Verbum XXXIV, 2012, nº 2, 257-276

INTERPRÉTABILITÉ ET LEXIQUE SPÉCIALISÉ

Natalia GRABAR
UMR 8163 STL CNRS & Université Lille 1 & 3

Thierry HAMON
LIMSI-CNRS & Université Paris 13

Dany AMIOT
UMR 8163 STL CNRS & Université Lille 1 & 3

RÉSUMÉ

Nous proposons d'étudier l'interprétabilité du lexique spécialisé provenant du vocabulaire biomédical. Plus particulièrement, nous mettons au jour les paramètres qui peuvent influencer cette interprétabilité. L'étude est basée sur les travaux de recherche en linguistique (proposition de paramètres et interprétation des résultats) et en Traitement Automatique de Langues (calcul des paramètres et traitement d'un gros volume de données). Nous faisons l'hypothèse que de tels paramètres peuvent être utilisés par les patients, pour accéder au sens du lexique médical, et par les acteurs de santé, pour mieux expliquer aux patients les notions médicales. Parmi les paramètres se trouvent par exemple la présence de lexèmes dans les lexiques de référence, leurs catégories syntaxiques, leur longueur et complexité morphologique, leurs fréquences, et les chaînes de caractères initiales et finales.

ABSTRACT

We propose to study the interpretability of specialized lexicon from the biomedical field. More precisely, we propose to update the parameters that can impact this interpretability. Our study is based on research in linguistics (proposal of the parameters and interpretation of the results) and in Natural Language Processing (computing of the parameters and processing of large data). According to our hypothesis, such parameters can be used by patients, for accessing the meaning of medical lexicon, and by health professionals, for providing a better explanation of medical notions to patients. Among the parameters, we can find for instance the presence of lexemes in the reference lexica, their syntactic categories, their length and morphological complexity, their frequency, and their final and initial substrings.

1. INTRODUCTION

Dans cet article, nous nous proposons de mettre au jour les paramètres qui peuvent avoir une influence sur l'interprétabilité des lexèmes des lexiques spécialisés, plus précisément ici, des lexèmes appartenant au vocabulaire biomédical (noms de maladies, de médicaments, de substances, etc.). L'intérêt porté à ce domaine trouve sa justification dans le fait que, dans leur vie quotidienne, les citoyens sont systématiquement confrontés à ce type de terminologies, que ce soit dans leur relation avec le médecin, dans les informations, entendues ou lues, dans les émissions de télévision ou de radio, les journaux ou sur la Toile, etc. Cependant, un contact fréquent avec ce type de terminologie n'en garantit pas l'intelligibilité. Comme cela a souvent été remarqué (AMA, 1999 ; McCray, 2006 ; Tran *et al*., 2009), ce vocabulaire renvoie souvent à des notions qui restent incomprises des patients alors qu'elles sont nécessaires au bon suivi des pathologies et à la réussite des processus de soins. Quant à l'intérêt de mettre au jour les paramètres influençant l'interprétabilité des lexèmes, on peut faire l'hypothèse que s'ils sont disponibles, ils peuvent être utilisés par les acteurs de santé, dans les travaux de recherche et/ou de vulgarisation, afin d'expliquer aux personnes intéressées (fondamentalement les patients ou leurs proches), des notions médicales importantes dans la compréhension des maladies et des soins proposés.

Nous avons souhaité réaliser une étude restreinte (uniquement donc sur le vocabulaire médical) mais à grande échelle, c'est pourquoi nous avons choisi de travailler à partir d'un gros corpus, *Snomed International* (Côté, 1996), une terminologie qui contient 151'104 termes spécialisés et qui a pour vocation de recenser et décrire le vocabulaire médical aussi exhaustivement que possible. Ce travail sur les données très volumineuses a été réalisé en exploitant les méthodes de Traitement Automatique des Langues (TAL).

L'article s'organise comme suit. Tout d'abord, sont présentés les paramètres linguistiques pouvant influencer l'interprétabilité du lexique biomédical (section 2), ainsi que les méthodologies et paramètres utilisés en TAL (section 3). Dans la section 4, le bilan des paramètres utilisés est effectué et les objectifs de l'étude sont précisés. La section 5 apporte des précisions sur la collecte des données et sur leurs annotations, en particulier sur la création des données de référence concernant l'interprétabilité des lexèmes. La section 6 décrit quels sont les paramètres qui ont été choisis et mis en œuvre ; les résultats et observations sur la pertinence et l'importance des paramètres étudiés sont alors présentés section 7. La conclusion est l'objet de la section 8.

2. LE CHOIX DES PARAMÈTRES LINGUISTIQUES À TESTER

Comme nous adoptons une perspective TAL et non purement linguistique, nous n'avons pas pris en compte les paramètres linguistiques

classiques pour étudier le degré d'interprétabilité des lexèmes, c'est-à-dire, parmi les plus souvent cités :

(i) le *degré de figement*, pour lequel il existe une abondante littérature ; cf. par exemple, et uniquement sur le français, Gross M. (1988), Gross G. (1996), Klein (2007), Lamiroy (2003, 2008), Mejri (2008), etc., en relation, ou non, avec des phénomènes métaphoriques (Martin 1997, Svensson 2004, Augustyn 2009), que ce soit dans une perspective diachronique (Fónagy 1997, Klein 2007) ou dans une perspective phraséologique (Nunberg, Sag & Wasow 1994, Hunston & Francis 2000, Lamiroy 2008) ;

(ii) le degré de *compositionnalité sémantique* des expressions complexes, « [l]'interprétation d'une expression complexe [étant] une fonction de l'interprétation de ses parties et de la manière dont elles sont assemblées » (Godard 2012). Là aussi, la littérature est importante ; si nous nous limitons au lexique, et notamment au lexique construit, nous pouvons par exemple citer Schreuder, Burani, & Baayen (2003), ou Ronneberger-Sibold (2006).

Dans l'approche adoptée ici, abstraction a été faite du sens pour n'utiliser que des paramètres « formels » en quelque sorte, par exemple la taille des lexèmes ou la récurrence de chaînes en fin ou en début de mot.

Comme nous le verrons, les raisons qui peuvent entraver l'interprétation des termes du vocabulaire médical sont complexes, multifactorielles, et interagissent les unes avec les autres ; elles nous semblent cependant, d'un point de vue linguistique, être principalement de deux types, externes ou internes.

2.1. Les causes externes

Celles-ci peuvent concerner :

(i) *l'origine des termes*. Ceux-ci sont en effet très souvent issus des langues classiques, du latin (*cardia, sagittal, endoscopique*) ou du grec, que ce soit dans des lexèmes dérivés (*hépatite*), composés (*thrombophlébite*), ou construits par les deux types de procédés (*trachéodynie*) ; ils sont souvent ininterprétables pour un non spécialiste.

