Analyse et évolution de la compréhension de termes techniques

Natalia Grabar[1] Thierry Hamon[2]
(1) CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
natalia.grabar@univ-lille3.fr
(2) LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France;
Université Paris 13, Sorbonne Paris Cité, F-93430 Villetaneuse, France
hamon@limsi.fr

RÉSUMÉ

Nous faisons l'hypothèse que les mots techniques inconnus dotés d'une structure interne (mots affixés ou composés) peuvent fournir des indices linguistiques à un locuteur, ce qui peut l'aider à analyser et à comprendre ces mots. Afin de tester notre hypothèse, nous proposons de travailler sur un ensemble de mots techniques provenant du domaine médical. Un grand ensemble de mots techniques est annoté par cinq annotateurs. Nous effectuons deux types d'analyses : l'analyse de l'évolution des mots compréhensibles et incompréhensibles (de manière générale et en fonction de certains suffixes) et l'analyse des clusters avec ces mots créés par apprentissage non-supervisé, sur la base des descripteurs linguistiques et extra-linguistiques. Nos résultats indiquent que, selon la sensibilité linguistique des annotateurs, les mots techniques peuvent devenir décodables et compréhensibles. Quant aux clusters, le contenu de certains reflète la difficulté des mots qui les composent et montre également la progression des annotateurs dans leur compréhension. La ressource construite est disponible pour la recherche : http://natalia.grabar.free.fr/rated-lexicon.html.

ABSTRACT

Analysis and Evolution of Understanding of Technical Terms.

We propose a hypothesis according to which unknown words with internal structure (affixed words or compounds) can provide the speaker with linguistic cues, which can help with analysis and understanding of these words. To evaluate our hypothesis, we propose to work with a set of technical words from the medical area. A large set of technical words is annotated by five annotators. We perform two kinds of analysis : an analysis of the evolution of understandable and non-understandable words (globally and according to some suffixes) and an analysis of clusters created with these words by unsupervised machine learning approach on the basis of linguistic and extra-linguistic descriptors. Our results suggest that, according to the linguistic sensibility of annotators, technical words can become decodable and understandable by them. As for the clusters, the content of some of them is representing their understanding difficulty and also indicates the evolution of annotators in their understanding. The resource is available for the research purposes : http://natalia.grabar.free.fr/rated-lexicon.html.

MOTS-CLÉS : Mots inconnus, domaine de spécialité, décodage linguistique des mots, régularités linguistiques.

KEYWORDS: Unknown words, specialized area, linguistic decoding of words, linguistic regularities.

1 Introduction

Dans sa vie quotidienne, un locuteur peut rencontrer des mots qui lui sont inconnus. Cela peut concerner les néologismes (comme en (1) et (2)) ou bien les mots techniques provenant de domaines de spécialité (comme en (3) et (4)). Dans les deux cas, la compréhension de ces mots n'est pas évidente et leur sémantique reste plus ou moins opaque.

(1) *Le concept de l'année qui fait bonne fortune dans les sciences humaines est un mot impossible à mémoriser. « Sérendipité » ne figure même pas dans les dictionnaires français. (www.scienceshumaines.com, 26/07/2016)*

(2) *À l'issue de ce parcours, l'auteur formule un certain nombre de propositions d'actions pour transformer les frontages français. Elles sont regroupées en deux grandes catégories : « rééquilibrer le partage modal de la rue » (diminuer la place de la voiture au profit d'une pluralité de modes de transport, des piétons, au vélo) ; et « rééquilibrer le partage de la rue », en favorisant la plurifonctionnalité de ces espaces publics. (www.metropolitiques.eu, 03/05/2013)*

(3) *Jacques Chirac souffre depuis plusieurs mois d'anosognosie, indique le rapport médical signé par le professeur Olivier Lyon-Caen et remis vendredi au juge Pauthe et au parquet de Paris, que s'est procuré Le Journal du Dimanche. (Le Monde.fr, 04.09.2011)*

(4) *Ces deux pays ont commencé à manger du fugu pour une raison très simple : toutes espèces confondues, il est d'une rare abondance dans cette région du monde car il a très peu de prédateurs. (www.lemanger.fr/index.php/fugu-poisson-mortel/)*

Plusieurs procédés linguistiques sont disponibles pour enrichir le lexique d'une langue. En français, parmi les plus usités, nous pouvons mentionner les procédés propres à une langue (affixation, composition y compris composition néoclassique, comme dans l'exemple (3)) ou les procédés d'emprunt et d'adaptation de lexèmes venant d'une autre langue, comme dans les exemples (1), (2) et (4) (Guilbert, 1971). Nous nous intéressons plus particulièrement aux mots qui disposent d'une structure interne, comme *frontage* ou *anosognosie*, car nous supposons que les régularités linguistiques (composants et affixes des mots, règles exploitées pour former leur structure interne, etc.) peuvent aider le locuteur à déduire la structure et la sémantique de tels mots. Il s'agit alors de régularités et similarités pouvant être observées au niveau de traits linguistiques que le locuteur peut dégager et intégrer. En effet, en linguistique, la question de compréhension est liée à différents facteurs, comme :

– connaissance et reconnaissance des composants des mots complexes : la manière de segmenter un mot, par exemple *anosognosie*, en composants ;
– patrons morphologiques et de relations entre les composants : la manière d'articuler ses composants et de former la sémantique du mot (Iacobini, 2003; Amiot & Dal, 2008; Namer, 2003).

