA*, **285**(20), 2612–2621.

BOOIJ G. (2010). *Construction Morphology*. Oxford : Oxford University Press.

CARTONI B. & DELÉGER L. (2011). Découverte de patrons paraphrastiques en corpus comparable : une approche basée sur les n-grammes. In *TALN*.

CLAVEAU V. & KIJAK E. (2014). Generating and using probabilistic morphological resources for the biomedical domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 3348–3354.

CÔTÉ R. A., BROCHU L. & CABANA L. (1997). *SNOMED Internationale – Répertoire d'anatomie pathologique*. Secrétariat francophone international de nomenclature médicale, Sherbrooke, Québec.

DAILLE B. (2003). Conceptual structuring through term variations. In *Proceedings of the ACL Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, p. 9–16.

DELÉGER L. & ZWEIGENBAUM P. (2008). Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In *AMIA 2008*, p. 146–50.

DUBAY W. H. (2004). The principles of readability. *Impact Information*. Available at *http ://almacenplantillasweb.es/wp-content/uploads/2009/11/The-Principles-of-Readability.pdf*.

DUJOLS P., AUBAS P., BAYLON C. & GRÉMY F. (1991). Morphosemantic analysis and translation of medical compound terms. *Methods in Informatics and Medicin (MIM)*, **30**, 30–35.

ELHADAD N. & SUTARIA K. (2007). Mining a lexicon of technical terms and lay equivalents. In *BioNLP*, p. 49–56.

EYSENBACH G. (2007). Poverty, human development, and the role of ehealth. *J Med Internet Res*, **9**(4), e34.

FERNÁNDEZ-SILVA S., FREIXA J. & CABRÉ M. (2011). A proposed method for analysing the dynamics of cognition through term variation. *Terminology*, **17**(1), 49–73.

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

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

FRANÇOIS T. & FAIRON C. (2013). Les apports du TAL à la lisibilité du français langue étrangère. *TAL*, **54**(1), 171–202.

GRABAR N. & HAMON T. (2006). Terminology structuring through the derivational morphology. In T. SALAKOSKI, F. GINTER, S. PYYSALO & T. PAHIKKALA, Eds., *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, number 4139 in LNAI, p. 652–663 : Springer.

GRABAR N. & HAMON T. (2014a). Automatic extraction of layman names for technical medical terms. In *ICHI 2014*, Pavia, Italy.

GRABAR N. & HAMON T. (2014b). Unsupervised method for the acquisition of general language paraphrases for medical compounds. In *Computerm 2014*, Dublin, Ireland.

GRABAR N., HAMON T. & AMIOT D. (2014). Automatic diagnosis of understanding of medical words. In *Workshop on Predicting and Improving Text Readability for Target Reader Populations*, p. 11–20, Gothenburg, Sweden.

GRABAR N., VAROUTAS P., RIZAND P., LIVARTOWSKI A. & HAMON T. (2009). Automatic acquisition of synonym ressources and assessment of their impact on the enhanced search in EHRs. *Methods of Information in Medicine*, **48**(2), 149–154. PMID 19283312.

GRABAR N. & ZWEIGENBAUM P. (2000). A general method for sifting linguistic knowledge from structured terminologies. *JAMIASUP*, p. 310–314.

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

HAHN U., HONECK M., PIOTROWSKY M. & SCHULZ S. (2001). Subword segmentation - leveling out morphological variations for medical document retrieval. In *AMIA*, 229-233.

HARGRAVE D., BARTELS U., LAU L., ESQUEMBRE C. & BOUFFET E. (2003). évaluation de la qualité de l'information médicale francophone accessible au public sur internet : application aux tumeurs cérébrales de l'enfant. *Bulletin du Cancer*, **90**(7), 650–5.

IACOBINI C. (1997). Distinguishing derivational prefixes from initial combining forms. In *First mediterranean conference of morphology*, Mytilene, Island of Lesbos, Greece.

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.

JESSOP D., ADAMS S., WILLIGHAGEN E., HAWIZY L. & MURRAY-RUST P. (2011). Oscar4 : a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, **3**(41).

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.