(ii) *la catégorie*. Il arrive que des noms propres entrent dans des unités lexicales complexes dénotant par exemple des pathologies, comme dans *maladie de Charcot* ou *syndrome de Stockholm*. Le nom propre ne nous donne aucune indication sur la maladie elle-même, mais uniquement sur des circonstances extérieures, à savoir ici le découvreur (*Charcot*) ou le lieu de l'événement qui a permis d'identifier la pathologie (*Stockholm*).

(iii) *la non intégrité des lexèmes*, dans les abréviations et les sigles, par exemple *Cu, EnaKt, Fab, Cr, Hbc, ADN*. Dans le lexique spécialisé, comme d'ailleurs dans le lexique général, ceux-ci sont généralement indécodables sans connaissances particulières.

2.2. Les causes internes

Le second type de causes est lié à la complexité structurelle des termes eux-mêmes, ce qui, nous en faisons l'hypothèse, peut être une entrave à l'interprétabilité, que l'unité complexe soit polysegmentale ou non[1].

• Les unités polysegmentales que l'on trouve dans la terminologie médicale sont de plusieurs types : elles peuvent instancier les structures classiques que l'on trouve dans la langue générale, par exemple [N1 prép N2] (*syndrome de Beyrouth*), [NA] (*cystite chronique, acide aminé, échinococcose multiloculaire*), mais elles peuvent aussi manifester une plus grande complexité, ainsi par exemple *thrombophlébite du sinus sagittal supérieur, dermite psoriasiforme spongiotique, pyodermite simulant une blastomycose* ou encore *fibrofolliculomes multiples,* que la terminologie *Snomed International* recense comme unités lexicales.

• Les unités monosegmentales complexes regroupent les lexèmes dérivés et composés.

– Les dérivés regroupent bien sûr les suffixés (*cathétérisation, basaloïde*), les préfixés (*prédiabète, pseudarthrose*), ou des lexèmes ayant subi plusieurs opérations de dérivation (*multiloculaire, catathymie*). Ces quelques exemples suffisent à montrer que ces lexèmes sont très fréquemment construits sur base néoclassique. Par ailleurs, certains suffixes n'apparaissent que dans des lexèmes des langues de spécialité, ce qui ne facilite là non plus pas nécessairement l'interprétation des lexèmes construits ; en voici deux exemples (sur le sujet, cf. Corbin & Paul 2000) : *-ite* (*otite, phlébite, dermatite*, etc.)[2], *-ase* (*lithiase, lyase, coagulase*, etc.).

– Les composés du domaine médical sont principalement des composés néoclassiques[3] ; ceux-ci nous semblent être particulièrement source de difficultés interprétatives dans la mesure où (i) ils mettent en jeu des lexèmes d'origine grecque ou latine (*cérébro-spinal, ammoniophanérèse*), (ii) leur ordre de lecture se fait de droite à gauche (une arthroplastie par exemple est une plastie 'opération chirurgicale' (constituant droit) concernant une arthr(o) 'articulation' (constituant gauche)) alors que celui des composés standards se fait de gauche à droite (un pneu-neige est un pneu (constituant gauche) utilisé lorsqu'il y a de la neige (constituant droit)) ; par ailleurs, (iii), les composés néoclassiques du domaine médical peuvent comporter non seulement deux constituants (par ex. *abdominoplastie*), comme les composés

---

1 Comme le critère graphique sera extrêmement important par la suite, nous distinguons ce que nous appelons ici des unités polysegmentales, composées graphiquement de plusieurs séquences, et des unités monosegmentales, constituées graphiquement d'un seul segment.

2 Le suffixe *-ite* est cependant en train d'entrer dans la langue générale, par exemple *réunionite* ; sur ce sujet, cf. Pacak *et al*., 1980 ; Manuila *et al*., 2001.

3 Sur la composition néoclassique, cf. par ex. Amiot & Dal 2005 ; Bauer 2009 ; Bisetto & Scalise 2009 ; Dal & Amiot 2008 ; Iacobini 2011 ; Lüdeling 2006 ; Lüdeling, Schmidt & Kiokpasoglou 2002 ; Namer & Villoing 2005.

standards), mais aussi trois (*aérodontalgie* ‘Douleur (*algie*) dentaire (*dont-*) provoquée par les changements de pression atmosphérique (*aéro*), quatre (*uvulo-palato-pharyngoplastie* qui est une opération (*plastie*) qui se fait par ablation partielle de la luette (*uvul(o)*), du voile du palais (*palat(o)*), et des amygdales (*pharyng(o)*), et même davantage (*canaliculo-dacryocystorhino-stomie*). Dans le vocabulaire médical, il est très fréquent de trouver des composés néoclassiques extrêmement complexes ; sur la composition néoclassique dans le vocabulaire médical, cf. Namer (2005a, 2005b), Namer & Baud (2005), Namer & Villoing (2005).

3. MÉTHODOLOGIE EN TAL ET OBJECTIFS DE L’ÉTUDE

En TAL et dans les travaux liés à la terminologie, l’identification automatique de la complexité sémantique des termes est fondamentale, elle reste cependant souvent implicite (Wüster 1981, Cabré & Estopà 2000, Cabré 2002). De manière générale, c’est la spécificité des termes associés à un domaine de spécialité qui est étudiée, la possibilité de leur interprétation étant alors tout au plus une problématique secondaire.

Le plus souvent, dans les approches proposées, il s’agit de cerner le degré de spécialisation du terme, considéré généralement en fonction de son adéquation au domaine, et ceci en vue de filtrer les termes extraits de corpus spécialisés (Korkontzelos & al. 2008). Pour cela, les approches proposées exploitent les formes de surface des termes, comme par exemple, la présence et la spécificité de mots pivots (Drouin 2002, Drouin & Langlais 2006), le voisinage du terme dans le corpus, la diversité de ses composants par le biais de mesures statistiques comme la *C-Value* ou le *PageRank* (Daille 1995, Frantzi *et al*. 1997, Maynard & Ananiadou 2000), ou de divers critères statistiques (Drouin 2002). Un autre moyen consiste à vérifier si les unités lexicales sont encodées dans les terminologies de référence ; si tel est le cas, ces unités sont alors considérées comme ayant un sens spécialisé (Elhadad & Sutaria 2007). Finalement, l’application de mesures de lisibilité et d’évaluation de la complexité des mots en comptabilisant le nombre de lettres ou de syllabes est largement utilisée dans les pays scandinaves et aux États-Unis (Flesch 1948, Gunning 1973, Bjornsson 1979).

Notons que, même si la linguistique et le TAL se focalisent sur des paramètres qui leur sont propres, il est possible de faire le lien entre les deux ensembles de paramètres ; ainsi, par exemple, la mesure de lisibilité peut permettre de calculer la complexité morphologique (ou syntaxique) des lexèmes ou des termes.