Pour étudier notre hypothèse, nous proposons de travailler sur un ensemble de mots techniques pris hors contexte issus d'une terminologie médicale. Le choix de travailler hors contexte est fait pour plusieurs raisons :

1. lorsque les mots nouveaux apparaissent, ils bénéficient souvent de contextes peu nombreux et relativement pauvres, qui ne peuvent pas toujours aider la compréhension ;
2. de manière similaire, dans un domaine de spécialité, les contextes, et surtout les contextes non définitoires, apportent souvent très peu d'aide pour la compréhension d'un terme ;

3. finalement, le travail avec des mots hors contexte permet de traiter un ensemble de mots plus grand et donc d'avoir des observations obtenues sur du matériel linguistique plus important ;
4. d'un autre point de vue, l'analyse de mots en contexte correspond à sa perception *en extension* qui se base sur des indices environnants, alors que l'analyse de mots hors contexte correspond à sa perception *en intention* qui se base sur des indices internes au mot.

Pour ces différentes raisons, nous supposons que la structure interne de mots inconnus peut également aider dans leur compréhension. Selon notre hypothèse, les mots affixés ou composés, qui par définition ont une structure interne, peuvent ainsi fournir des indices linguistiques requis et le locuteur peut ainsi analyser les mots inconnus en se basant sur la structure qu'il arrive à dégager. Un autre aspect de notre hypothèse est que les régularités dans la langue, en ce qui concerne l'emploi des composants et des affixes, peuvent être détectées par les locuteurs, sans que les informations explicites (sémantique des composants, règles de formation de mots, etc.) leur soient fournies.

Afin de tester notre hypothèse, nous proposons de travailler sur un ensemble de mots techniques provenant d'un domaine de spécialité, qui est la médecine dans notre cas. Malgré la présence de plus en plus importante de ce domaine dans notre vie, la médecine manipule beaucoup de termes techniques qui restent inconnus aux locuteurs qui ne sont pas spécialistes de ce domaine.

Dans la suite de ce travail, nous présentons quelques travaux existants (section 2), les données sur lesquelles nous travaillons (section 3) et la méthode que nous proposons pour vérifier notre hypothèse (section 4). Nous décrivons et discutons les résultats obtenus (section 5) et terminons avec les pistes pour les travaux futurs (section 6).

2 Travaux existants

Nous nous attachons de présenter ici les travaux existants qui portent sur la compréhension de textes par les locuteurs. Il existe également de nombreux travaux sur le traitement de mots inconnus des dictionnaires dans les applications automatiques, que nous ne présentons pas ici.

En Traitement Automatique des Langues, il existe une grande variété de travaux et d'approches consacrés à l'étude de la compréhension ou de la lisibilité des mots. Les travaux en lisibilité cherchent à effectuer une analyse des textes afin de déterminer s'ils sont accessibles ou non pour un lecteur donné. Les mesures de lisibilité sont communément utilisées pour évaluer la complexité de documents. Parmi les mesures existantes, on distingue les mesures classiques de lisibilité et les mesures computationnelles de lisibilité (François, 2011). Les mesures classiques exploitent d'habitude les informations sur le nombre de caractères et/ou de syllabes des mots, et des modèles de régression linéaire (Flesch, 1948; Gunning, 1973), tandis que les mesures computationnelles peuvent impliquer les modèles vectoriels et une grande variété de descripteurs, parmi lesquels les suivants ont été largement utilisés, en particulier sur des données biomédicales : la combinaison de mesures de lisibilité classiques avec des terminologies médicales (Kokkinakis & Toporowska Gronostaj, 2006) ; les n-grammes de caractères (Poprat *et al.*, 2006) ; les descripteurs stylistiques (Grabar *et al.*, 2007) ou discursifs (Goeuriot *et al.*, 2007) ; le lexique (Miller *et al.*, 2007) ; les informations morphologiques (Chmielik & Grabar, 2011) ; ou la combinaison de différents types de descripteurs (Wang, 2006; Zeng-Treiler *et al.*, 2007; Leroy *et al.*, 2008; François & Fairon, 2013; Gala *et al.*, 2013). Les différents travaux en lisibilité computationnelle des documents montrent souvent de très bons résultats, pouvant par exemple dépasser 90 % de précision et de rappel avec des descripteurs morphologiques (Chmielik & Grabar, 2011). Ce type de travaux est essentiellement effectué avec les textes et mots de domaines

spécialisés, comme en témoigne la plupart de travaux cités ici. Cependant, quelques travaux proposent également de travailler sur la langue générale et la compréhension ou l'acquisition du lexique par les enfants (François & Fairon, 2013; Gala *et al.*, 2013).