KLINGER R., KOLÁRIK C., FLUCK J., HOFMANN-APITIUS M. & FRIEDRICH C. (2008). Detection of iupac and iupac-like chemical names. In *ISMB 2008*, p. 268–276.

KUSEC S. (2004). Les sites web relatifs au diabète, sont-ils lisibles ? *Dibète et société*, **49**(3), 46–48.

LAURENT D., NÈGRE S. & SÉGUÉLA P. (2009). Apport des cooccurrences à la correction et à l'analyse syntaxique. In *TALN*.

LEROY G., HELMREICH S., COWIE J., MILLER T. & ZHENG W. (2008). Evaluating online health information : Beyond readability formulas. In *AMIA 2008*, p. 394–8.

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.

LINDBERG D., HUMPHREYS B. & MCCRAY A. (1993). The unified medical language system. *Methods Inf Med*, **32**(4), 281–291.

LOGINOVA-CLOUET E. & DAILLE B. (2013). Segmentation multilingue des mots composés. In *TALN 2013*, p. 564–571.

LOVIS C., MICHEL P.-A., BAUD R. & SCHERRER J.-R. (1995). Word segmentation processing : a way to exponentially extend medical dictionaries. In *Medical Informatics in Europe (MIE)*, p. 28–32.

MAX A., BOUAMOR H. & VILNAT A. (2012). Generalizing sub-sentential paraphrase acquisition across original signal type of text pairs. In *EMNLP*, p. 721–31.

MCCRAY A. (2005). Promoting health literacy. *J of Am Med Infor Ass*, **12**, 152–163.

MCCRAY A. T., BROWNE A. C. & MOORE D. (1988). The semantic structure of neo-classical compounds. In *Proceedings of the Annual SCAMC*, p. 165–168.

NAMER F. (2003). Automatiser l'analyse morpho-sémantique non affixale : le système DériF. *Cahiers de Grammaire*, **28**, 31–48.

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

PACAK M. G., NORTON L. M. & DUNHAM G. S. (1980). Morphosemantic analysis of -itis forms in medical language. *Methods in Medical Informatics (MIM)*, **19**(2), 99–105.

PATEL V., BRANCH T. & AROCHA J. (2002). Errors in interpreting quantities as procedures : The case of pharmaceutical labels. *International journal of medical informatics*, **65**(3), 193–211.

PÉRY-WOODLEY M. & REBEYROLLE J. (1998). Domain and genre in sublanguage text : definitional microtexts in three corpora. In *First International Conference on Language Resources and Evaluation*, p. 987–992.

SCHULZ S., ROMACKER M., FRANZ P., ZAISS A., KLAR R. & HAHN U. (1999). Towards a multilingual morpheme thesaurus for medical free-text retrieval. In *Medical Informatics in Europe (MIE)*, p. 891–894.

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

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

TRAN T., CHEKROUD H., THIERY P. & JULIENNE A. (2009). Internet et soins : un tiers invisible dans la relation médecine/patient ? *Ethica Clinica*, **53**, 34–43.

WANG Y. (2006). Automatic recognition of text difficulty from consumers health information. In IEEE, Ed., *Computer-Based Medical Systems*, p. 131–136.

WILLIAMS M., PARKER R., BAKER D., PARIKH N., PITKIN K., COATES W. & NURSS J. (1995). Inadequate functional health literacy among patients at two public hospitals. *JAMA*, **274**(21), 1677–1682.

WOLFF S. (1987). Automatic coding of medical vocabulary. In N. SAGER, C. FRIEDMAN & M. S. LYMAN, Eds., *Medical Language Processing. Computer Management of Narrative Data*, chapter 7, p. 145–162. New-York : Addison-Wesley.

ZENG Q. & TSE T. (2006). Exploring and developing consumer health vocabularies. *JAMIA*, **13**, 24–29.

ZENG-TREILER Q., KIM H., GORYACHEV S., KESELMAN A., SLAUGTHER L. & SMITH C. (2007). Text characteristics of clinical reports and their implications for the readability of personal health records. In *MEDINFO*, p. 1117–1121, Brisbane, Australia.

ZWEIGENBAUM P. & GRABAR N. (2003). Corpus-based associations provide additional morphological variants to medical terminologies. In *AMIA*, p. 768–772.