Nous proposons de prendre en compte les différents paramètres mentionnés et de les exploiter avec les méthodes propres au TAL afin de pouvoir traiter de gros volumes de données. Dans cette première étude, nous travaillons avec les paramètres positionnés au niveau des lexèmes, ce qui signifie que nous ne prenons pas en considération les paramètres positionnés au

niveau des constituants (fréquence des constituants dans la langue générale ou dans le domaine de spécialité étudié, fréquence en tant que constituant droite / gauche, origine des constituants, etc.). D'autres paramètres seront mentionnés ultérieurement (voir section 6).

Nous faisons par ailleurs l'hypothèse que la modélisation des phénomènes associés à certains facteurs décrits ci-dessus, peut être mise en œuvre dans des approches exploitant par exemple la structure interne des lexèmes composés complexes, des mesures de lisibilité (Flesch 1948, Gunning 1973, Bjornsson 1979), le nombre d'affixes et de constituants ainsi que différents critères statistiques associées aux lexèmes. Dans cette démarche, nous nous appuierons, entre autre, sur les travaux de décomposition des composés en leurs constituants (Dujols 1991, Lovis 1995, Hahn 2001) et l'analyseur morphologique Dérif (Namer & Zweigenbaum 2004, Namer 2009) pour traiter morphologiquement des lexèmes extraits d'une terminologie en vue de juger de leur degré de complexité et/ou d'interprétabilité.

Dans la suite de ce travail, nous présentons d'abord les données avec lesquelles nous travaillons, puis la manière dont nous les traitons, afin de déterminer (i) si les lexèmes sont interprétables pour un locuteur non spécialiste et (ii) quels sont les paramètres les plus importants intervenant dans cette interprétation.

4. PRÉPARATION DES DONNÉES

4.1. Sélection des données

Comme nous l'avons signalé dans l'introduction, les données sur lesquelles nous travaillons ont été obtenues à partir de la terminologie médicale *Snomed International*. Les 151'104 termes spécialisés qu'elle contient sont répartis en 11 axes sémantiques, par exemple « maladies » et « anomalies » (*exanthème vésiculobulleux, dextroversion, thrombophlébite du sinus sagittal supérieur, hépatomégalie congénitale*), « procédures médicales » (*acte microchirurgical, tumorectomie, herniorrhaphie*), « produits chimiques » et « médicaments » (*bactéricide, isocyanate de méthyle, aspirine d'aluminium*), « fonctions de l'organisme » (*épidermopoïèse anormale, rigidité mésencéphalique, métatarsalgie*), « organismes vivants » (*bactérie, Clostridium thermoautotrophicum, Aspergillus deflectus*), « statut social » (*belle-mère, parent diabétique, donneur de sang, condition de vie*) ou « anatomie » (*pannicule adipeux, mésobronche aviaire, veine stylomastoïdienne*). Pour notre étude, nous avons retenu les termes des cinq axes que nous avons considéré comme étant majeurs pour le domaine médical : « maladies », « anomalies », « fonctions », « procédures » et « anatomie ». Ce sont en effet ceux qui intéressent le plus les personnes non spécialistes qui consultent les sites dédiés au domaine médical dans leurs recherches d'informations sur des maladies elles-mêmes, ou sur les traitements et les interventions susceptibles de les enrayer. Cette première sélection nous permet d'obtenir

un ensemble de 104’649 termes, qui sont ensuite segmentés en mots graphiques.

Par ailleurs, les termes complexes recensés dans la terminologie sont ensuite segmentés en autant de mots graphiques. Ainsi, une expression complexe comme *thrombophlébite du sinus sagittal supérieur* a-t-elle été segmentée en cinq mots. Précisons que ce critère graphique, qui est un critère négatif, est fondamental pour le repérage des composés néoclassiques. En effet, lorsqu’il est associé à d’autres paramètres (notamment le nombre de constituants ; cf. les différents paramètres sous 5), il permet d’éliminer les unités polysegmentales en tant que telles pour ne garder que les unités monosegmentales : les lexèmes non construits (au moins en synchronie), que ce soit des noms communs ou des noms propres, les lexèmes construits par dérivation et composition néoclassique, les emprunts, les abréviations et les sigles.

Une fois la segmentation terminée, nous avons obtenu au total 29’641 mots, auxquels nous avons appliqué un ensemble de règles pour détecter et supprimer les marques flexionnelles, ce qui a fourni une première lemmatisation automatique, qui a dû ensuite être vérifiée, complétée et corrigée manuellement.

L’ensemble de ces lexèmes (29’641) est supposé représenter des notions médicales fondamentales. Comme nous voulions effectuer une étude aussi exhaustive que possible, nous avons gardé l’ensemble de ces lexèmes pour la suite de l’étude.

4.2. Étiquetage et annotation

Nous avons assigné à chacun de ces lexèmes une catégorie syntaxique, comme par exemple Nom, Adjectif, Verbe, Adverbe, Préposition ou Déterminant. Lorsque cela était pertinent, i.e. pour les noms propres, les emprunts et les abréviations[4], nous avons assigné une annotation spécifique. L’assignation de catégories syntaxiques était corrigée manuellement parce que les outils automatiques d’annotation morphosyntaxiques peuvent difficilement traiter ce type de matériel pour les raisons suivantes : (1) les lexèmes traités étaient considérés hors contexte, ce qui rendait leur étiquetage automatique difficile ; (2) il s’agissait de lexèmes spécifiques du domaine médical, souvent absents des dictionnaires classiques ; (3) les catégories choisies pour l’annotation étaient d’ordre syntaxique mais aussi sémantique et n’étaient donc pas couvertes par les outils automatiques. Par ailleurs, la lemmatisation effectuée a été vérifiée et complétée afin d’obtenir la forme citationnelle de ces mots (singulier pour les noms, singulier masculin pour les adjectifs, infinitif pour les verbes). La suite du travail a été effectuée sur les lemmes.

---

4 Dans *SI*, la catégorie « Abréviation » regroupe les abréviations et les sigles.

Ces données ont aussi été annotées selon les objectifs de notre étude. Trois locuteurs, entre 25 et 40 ans, sans aucune formation médicale ont été impliqués dans cette annotation. Les annotateurs devaient effectuer le travail de manière individuelle, en utilisant leurs connaissances et intuitions linguistiques de locuteur, la consultation de dictionnaires étant bien entendu interdite. A chaque locuteur, il était demandé d'analyser l'ensemble de lexèmes de la liste. Pour chaque lexème, l'un des trois jugements était possible :

1) je comprends le lexème
2) j'ai une vague idée de la signification du lexème
3) je ne comprends pas le lexème

1) regroupe ainsi les lexèmes connus et/ou interprétables, sous 3) figurent les lexèmes inconnus et/ou non interprétables. Sous 2) en revanche peuvent être regroupés différents types de lexèmes : des lexèmes dont l'annotateur a un souvenir plus ou moins vague, par exemple parce qu'il a l'impression de les avoir déjà vus / entendus / lus, des lexèmes que l'annotateur a la possibilité de décomposer et d'interpréter sans toutefois y parvenir réellement, etc. Nous avons cependant préféré proposer une seule possibilité pour l'annotation de ces types de situations, car nous nous sommes rendu compte dans un travail précédent (Grabar, 2013) que la proposition d'une grille comportant des catégories trop proches pouvait introduire de la confusion chez les annotateurs et donc dans les annotations.