En psycholinguistique, les questions de compréhension du lexique se focalisent sur différents aspects de la lecture et de l'activité linguistique des sujets :

1. connaissance des composants des mots complexes et leur décomposition : la question de fond concerne avant tout la manière dont les mots complexes (affixés ou composés) sont stockés dans le cerveau, ce qui permet ensuite de décoder ces mots ou bien de les produire. Il a été ainsi démontré que plusieurs indices peuvent faciliter la lecture ou la production de mots complexes. Nous pouvons distinguer les indices internes et externes :
   (a) les indices internes permettent de rendre la structure interne des mots plus explicite. Parmi ces indices se trouvent par exemple : la présence d'un tiret (Bertram *et al.*, 2011) ou d'un espace (Frisson *et al.*, 2008), la présentation des mots avec d'autres mots qui leurs sont proches morphologiquement (Lüttmann *et al.*, 2011) ou des amorces (linguistiques, pseudosuffxes, etc.) (Bozic *et al.*, 2007; Beyersmann *et al.*, 2012) ;
   (b) les indices externes permettent d'expliciter la sémantique ou la structure des mots grâce aux informations externes à ce mot. Parmi les indices externes se trouvent par exemple : l'emploi des images (Dohmes *et al.*, 2004; Koester & Schiller, 2011) ou des contextes favorables (Cain *et al.*, 2009) ;
2. ordre de composants et variété de patrons morphologiques : une grande attention est également portée à l'ordre des composants et à la position (tête ou modifieur) de ces composants. Il a en effet été observé que ces facteurs ont une influence stable dans la reconnaissance de mots complexes (Libben *et al.*, 2003; Holle *et al.*, 2010; Feldman & Soltano, 1999). La notion de la transparence sémantique de la tête morphologique (*morphological headedness*) a été ainsi isolée (Jarema *et al.*, 1999; Libben *et al.*, 2003) : les travaux autour de cette notion indiquent qu'elle joue un rôle important dans la décomposition des mots complexes, dans la reconnaissance de patrons de décomposition, et dans l'activité lexicale de manière générale ;
3. influence de la longueur des mots ou des types d'affixes (Meinzer *et al.*, 2009) ;
4. impact de la fréquence des bases (Feldman *et al.*, 2004) ;
5. stockage et traitement des informations linguistiques : les expériences montrent que les processus lexicaux et morphologiques semblent être traités dans plusieurs zones du cerveau, comme le lobe frontal inférieur gauche, le lobe pariétal droit et le lobe temporo-occipital de manière bilatérale (Bozic *et al.*, 2007; Meinzer *et al.*, 2009; Koester & Schiller, 2011).

Notre hypothèse, selon laquelle l'analyse et l'accès aux règles internes d'un mot peut améliorer la compréhension de ce mot, a été également débattue en psycholinguistique (Baumann *et al.*, 2003; Kuo & Anderson, 2006; McCutchen *et al.*, 2014). Typiquement, cette hypothèse correspond aux points 1(a), 2, 3 et peut-être 4 ci-dessus. Elle s'oppose à deux autres hypothèses : l'acquisition en contexte, qui correspond au point 1(b), et l'information explicite fournie sur la sémantique des composants ou sur les règles linguistiques impliquées dans la formation des mots. L'efficacité de l'analyse de la structure morphologique des mots semble aujourd'hui être acceptée par les psycholinguistes (Bowers & Kirby, 2010). Ceci va dans le sens de notre hypothèse. Cependant, dans notre travail, pour vérifier cette hypothèse, nous exploitons les méthodes de TAL et les descripteurs générés grâce aux méthodes de TAL. De cette manière, nous pouvons (1) travailler avec un gros volume de données linguistiques et de descripteurs, et (2) exploiter des méthodes quantitatives et non-supervisées.

3 Données exploitées

Les données linguistiques sont obtenues à partir de la terminologie médicale Snomed International (Côté, 1996), dont la vocation consiste à décrire aussi exhaustivement que possible le domaine médical. La terminologie contient 151 104 termes médicaux structurés en onze axes sémantiques (*e.g.* maladies et anomalies, actes médicaux, produits chimiques, organismes vivants, anatomie). Nous gardons les termes provenant de cinq axes (maladies, anomalies, actes médicaux, fonctions et anatomie), que nous considérons comme des axes contenant les termes centraux de la médecine, auxquels les locuteurs sont le plus souvent confrontés. Ainsi, nous ne souhaitons pas nous concentrer sur les termes et mots très spécialisés et le plus souvent inconnus des locuteurs, comme par exemple les produits chimiques (*trisulfure d'hydrogène*) ou les organismes vivants (*Sapromyces, Acholeplasma laidlawii*). Cependant, de tels mots peuvent faire partie des termes étudiés ici. Les termes sélectionnés (104 649) sont segmentés en mots pour obtenir une liste de 29 641 mots uniques, qui constituent notre matériel de travail. Cet ensemble contient des mots composés (*abdominoplastie, dermabrasion*), construits (*cardiaque, acineux, lipoïde*) et simples (*acné, fragment*), mais aussi des abréviations (*ADPase, ECoG, Fya*) et des emprunts (*stripping, Conidiobolus, stent, blind*). Ces données sont annotées par cinq locuteurs natifs du français, agés de 25 à 60 ans, sans formation en médecine et de statuts socio-professionnels différents. Chaque annotateur a reçu un ensemble avec les mêmes 29 641 mots, mais ordonnés différemment de manière aléatoire à chaque fois. Il a été indiqué aux annotateurs que, lors de l'annotation, l'utilisation de toute source d'information explicite (dictionnaires, encyclopédies, etc.) n'est pas autorisée. De même, ils n'ont pas le droit de revenir sur les annotations effectuées auparavant. La tâche présentée aux annotateurs consiste à assigner chaque mot à une des trois catégories : *1. Je peux comprendre le mot*, *2. Je ne suis pas sûr*, *3. Je ne peux pas comprendre le mot*. D'une part, nous supposons que ces annotateurs représentent un niveau modéré de « lisibilité » des locuteurs (Schonlau *et al.*, 2011) et que nous pourrons généraliser nos observations sur le même type de population. D'autre part, nous supposons que sur la base de ces annotations, nous pouvons observer s'il existe une progression dans la compréhension de mots techniques de manière générale et d'un type donné. Ces annotations manuelles correspondent donc aux données de référence que nous analysons et sur lesquelles nous faisons des observations.

4 Méthode

Notre méthode comporte deux aspects principaux :

1. analyse de la progression de la compréhension des mots au sein des catégories *1 (Je peux comprendre le mot)* et *3 (Je ne peux pas comprendre le mot)*, globalement et en fonction de quelques composants (section 4.1) ;
2. classification non supervisée des mots sur base d'un ensemble de descripteurs, analyse du contenu des clusters et comparaison avec les annotations manuelles (section 4.2).

4.1 Progression globale de la compréhension de mots

La progression globale de la compréhension de mots correspond au taux de mots compréhensibles ou incompréhensibles annotés à un moment t. Cela permet de voir de manière globale si les locuteurs arrivent à maîtriser plus de mots au fur et à mesure qu'ils effectuent les annotations. Cette analyse est

effectuée de manière générale, sur tous les mots de l'ensemble, et de manière plus ciblée, sur les mots qui comportent un composant donné.

4.2 Classification non supervisée des mots

La classification est effectuée de manière non supervisée. Nous testons plusieurs algorithmes implémentés dans Weka : `SOM` (Self-Organizing Map) (Kohonen, 1989), `Canopy` (McCallum *et al.*, 2000), `Cobweb` (Fisher, 1987), `EM` (Expectation Maximization) (Dempster *et al.*, 1977), `SimpleKMeans` (Witten & Frank, 2005). Sauf avec `SimpleKMeans` et `EM`, il n'est pas nécessaire d'indiquer le nombre de clusters attendus. Dans tous les cas, les algorithmes se basent sur les descripteurs présentés ci-dessous, et les régularités détectées pour créer les clusters.

Chaque mot est décrit avec un ensemble de 23 descripteurs linguistiques et extra-linguistiques liés à la langue générale et spécialisée. Les descripteurs, calculés automatiquement, peuvent être groupés en 8 classes :

- *Catégories syntaxiques.* Les catégories syntaxiques et les lemmes sont calculés par `TreeTagger` (Schmid, 1994) et corrigés par `Flemm` (Namer, 2000). Les catégories syntaxiques sont assignées aux mots dans les contextes de leurs termes. Si un mot donné reçoit plus d'une catégorie, c'est la plus fréquente qui est retenue comme descripteur. Parmi les catégories principales, nous trouvons les noms, les adjectifs, les noms propres, les verbes et les abréviations. Au total, nous avons 11 catégories syntaxiques : abréviation *ABR*, adjectif *Adj*, adverbe *Adv*, préposition *D*, déterminant *DET*, emprunt *F*, nom *N*, nom propre *NAM*, numéral *NUM*, pronom *PRO*, verbe *V* ;
- *Présence des mots dans les lexiques de référence.* Nous exploitons deux lexiques de référence du français : TLFi [1] et `lexique.org` [2]. TLFi est un dictionnaire de la langue française, qui couvre XIXème et XXème siècles. Ils contient presque 100 000 entrées. `lexique.org` est un lexique créé pour les travaux en psycholinguistique. Il contient plus de 135 000 entrées, dont les formes flexionnelles de verbes, adjectifs et noms (presque 35 000 lemmes) ;
- *Fréquence des mots dans un moteur de recherche généraliste.* Nous interrogeons le moteur de recherche Google pour obtenir une indication de la fréquence des mots sur la Toile. Les mots qui sont plus fréquents peuvent également être plus faciles à comprendre ;
- *Fréquence des mots dans la terminologie médicale.* Nous calculons la fréquence de chaque mot dans la terminologie Snomed International. Les mots qui font partie d'un plus grand nombre de termes dans cette terminologie peuvent être plus centraux et donc plus facilement compréhensibles par les locuteurs ;
- *Nombre d'axes sémantiques associés aux mots, et leurs types.* Nous exploitons l'information sur les axes sémantiques de la Snomed International associés à chaque terme et répercutés sur les mots composant ces termes, parce que nous nous attendons à ce que les termes qui apparaissent dans différents axes sémantiques sont également plus fondamentaux et sans doute mieux compréhensibles par les locuteurs ;
- *Longueur des mots en nombre de caractères et de syllabes.* Nous calculons le nombre de caractères et de syllabes [3] de chaque mot. Nous nous attendons à ce que les mots longs soient plus difficiles à comprendre ;
- *Nombre de bases et d'affixes.* Chaque lemme est traité avec l'analyseur morphologique `Dérif`