La distinction dans les jugements peut être faite au niveau des annotations individuelles mais aussi en prenant en compte l'ensemble des annotations. Nous expliciterons ce point dans la section 5.

Comme le nombre de lexèmes retenus est très important (presque 30'000 entrées) et que la période d'annotation a duré un certain temps (1 à 2 mois, selon les annotateurs), il est possible que les annotations présentent des inconsistances intra-annotateur, celles-ci pouvant être dues à plusieurs raisons : l'instabilité dans le temps, l'influence de la fatigue, l'effet d'apprentissage, ou l'évolution des connaissances médicales elles-mêmes.

Les annotations effectuées manuellement correspondent aux données de référence par rapport auxquelles nous avons étudié les paramètres qui jouent un rôle dans la détection automatique de l'interprétabilité sémantique, ceci avec les outils que nous offre le TAL.

5. RECENSEMENT DES PARAMÈTRES ET ANALYSE AUTOMATIQUE DE L'INTERPRÉTABILITÉ DES LEXÈMES

Les données linguistiques ont ensuite été traitées automatiquement avec des méthodes du TAL afin d'étudier l'importance relative des paramètres (i) pour distinguer les lexèmes interprétables des lexèmes non interprétables, et (ii) pour étudier quels sont les paramètres les plus saillants pour cette

distinction. L'approche que nous proposons se compose de deux types d'éléments : les paramètres des lexèmes étudiés et une approche dite par apprentissage automatique. Nous ne présenterons que le premier dans le cadre de cet article, mais ces deux éléments sont utilisés de manière à ce qu'ils soient indicatifs de l'interprétabilité des lexèmes.

Nous présentons ici les paramètres que nous avons choisi d'exploiter, de même que leur motivation vis-à-vis des objectifs poursuivis. Les paramètres recensés dans l'état de l'art (sections 2 et 3) ont été enrichis : nous avons défini 24 paramètres au total, calculés automatiquement, sauf la correction de catégories syntaxiques. Comme nous l'avons mentionné supra, dans le choix des paramètres, nous avons privilégié ceux qui se positionnent au niveau des lexèmes et non ceux qui se positionnent au niveau des constituants, bien que ceux-ci aient aussi été calculés et qu'ils soient déjà disponibles. Il s'agit en effet de deux niveaux différents, celui du lexème et celui de ses constituants morphologiques.

Les paramètres utilisés peuvent être regroupés en 10 classes suivantes :

1) *Catégories syntaxiques*. Les catégories syntaxiques sont exploitées afin de faire la distinction entre les lexèmes empruntés (souvent des mots latins ou grecs ; cf. supra sous 5.), les abréviations et les noms propres d'une part, et les lexèmes appartenant aux autres catégories. Par ailleurs, parmi les autres catégories, nous supposons que l'opacité concerne davantage les noms et les adjectifs, qui correspondent généralement aux notions médicales, que les verbes, les adverbes (et éventuellement les autres catégories syntaxiques), qui appartiennent souvent à la langue générale, et qui donc devraient être plus familiers aux locuteurs.

2) *Présence des lexèmes dans les lexiques de référence*. Nous exploitons deux lexiques de référence du français : le *Trésor de la Langue Française informatisé* (*TLFi*) et le *Lexique.org* (lexique.org). Le *TLFi* est un dictionnaire de la langue française des XIX^{e} et XX^{e} siècles, il contient environ 100 000 entrées. Le *Lexique.org* est un lexique créé pour les travaux en psycholinguistique, il contient environ 135'000 entrées, dont les formes fléchies des verbes, adjectifs et noms, ce qui correspond à environ 35'000 lemmes. Le contenu de ces deux lexiques dépasse, et de loin, la compétence lexicale commune des locuteurs ; de ce fait, nous nous attendons à ce que les lexèmes qui ne figurent ni dans l'un ni dans l'autre soient des termes très spécialisés, potentiellement inconnus d'un locuteur non spécialiste, et donc sources d'opacité sémantique.

3) *Fréquence des lexèmes attestée sur un moteur de recherche généraliste*. Pour chaque lexème, nous interrogeons le moteur de recherche Google afin de connaître sa fréquence sur la Toile. Les lexèmes fréquents sont supposés être mieux connus par les locuteurs, et donc plus facilement interprétables. Comme nous le verrons plus loin, les fréquences varient énormément en fonction des lexèmes : elles peuvent aller de plusieurs centaines de millions d'occurrences (230'000'000 pour *activation*, 175'000'000 pour

drain) à quelques milliers (9'394 pour *cholangiocholangiostomie,* 4'820 pour *flagellantisme*) ou dizaines (10 pour *cholécystocolostomie*).

4) *Fréquence des lexèmes dans la terminologie*. Nous calculons aussi la fréquence des lexèmes dans la terminologie médicale *Snomed International*. De manière similaire, nous supposons que les lexèmes qui apparaissent dans un nombre élevé de termes, peuvent être plus facilement compris par les locuteurs.

5) *Nombre et types de catégories sémantiques associées aux lexèmes* dans la terminologie *Snomed International.* Comme nous l'avons indiqué, la terminologie *Snomed International* est organisée en 11 catégories sémantiques, dont nous étudions les termes provenant de cinq catégories. Par contre, nous calculons les fréquences de lexèmes composant ces termes dans l'ensemble de la terminologie. Nous proposons d'exploiter cette information, car il est possible que les lexèmes appartenant à plusieurs catégories sémantiques, correspondent aussi à des notions assez fondamentales de médecine. De tels lexèmes seraient alors susceptibles d'être mieux connus et compris des locuteurs, sauf lorsqu'il s'agit de lexèmes polysémiques, qui ne sont pas traités ici. Par ailleurs, il est possible que les termes de certaines catégories de la *Snomed International* contiennent davantage de lexèmes interprétables que d'autres.

6) *Longueur des lexèmes en nombre de caractères et de syllabes*. Pour chaque lexème, nous calculons le nombre de caractères et de syllabes graphiques. Nous nous attendons à ce que les lexèmes plus longs soient potentiellement plus difficiles à comprendre, car ils peuvent correspondre à des lexèmes relativement complexes, notamment à des composés néoclassiques.

7) *Nombre de bases et d'affixes*. Chaque lexème est analysé par l'analyseur morphologique Dérif (Namer 2009), adapté au traitement de lexèmes médicaux, qui en effectue une décomposition en bases et affixes connus. Nous exploitons cette information aussi, car les lexèmes décomposables en plusieurs bases morphologiques correspondent souvent, dans la terminologie médicale, à des composés néoclassiques.

8) *Chaînes de caractères initiales et finales des lexèmes*. Pour chaque lexème, nous calculons les chaînes initiales et finales allant de trois à cinq caractères graphiques. Nous supposons que ces chaînes de caractères peuvent être évocatrices de bases récurrentes ou éventuellement d'affixes, préfixes ou suffixes. La motivation principale est essentiellement liée aux chaînes de caractères en fin de lexèmes, qui peuvent alors correspondre à des constituants sémantiquement recteurs de composés néoclassiques.

9) *Nombre et pourcentage de consonnes, de voyelles et d'autres caractères*. Nous calculons également le nombre et le pourcentage de consonnes, de voyelles et de signes diacritiques (par exemple, les tirets, les apostrophe ou les virgules dans les noms de produits chimiques), car de tels signes peuvent être des indices de complexité potentielle.

10) *Scores de lisibilité des lexèmes*. Nous appliquons deux mesures de lisibilité : Flesch (Flesch, 1948) et sa variante Flesch-Kincaid (Kincaid *et al*., 1975). De telles mesures sont typiquement utilisées pour évaluer le niveau de difficulté d'un texte en fonction du niveau d'études d'un locuteur (études secondaires, supérieures...). Très populaires aux États-Unis et dans les pays Scandinaves, ces mesures sont peu connues et utilisées ailleurs en Europe et en France. Typiquement, ces mesures exploitent les indices de surface des lexèmes (nombre de caractères et/ou de syllabes) et cherchent à les normaliser avec des coefficients adaptés en fonction de paramètres sociolinguistiques. Nous proposons d'exploiter ces mesures pour voir si, de la même manière que d'autres paramètres (nombre de bases et d'affixes, longueur du lexèmes, etc.), elles peuvent constituer un paramètre supplémentaire, non linguistique, pour mesurer l'interprétabilité des lexèmes des vocabulaires spécialisés.

Comme nous pouvons le voir, nous avons essayé de « traduire » les spécificités des termes retenus, notamment les néoclassiques, en des paramètres linguistiques et extralinguistiques calculables automatiquement. Cet aspect (calcul automatique possible) est en effet important pour traiter un gros volume de données linguistiques (presque 30'000 lexèmes dans notre étude).

6. RÉSULTATS ET OBSERVATIONS

Nous présentons et analysons dans cette section les résultats obtenus selon différents points de vue : la création des données de référence et l'accord inter-annotateur, les résultats globaux obtenus automatiquement et la pertinence des paramètres, l'analyse détaillée de l'influence et de la pertinence des paramètres.

6.1. Annotations et accord inter-annotateur

	Annot. 1	Annot. 2	Annot. 3	Unanimité	Majorité
1) Je comprends le lexème	8 099	8 625	7 529	5 960	7 655
2) J'ai une vague idée de la signification du lexème	1 895	1 062	1 431	61	597
3) Je ne comprends pas le lexème	19 647	19 954	20 681	16 904	20 511
Total	29 641	29 641	29 641	22 925	28 763

Tableau 1.– Résultats des annotations, différents jeux de données (par annotateur, unanimité et majorité)

Dans le tableau 1, nous présentons les résultats des annotations des trois annotateurs ; visiblement, ceux-ci ne sont pas totalement convergents : alors que les annotateurs 1 et 2 produisent des jugements assez similaires sur la possibilité, ou non, d'interpréter les lexèmes qui leurs sont présentés, le troisième se démarque sensiblement : par rapport aux deux autres, il y a environ 1'000 lexèmes supplémentaires qu'il a déclaré ne pas être en mesure d'interpréter. De manière générale cependant, les lexèmes ininterprétables sont majoritaires et représentent plus de deux tiers des données et ceci quel que soit l'annotateur. L'accord inter-annotateur, calculé avec le Kappa de Fleiss (Fleiss & Cohen, 1973), est bon, avec un kappa de 0.73. Ceci correspond en effet à un niveau d'accord élevé, surtout pour un travail sur des données linguistiques, où les accords entre les locuteurs sont difficiles à obtenir.

En nous fondant sur les mêmes annotations fournies par les trois annotateurs, nous avons aussi calculé les ensembles sur lesquels tous les annotateurs sont d'accord entre eux (*unanimité*). De la même manière, nous avons calculé un autre sous-ensemble pour lequel les annotations ont un avis majoritaire *(majorité)*. Par définition, ces deux sous-ensembles montrent moins d'ambiguïté d'annotations car les ambiguïtés potentielles sont « résolues ».

6.2. Observations générales sur les résultats et la pertinence des paramètres

L'exploitation de traitements automatiques a permis d'essayer de reproduire les données de référence grâce aux paramètres utilisés. Le succès de cette reproduction peut être mesurée : plus les valeurs sont proches de 1, plus la reproduction automatique est performante et plus les paramètres utilisés sont saillants.

	Annot. 1	Annot. 2	Annot. 3	Unanimité	Majorité
Précision	0.799	0.807	0.835	0.947	0.878
Rappel	0.828	0.825	0.861	0.950	0.894
F-mesure	0.810	0.813	0.846	0.948	0.885

Tableau 2. – Résultats de la reproduction automatique des annotations de référence, effectuée sur les cinq jeux de données (un jeu par annotateur, le jeu *unanimité* et le jeu *majorité)*

Dans le tableau 2, nous présentons les résultats de la reproduction automatique des données de référence. La réussite des résultats est exprimée avec les trois mesures classiques d'évaluation : la précision (ou l'exactitude), le rappel (ou la complétude) et la F-mesure (la moyenne harmonique de la précision et du rappel). Nous pouvons voir que, parmi les trois annotateurs, les annotations du troisième sont les plus faciles de reproduire (gain de 0.030 pour la F-mesure). Les résultats deviennent meilleurs avec l'ensemble

majorité, et atteignent 0.948 sur l'ensemble *unanimité*. Comme nous nous y attendions, ces deux ensembles (*majorité* et *unanimité)* présentent un accord plus important ; il y a moins d'ambiguïté dans les annotations, il est donc plus facile de détecter les régularités spécifiques à ces données linguistiques. Ces résultats globaux, très élevés, indiquent aussi la pertinence globale des paramètres utilisés. Nous allons maintenant analyser la pertinence des paramètres de manière plus détaillée.