1. `http://atilf.atilf.fr/tlf.htm`
2. `http://www.lexique.org/`
3. Module Perl `Lingua::EN::Syllable` (`http://search.cpan.org/~gregfast/Lingua-EN-Syllable-0.251/Syllable.pm`), adapté au français

(Namer, 2003), adapté au domaine médical. `Dérif` effectue entre autre la décomposition des lemmes en bases et affixes connus de sa base de données. Nous exploitons la décomposition morphologique pour avoir le nombre de bases et d'affixes des mots. Nous nous attendons à ce que les mots avec plus de composants et d'affixes soient plus difficiles à comprendre ;

– *Chaîne de caractères initiale et finale.* Nous calculons les chaînes de caractères initiales et finales de longueurs allant de trois à cinq caractères. Nous supposons que ces sous-chaînes peuvent évoquer les composants positionnés à la fin ou au début des mots. La motivation principale de ce descripteur est que les sous-chaînes finales peuvent correspondre la tête morphologique et sémantique des mots et donc permettre de les décoder et catégoriser ;
– *Nombre et pourcentage de consonnes, voyelles et autres caractères.* Nous calculons le nombre et le pourcentage de consonnes, voyelles et autres caractères (*i.e.* , tiret, apostrophe, virgule).

Dans le tableau 1, nous présentons des exemples de mots avec leurs descripteurs.

lemme	POS	l_1	l_2	f_g	f_t	nb_a	nb_s	initial	final	nb_c	nb_v
alarme	N	+	+	73400000	6	1	2	ala,alar,alarm	rme,arme,larme	3	3
hépatite	N	+	+	15300000	9	3	3	hép,hépa,hépat	ite,tite,atite	4	4
angiocholite	N	-	+	74700	12	1	5	ang,angi,angio	ite,lite,olite	6	6
desmodontose	N	+	-	2050	12	1	4	des,desm,desmo	ose,tose,ntose	7	5

TABLE 1 – Extrait des données traitées avec les descripteurs (catégorie syntaxique *POS*, présence dans les lexiques de référence (TLFI l_1 et `lexique.org` l_2), fréquence dans un moteur de recherche f_g, fréquence dans la terminologie f_t, nombre d'axes sémantiques nb_a, nombre de syllabes nb_s, chaînes initiales et finales (*initial, final*), nombre de consonnes nb_c, nombre de voyelles nb_v).

Nous effectuons plusieurs expériences pour lesquelles nous varions les descripteurs exploités :

– l'ensemble des 23 descripteurs,
– un ensemble de descripteurs réduit aux propriétés linguistiques : catégorie syntaxique, nombre de syllabes, chaînes de caractères initiales et finales, ce qui permet de prendre en compte les observations faites dans les travaux psycholinguistiques (Jarema *et al.*, 1999; Libben *et al.*, 2003; Meinzer *et al.*, 2009),
– l'ensemble réduit de descripteurs linguistiques avec la fréquence dans un moteur de recherche généraliste, ce qui permet de prendre en compte d'autres observations faites en psycholinguistique (Feldman *et al.*, 2004).

Avec `SimpleKMeans` et `EM`, nous effectuons deux séries d'expériences, où le nombre de clusters est fixé à 1 000 et 2 000 (pour presque 30 000 individus à clusteriser). Au sein des clusters, nous supposons pouvoir retrouver les régularités linguistiques des mots, en fonction des descripteurs considérés. Nous voulons également observer si les clusters reflètent la compréhension de mots par les annotateurs, éventuellement avec la prise en compte de l'évolution chronologique de leurs annotations.

5 Résultats et discussion

Nous présentons et discutons d'abord l'annotation manuelle (section 5.1), et ensuite les deux aspects de la méthode : (1) l'analyse de la progression de la compréhension des mots globalement et en fonction de quelques composants (section 5.2) et (2) la classification non supervisée des mots (section 5.3).

5.1 Annotation manuelle

L'annotation de l'ensemble des mots par les différents annotateurs a nécessité entre 3 semaines et 3 mois. Notons que le temps nécessaire pour l'annotation dépend surtout de la motivation des annotateurs, et non pas de leur âge ou statut socio-professionnel. Par ailleurs, il n'était pas question d'annoter l'ensemble du matériel en une seule fois : les annotateurs étaient libres de choisir leur propre rythme afin de ne pas ressentir trop de fatigue.

Lorsque l'on travaille avec les données provenant d'un domaine de spécialité, l'accord inter-annotateur (Cohen, 1960) sur la compréhension des mots est très élevé : il est supérieur à 0.730. Nous avons fait une observation similaire dans un autre travail (Grabar *et al.*, 2014).