6.3. Analyse détaillée de l'importance des paramètres

Les résultats montrent que plusieurs paramètres interagissent et aucun ne semble pouvoir jouer, à lui seul, un rôle explicatif dans la possibilité, ou non, d'interpréter un lexème. Le paramètre le plus efficace (i.e. la présence des lexèmes dans la ressource lexique.org) offre des résultats inférieurs à ceux obtenus avec différentes combinaisons de paramètres. Il est tout de même important de noter que l'un des descripteurs semble ne jouer aucun rôle, les tests de lisibilité de Flesch et de Flesch-Kincaid : les scores ne sont pas utilisés par le système automatique pour la détection de l'interprétabilité, ce qui est une observation inattendue car ce type de scores est utilisé assez communément pour statuer sur la difficulté des documents et de leur niveau de compréhension. Il est possible que les informations encodées par ce paramètre soient déjà représentées par d'autres paramètres (la taille des lexèmes en particulier) ou bien qu'il soit réellement non pertinent pour évaluer le caractère plus ou moins interprétable des lexèmes. Ce que l'on constate en revanche, c'est que les lexèmes qui reçoivent systématiquement la réponse 1) (cf. tableau 1) de la part des trois annotateurs possèdent les propriétés suivantes (comme précédemment, nous distinguerons les propriétés externes et internes[5]).

Propriétés externes :

• les lexèmes sont présents dans les deux lexiques de référence, ce qui est le cas, par exemple, de lexèmes comme *abrasion, anémie* ou *cellulite*. Ce paramètre est très efficace et en général il existe une bonne cohésion entre la présence dans les lexiques de référence et la catégorisation par les annotateurs. Quelques exemples intéressants concernent les termes assez complexes, mais qui apparaissent dans les lexiques de référence et qui sont également connus par les annotateurs (*e.g. septicémie, prostatectomie, lombalgie, aménorrhée*). Lors du traitement automatique, ce paramètre ajoute aussi quelques points à l'efficacité du traitement ;

---

5 Alors que les propriétés internes concernent la manière dont les lexèmes sont formés (leur taille, s'ils possèdent des chaînes de caractères récurrentes, à droite ou à gauche, etc.), les propriétés externes renvoient à leur statut dans la langue (leur degré de lexicalisation par exemple).

• ils appartiennent assez fréquemment à la catégorie sémantique de *Snomed International* concernant les fonctions de l'organisme. Il s'agit par exemple de lexèmes comme *anabolisme, enfantement, articulatoire, croissance, déni.* Cet ensemble de lexèmes semble couvrir des notions médicales assez communes dans la langue ;

• ils sont généralement fréquents, mais le critère de la fréquence est le seul critère qui soit purement relatif, *i.e.* qu'il ne peut s'interpréter qu'à la lumière des autres critères ; c'est ce que nous montrent les quelques exemples que nous allons commenter ici. Ainsi, alors que des lexèmes comme *coccyx* ou *drain* ont de fortes fréquences (1'800'000 et 175'000'000, respectivement) et sont en effet compris des annotateurs, il existe aussi des lexèmes comme *colique* ou *clitoridien* montrant quant à eux des fréquences beaucoup plus faibles (respectivement 807'000 et 9'821), et qui cependant sont compris des trois annotateurs. En revanche, certains lexèmes aux fréquences importantes, par exemple *coagulase*, *clivage* ou *douve* (655'000, 1'350'000 et 1'030'000, respectivement) n'ont pas été interprétés par les annotateurs. Si la fréquence est bien un paramètre important, elle ne montre cependant son potentiel et son efficacité qu'en combinaison avec d'autres paramètres.

Propriétés internes :

• les lexèmes appartiennent aux catégories du verbe, de l'adjectif et de l'adverbe. L'hypothèse d'une différence d'interprétabilité entre ces catégories d'une part, les abréviations, les emprunts et les noms propres d'autre part est donc confirmée : les premières sont sémantiquement plus opaques que les secondes. Notons par ailleurs que les noms ne représentent pas une catégorie homogène quant à leur interprétabilité. Si ce paramètre était utilisé seul, il pourrait fournir des résultats d'une performance assez élevée entre 0.68 et 0.7 ;

• ils ne doivent être ni trop courts (pas moins de quatre ou six caractères, contrairement aux abréviations), ni trop longs (il s'agit dans ce cas généralement de composés néoclassiques, notamment lorsqu'ils possèdent plus de deux constituants). Ceci transparaît quel que soit le paramètre en cause : nombre de caractères ou de syllabes, nombre de voyelles ou de consonnes, nombre de bases et d'affixes. Notons que l'efficacité de ces paramètres ressort surtout lorsqu'ils sont combinés entre eux ou avec d'autres (présence dans le lexique de référence, chaîne finale des lexèmes, catégorie grammaticale...) ;

• ils possèdent des chaînes finales récurrentes de trois ou quatre caractères, qui peuvent être associées à des suffixes ou à ces constituants recteurs de lexèmes néoclassiques ; parmi celles-ci on trouve par exemple : *-lgie* (de *-algie*), *-stie* (de *plastie*), *-omie* (de *-ectomie*). En revanche, les chaînes finales de cinq caractères brouillent les résultats (certainement parce qu'elles suppriment la généralisation qui a été apportée par les chaînes de caractères

plus courtes) et les chaînes de caractères initiales sont très peu exploitées[6], elles apportent cependant un gain pour l'efficacité du traitement automatique.

Les lexèmes qui en revanche n'ont été interprétés par aucun des annotateurs possèdent souvent les propriétés inverses :

• ils ne sont référencés ni dans le *TLFi*, ni dans *lexique.org* ; c'est le cas par exemple de *dactylolyse, cestodose, laminotomie,* ou *galactosurie.* L'association entre leur absence des lexiques de référence et l'impossibilité de les interpréter est souvent très importante ;

• ils peuvent appartenir à une ou plusieurs catégories sémantiques de la terminologie *Snomed International.* Il s'agit typiquement des pathologies, procédures, anatomie et fonctions (avec des fréquences plus faibles). Nous pouvons ainsi remarquer que dans ces domaines médicaux, les termes sont moins interprétables que dans d'autres, et cela peut être dû à leurs propriétés internes. Par exemple, les termes relatifs aux pathologies et procédures médicales sont souvent très complexes, et formés par composition néoclassique, ce qui les rend plus difficiles à interpréter. Ce critère nous sera utile dans la poursuite de nos analyses ;

• ils appartiennent à des catégories non canoniques comme les sigles (*CaOH, PTHrP, NADp+, Cu, Hb*) ou les abréviations (*Cys-arg, MétHb, Micro-ID, Gln-glu*) ;

• ce sont généralement des lexèmes peu fréquents (*acétylase* avec 5 706 occurrences, *cheilotomie* avec 9 272 occurrences, *cruralgie* avec 10 672 occurrences), mais cf. ce qui a été dit sur la relativité de ce critère *supra*.

Ces analyses indiquent que quelques paramètres, comme la catégorie syntaxique ou la présence dans les lexiques de référence, sont assez fiables pour statuer sur l'interprétabilité des lexèmes, même lorsque ces paramètres sont pris de manière isolée. Quant aux autres paramètres, bien qu'étant efficaces, cette efficacité est moindre et elle n'apparaît souvent que lorsque les paramètres sont combinés.

Avant de conclure, nous voudrions faire une dernière remarque, relative au traitement automatique et à une généralisation possible : ces différents paramètres sont présents (ou absents) de manière homogène dans les différents jeux de données testés (trois annotateurs, *unanimité* et *majorité*). En effet, plusieurs paramètres et leurs combinaisons ressortent de manière très stable : le système automatique les associe régulièrement à l'interprétabilité ou à la non interprétabilité des lexèmes, et ceci pour chaque

6 Sur ce dernier point, nos résultats semblent rejoindre ceux présentés dans des travaux de psycholinguistiques sur la conscience morphologique (Colé 1988, Grainger, Colé & Segui 1991), qui s'accordent généralement sur le fait que les suffixes jouent un rôle beaucoup important que les préfixes, notamment dans la décomposition lexicale et la construction de l'interprétation.

annotateur, et pour les cinq ensembles de jeux de données. Une fois encore, ceci indique, entre autres, que le degré d'interprétabilité s'obtient en combinant plusieurs critères. Rappelons aussi que, comme les résultats obtenus automatiquement sont très proches des annotations de référence (précision élevée), cela donne une fiabilité assez importante à la saillance des descripteurs et de leurs combinaisons.

7. CONCLUSION ET TRAVAUX FUTURS

Dans notre travail, nous avons essayé de déterminer l'importance relative des différents paramètres pouvant entraver l'interprétabilité des lexèmes de la terminologie médicale. L'approche proposée repose à la fois sur des hypothèses établies à partir de données linguistiques et sur les méthodes utilisées en Traitement Automatique des Langues, permettant de générer les paramètres automatiquement, d'analyser un grand volume de données, et ainsi de faire émerger les paramètres saillants pour la tâche. Parmi ceux qui s'avèrent les plus saillants, nous trouvons surtout l'attestation des lexèmes dans les lexiques de référence et la catégorie grammaticale à laquelle ils appartiennent. D'autres paramètres ressortent, mais dans une moindre mesure, et en association : le degré de complexité des lexèmes, leur fréquence sur la Toile, leur distribution dans la terminologie médicale, chaînes finales de trois ou quatre caractères. Ceci montre à quel point l'interprétabilité des lexèmes fait intervenir de nombreux paramètres.

Ce premier travail, pour limité qu'il soit, nous semble riche de potentialités. Une des pistes pour de futurs travaux concerne l'exploitation de corpus textuels, alors que pour cette étude, nous n'avons utilisé que des paramètres de lexèmes « hors contexte ». Nous pensons en effet que les corpus textuels peuvent fournir d'autres types d'informations pertinentes (sémantiques ou statistiques), qui peuvent aussi être utilisées efficacement pour l'étude de l'interprétabilité. Par ailleurs, la présentation des lexèmes en contexte peut aussi influencer leur degré d'interprétabilité par les locuteurs. À un niveau infra-lexical, nous prévoyons aussi d'étudier l'interprétabilité des lexèmes, ou leur non-interprétabilité, en fonction de celle de leurs constituants. Il nous faudra alors identifier et exploiter des paramètres spécifiques aux constituants eux-mêmes (leur fréquence, leur origine par exemple), ainsi qu'à leurs combinaisons. Des expériences complémentaires, pour certaines plus ciblées et plus circonscrites, devront être envisagées pour chacune de ces approches.

BIBLIOGRAPHIE

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) : 552-557.

AMIOT D., DAL G. (2005). Integrating Combining Forms into a Lexeme-based Morphology. *Paper presented at 5th Mediterranean Morphology Meeting (MMM5)*, Fréjus : 323-336.

ANSCOMBRE J.-C. (2003). Les proverbes sont-ils des expressions figées ? *Cahiers de lexicologie* 82/1 : 159-173.

BAUER L. (2009). Typology of Compounds. In R. Lieber & Štekauer P. (eds), *The Oxford Handbook of Compounding*. Oxford : Oxford University Press : 343-356.

BJÖRNSSON H., HÄRD AF SEGERSTAD B. (1979). *Lix pä franska och tio andra språk*. Stockholm : Pedagogiskt centrum, Stockholms skolförvaltning.

BISETTO A., SCALISE S. (2009). *The classification of compounds*. In Lieber, R. & Štekauer, P. (eds), *The Oxford Handbook of Compounding*. Oxford : Oxford University Press : 34-53.

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

CABRÉ M., ESTOPÀ R. (2002). On the units of specialised meaning uses in professional communication. In *International Network for Terminology* : 217–237.

CABRÉ M.T. (2000). Terminologie et linguistique : la théorie des portes. *Terminologies nouvelles*, 21 : 10–15.

CHANG C.-C., CHIH-JEN L. (2001). LIBSVM : a library for support vector machines. Software available at : http://www.csie.ntu.edu.tw/~cjlin/libsvm.

COLÉ P. (1988). Le traitement des mots dérivés : une analyse morphologique sélective, *L'Année Psychologique,* 88 : 405-418.

CORBIN D., PAUL J. (2000). Aperçus sur la créativité morphologique dans la terminologie de la chimie. *La banque des mots* 60 : 51-68.

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

DAILLE B. (1995). Repérage et extraction de terminologie par une approche mixte statistique et linguistique. *TAL*, 36(1-2) : 101–118.

DAL G., AMIOT D. (2008). La composition néoclassique en français et ordre des constituants. In Amiot, D. (éd.), *La composition dans une perspective typologique.* Arras : Artois Presses Université : 89-113

D'ALESSANDRO D.M., KINGSLEY P., JOHNSON-WEST J. (2001). The readability of pediatric patient education materials on the World Wide Web. *Arch Pediatr Adolesc Med.*, 155/7 : 807-12.

DROUIN P. (2002). *Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés*. PhD thesis, Université de Montréal.

DROUIN P., LANGLAIS P. (2006). Évaluation du potentiel terminologique de candidats termes. In *Actes des 8e Journées internationales d'analyse statistique des données textuelles (JADT-2006)*. Besançon, France : 379–388.

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. *BIONLP* : 49-56.

FLEISS J. L., COHEN J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, *Educational and Psychological Measurement*, Vol. 33 : 613-619.

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

FÓNAGY I. (1997). Figement et changement sémantique, In Martins-Baltar, M. (éd.), *La Locution entre langue et usages*, Fontenay-Saint-Cloud : ENS Éditions : 131-164.