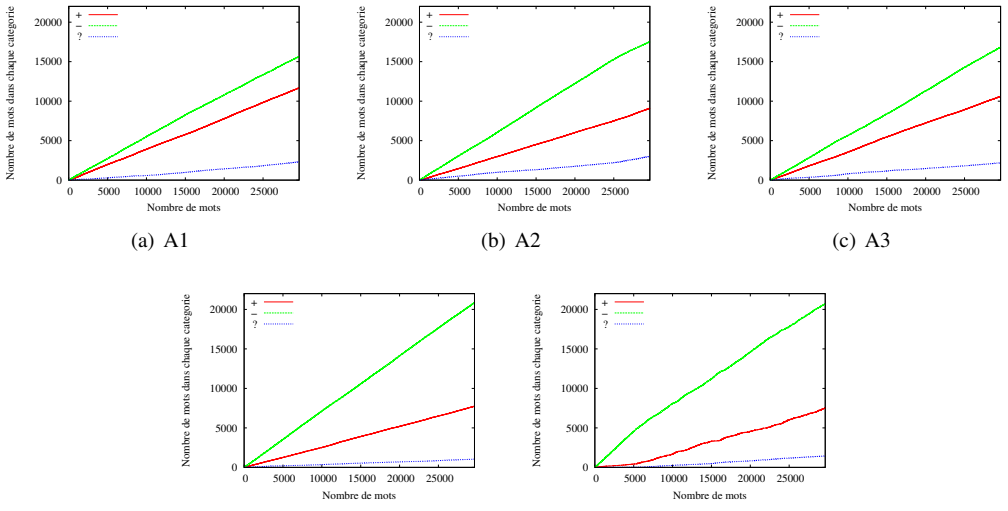
L'annotation manuelle permet également de dégager plusieurs types de mots difficiles à comprendre :
- les abréviations (*e.g.* , *OG, VG, PAPS, j, bat, cp*) ;
- les noms propres (*e.g.* , *Gougerot, Sjögren, Bentall, Glasgow, Babinski, Barthel, Cockcroft*), qui font souvent partie des noms d'examens, formules, indices, maladies, etc. ;
- les noms de médicaments ;
- de très nombreux termes médicaux techniques correspondant aux maladies, examens ou actes médicaux. Il s'agit le plus souvent de composés néoclassiques (*e.g. antihémophile, pseudohémophilie, sclérodermie, hydrolase, orthotopique, tympanectomie, arthrodèse, synesthésie*) ;
- les emprunts au latin ou à l'anglais ;
- les termes relatifs à l'anatomie humaine (*e.g. cloacal, pubovaginal, nasopharyngé, mitral, diaphragmatique, antre, inguinal, strontium, érythème, maxillo-facial, mésentérique, mésentère*).

5.2 Progression de la compréhension de mots

Les figures 1 et 2 indiquent l'évolution de croissance de deux catégories : *Je ne peux pas comprendre le mot* et *Je peux comprendre le mot*, qui totalisent plus de 90 % de mots. La catégorie *Je ne suis pas sûr* reste très minoritaire. La figure 1 indique l'évolution du nombre de mots dans les trois catégories pour chacun des cinq annotateurs. De tendance générale, le nombre de mots dans la catégorie *Je ne peux pas comprendre le mot* est plus important par rapport à deux autres catégories. Cependant, la maîtrise de ce lexique par les annotateurs est variable : l'annotateur A1 montre la meilleure maîtrise alors que pour les annotateurs A5 et A4 il s'agit d'un domaine beaucoup plus opaque.

La figure 2 indique l'évolution de trois catégories étudiées en pourcentage. La ligne correspondant à l'ensemble *Je ne peux pas comprendre le mot* est dans la partie supérieure des graphiques, alors que la ligne *Je peux comprendre le mot* inférieure. La catégorie *Je ne suis pas sûr* a toujours les valeurs les plus faibles. Nous pouvons distinguer deux tendances dans l'évolution de ce pourcentage :
- avec les annotateurs A5, A2 et A1, on peut observer la tendance de diminution de la proportion de mots inconnus. Ceci est le plus visible avec l'annotateur A5. Avec ces annotateurs, nous supposons qu'ils sont devenus plus familiers avec certains composants et bases, et qu'ils ont pu mieux maîtriser le lexique traité ;
- avec les annotateurs A3 et A4, après une augmentation de la proportion de mots inconnus au début de l'annotation, cette proportion reste assez stable par la suite. Avec ces annotateurs, nous pouvons supposer que le processus d'annotation d'un lexique volumineux ne leur a pas permis de mieux maîtriser les bases et composants véhiculés par ce lexique.

Une autre observation intéressante est visible avec A1 surtout, mais aussi A2 et A4 : le nombre de mots pour lesquels ils hésitent diminue sensiblement au profit des mots connus surtout et, dans une



(a) A1 (b) A2 (c) A3

(d) A4 (e) A5

Fig. 1 – Évolution globale du nombre de mots dans chaque catégorie.

moindre mesure, des mots inconnus. Ensuite, le nombre de mots connus ne semble plus augmenter, sauf pour A5. De plus, cet l'effet d'apprentissage opère surtout sur les 2 000 premiers mots et concerne le passage de mots pas sûr à des mots connus.

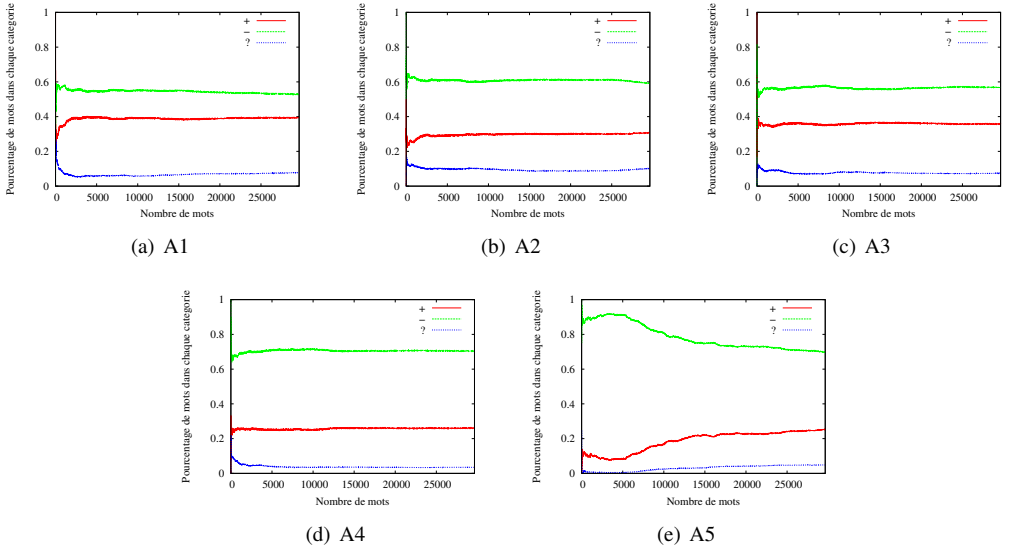
Les figures 3 et 4 indiquent l'évolution de la compréhension de mots en *-ite* et *-tomie*, respectivement. Pour les mots avec ces deux suffixes, nous voyons que l'annotateur A5 a le plus de difficultés à décoder et apprendre les mots : les mots non compréhensibles s'accumulent. Rappelons que la tendance de cet annotateur sur l'ensemble des mots traités était contraire : le pourcentage de mots inconnus diminuait progressivement. Les annotateurs A2 et A4 ont également des difficultés avec les mots en *-tomie* et *-ite*, respectivement. Concernant les autres annotateurs, ils montrent des progrès dans le décodage de mots en *-ite* et *-tomie*. Ils ont général une première vague importante d'amélioration de la compréhension au début, et ensuite une deuxième vague plus progressive.

Sur la base de ces observations, nous pouvons voir que, en fonction des types de mots, de leurs caractéristiques linguistiques et de la sensibilité des annotateurs, il est possible d'avoir une amélioration progressive dans la compréhension d'un lexique technique et *a priori* inconnu. Les régularités dans la langue peuvent donc aider à améliorer la compréhension des locuteurs, comme déjà remarqué dans les travaux en psycholinguistique (Lüttmann *et al.*, 2011). Nous allons maintenant voir si ces régularités peuvent aussi être exploitées par les algorithmes de clusterisation.

5.3 Classification non supervisée des mots

Dans le tableau 2, nous indiquons le nombre de clusters obtenus avec nos ensembles de descripteurs :

– `SOM` crée très peu de clusters : ils sont très grands et hétérogènes. Par exemple, avec E_f, les clusters contiennent 13 088, 4 840, 7 023 et 4 690 individus ;



(a) A1 (b) A2 (c) A3

(d) A4 (e) A5

Fig. 2 – Évolution globale du pourcentage de mots dans chaque catégorie.

Descripteurs	SOM	Canopy	Cobweb
E_c= Ensemble complet (23)	5	62	33853
E_r= Ensemble réduit (8)	4	28	12577
E_f= E_r et fréquence (9)	4	27	9861

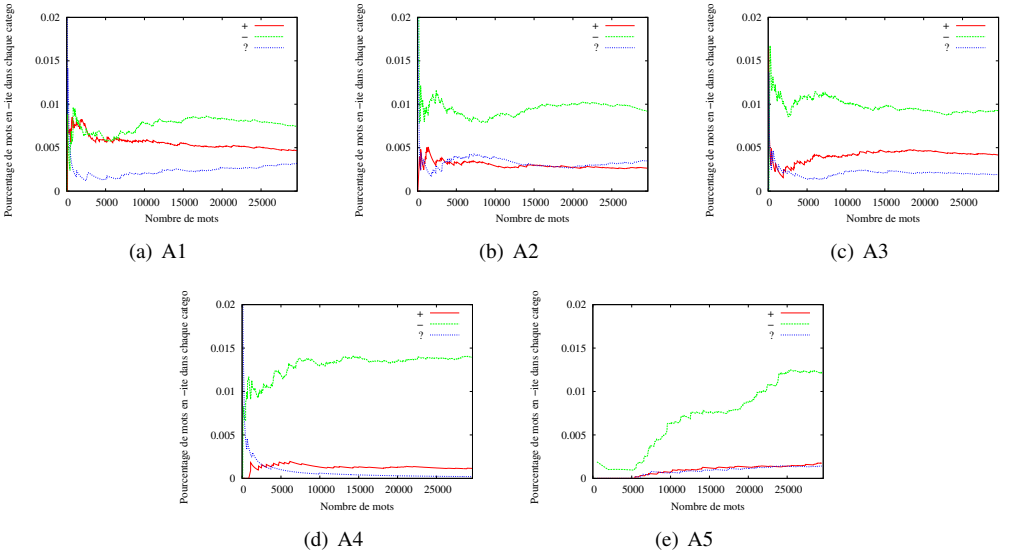
TABLE 2 – clusters

– `Cobweb` crée énormément de clusters dont la majeure partie sont des singletons. Par exemple, avec E_f, 9 374 clusters sur 9 861 sont singletons ;
– `EM` et `SimpleKMeans` créent le nombre de clusters indiqué : 1 000 et 2 000 ;
– `Canopy` crée entre 30 et 60 clusters, en fonction des descripteurs.