FRANTZI K.T., ANANIADOU S., TSUJII J. (1997). Automatic term recognition using contextual clues. In *Proceedings of the Second Workshop on Multilingality in software Industry : The AI Contribution (MULSAIC'97) – Workshop WL1, IJCAI'97*, Nagoya, Japan.

GODARD D. (2012). Compositionalité : questions linguistiques. *Sémanticlopédie :* http://www.semantique-gdr.net/dico/index.php/Compositionalité:_questions_linguistiques

GRABAR N. (2013). Patient-oriented indexing. *Deliverable 5.3 of the ANR TecSan project RAVEL.*

GODART-WENDLING B., ILDEFONSE F., PARIENTE J.C. , ROSIER I. (1998). Penser le principe de compositionnalité : éléments de réflexion historiques et épistémologiques. *Traitement Automatique des Langues (TAL),* 39(1) : 9–34

GRAINGER J., COLÉ P., SEGUI J. (1991). Masked morphological priming in visual word recognition. *Journal of Memory and Language*, 30 : 370-384.

GROSS M. (1988). Les limites de la phrase figée. *Langages*, 90 : 7-22.

GROSS G. (1996). *Les expressions figées en français, noms composés et autres locutions*. Paris : Ophrys.

GROSS G. (2003). Réflexions sur le figement. *CILL,* 31/2-4 : 45-69.

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 : *Annual Symposium of the American Medical Informatics Association (AMIA)* : 229-33.

HUNSTON S., FRANCIS G. (2000). *Pattern Grammar*. Amsterdam : Benjamins.

IACOBINI C. (2011). Elementi formativi. *Enciclopedia dell'italiano.* Roma: Istituto dell'Enciclopedia Italiana : 416-418.

KINCAID J.P., FISHBURNE R.P. JR, ROGERS R.L., CHISSOM B.S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. *Research Branch Report*, Millington, TN : Naval Technical Training, U.S. Naval Air

Station, Memphis, TN : 8-75.

KLEIN J.R. (2007). Le figement, un concept aussi essentiel que fluent. Réflexions à travers la synchronie et la diatopie, In : A. Haeki Buhofer and H. Burger (eds), *Phraseology in motion II.* (*Actes du Colloque EUROPHRAS, Bâle, août 2004*). Hohengren : Schneider Verlag : 75-84.

KOKKINAKIS D., TOPOROWSKA GRONOSTAJ M. (2006). Comparing Lay and Professional Language in Cardiovascular Disorders Corpora. *WSEAS Transactions on BIOLOGY and BIOMEDICINE* : 429-437

KORKONTZELOS I., KLAPAFTIS I.P., MANANDHAR S. (2008). Reviewing and evaluating automatic term recognition techniques. In B. Nordström and A. Ranta, (eds), *Advances in Natural Language Processing (6th International Conference on NLP, GoTAL 2008*) 5221 *LNAI,* Springer : 248–259.

LAMIROY B. (2003). Les notions linguistiques de figement et de contrainte, *Linguisticae Investigationes,* 26/1 : 1-14.

LAMIROY B. (2008). Les expressions figées : à la recherche d'une définition. In : P. Blumenthal et S. Mejri (éds), *Les séquences figées : entre langue et discours*. Stuttgart : Franz Steiner Verlag : 85–88.

LEROY G., HELMREICH S., COWIE J., MILLER T., ZHENG W. (2008). Evaluating Online Health Information : Beyond Readability Formulas. *AMIA 2008* : 394-398

lexique.org : *http://www.lexique.org/*

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*) : 28–32.

LÜDELING A. (2006). Neoclassical Compounding. In : K. Brown (ed.), *Encyclopedia of language and linguistics*, 2nd Edition. Oxford : Elsevier : 580–582.

LÜDELING A., SCHMIDT T., KIOKPASOGLOU S. (2002). Neoclassical word formation in German. *Yearbook of Morphology* : 253-283.

MANUILA L., MANUILA A., LEWALLE P., NICOULIN M. (2001). *Dictionnaire médical*. 9ème éd., Paris : Masson.

MARTIN R. (1997). Sur les facteurs du figement lexical. In M. Martins-Baltar (éd.), *La locution entre langue et usages*. Fontenay-St Cloud : ENS éditions : 291-305.

MAYNARD D., ANANIADOU S. (2000). Identifying terms by their family and friends. In : *Proceedings of COLING 2000*. Saarbrucken: Germany : 530-536.

MCCRAY A. (2005). Promoting Health Literacy. *Journal of American Medical Informatics Association*, 12 : 152-163.

MEJRI S. (2008). La place du figement dans la description des langues. In : P. Blumenthal et S. Mejri (éds), *Les séquences figées : entre langue et discours*. Stuttgart : Franz Steiner Verlag : 117–129.

NAMER F. (2005a). Guessing the meaning of neoclassical compounds within LG : the case of pathology nouns. *Generative Approaches to the Lexicon 2005*, (may 19-21), Geneva : 175-84.

NAMER F. (2005b). Morphosémantique pour l'appariement de termes dans le

vocabulaire médical : approche multilingue. *TALN 2005* (6-10 juin 2005), Dourdan : 63-72.

NAMER F., BAUD R. (2005). Predicting Lexical Relations between Biomedical Terms : towards a Multilingual Morphosemantics-based system. *Studies in Heath Technology and Informatics,* 116 : 793-798.

NAMER F., VILLOING F. (2005). Have cutthroats anything to do with tracheotomes ? Distinctive properties of VN vs NV compounds in French. *5th Mediterranean Morphology Meeting (MMM5)* (14 au 18 septembre 2005), Fréjus : 105-124.

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

NAMER F., ZWEIGENBAUM P. (2004). Acquiring meaning for French medical terminology : contribution of morphosemantics. In : *Annual Symposium of the American Medical Informatics Association (AMIA)*, San-Francisco.

NUNBERG G., SAG I., WASOW T. (1994). Idioms. *Language,* 70 : 491-538.

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

PARTEE B.H. (1984). *Compositionality. Varieties of formal semantics*. Landman F. & Veltman F, Pbsh.

QUINLAN J.R. (1993). *C4.5 : Programs for Machine Learning.* Morgan Kaufmann Publishers, San Francisco, CA, USA.

RONNEBERGER-SIBOLD E. (2006). Lexical Blends : Functionally Tuning the Transparency of Complex Words. *Folia linguistica* 40, 1-2 : 155-181

SVENSSON M. H. (2004). *Critères de figement*, Umea Universitet.

TLFi, *Trésor de la Langue Française informatisé* : http://www.atilf.fr/

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

WITTEN I.H., FRANK E. (2005). *Data mining : Practical machine learning tools and techniques.* Morgan Kaufmann, San Francisco.

WÜSTER E. (1981). L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses. In : V.I. Siforov (ed.), *Textes choisis de terminologie*, vol. I. *Fondements théoriques de la terminologie*, *GISTERM*. Québec : Université de Laval : 55–114.