Nous proposons de travailler avec les résultats de trois algorithmes : `EM`, `SimpleKMeans` et `Canopy`.

Avec les descripteurs des ensembles E_r et E_f, la création des clusters est essentiellement motivée par les chaînes initiales des mots (pas forcément selon les 3 à 5 premiers caractères) et dans une moindre mesure par leurs catégories syntaxiques et leurs fréquences. Par exemple, nous pouvons avoir des clusters de mots commençant par p ou a, ou encore des clusters qui regroupent des *phosphates* ou des enzymes en *-ase*. Ces groupements de substances chimiques deviennent intéressants. Cependant, globalement les clusters obtenus avec les descripteurs des ensembles E_r et E_f offrent peu d'intérêt. Nous proposons donc d'analyser plutôt les clusters obtenus avec l'ensemble de descripteurs E_c.

Avec `Canopy`, la taille des clusters va de 1 à 2 823 éléments. Plusieurs clusters sont dédiés aux deux catégories principales de l'annotation. Par exemple, 30 clusters contiennent au moins 80 % de mots de la catégorie *1 (Je peux comprendre le mot)*, alors que 6 clusters contiennent au moins 80 % de mots de la catégorie *3 (Je ne peux pas comprendre le mot)*. Parmi les clusters avec les mots compréhensibles nous avons par exemple :



(a) A1 (b) A2 (c) A3

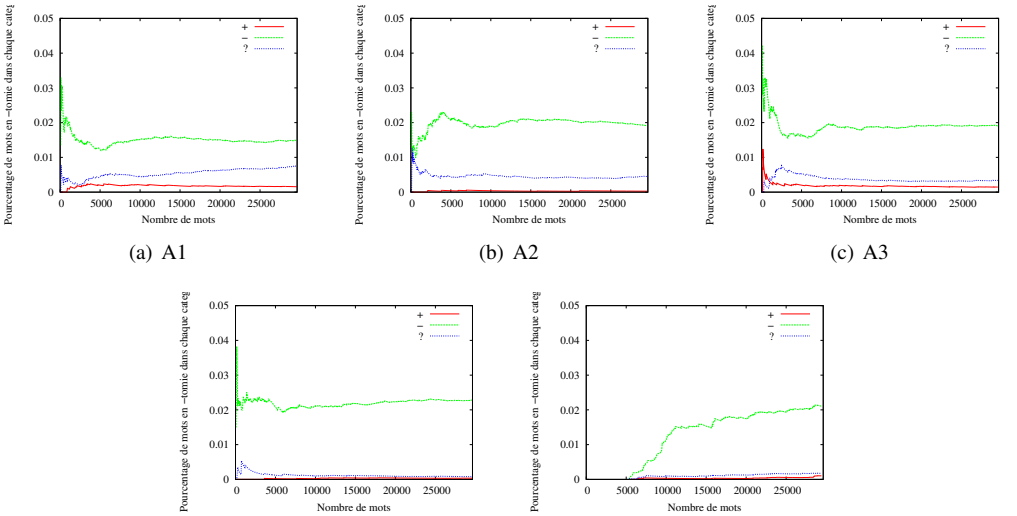
(d) A4 (e) A5

Fig. 3 – Évolution du pourcentage de mots en *-ite* dans chaque catégorie.

– des clusters séparés avec des numéraux (*mil, quinzième, treizième, deuxième, neuf, onzième*), des verbes (*allaite, étend, constitue, importe, analiser, assoir, contient*), et des adverbes (*massivement, rarement, probablement, occasionnellement, spontanément, proprement, tardivement, souvent, secondairement*) groupés selon leur catégorie syntaxique et éventuellement leurs chaînes finales ;
– des clusters de mots grammaticaux (*du, aucun, les, aux, des, au*) groupés selon leurs longueurs ;
– des adjectifs assez communs (*rudimentaire, prolongé, intact, perméable, hystérique, maximal, mobile, fracturé, instable, gros, narcissique, inadéquat, original, mort, sec, légal, manuel, traumatique, militaire, intrinsèque, large*) groupés selon leur catégorie syntaxique et leurs fréquences ;
– des adjectifs participiaux (*inapproprié, stratifié, relié, modifié, localisé, précisé, reliées, quadruplé*) groupés selon leur catégorie syntaxique, leurs fréquences et leurs chaînes finales ;
– des adjectifs plus spécialisés mais fréquents (*rotulien, vital, spasmodique, putréfié, zonale, redondant, structural, violacé, tremblant, vénal, synchrone, sensoriel*), également groupés selon leur catégorie syntaxique et leurs fréquences ;
– des noms spécialisés mais assez communs (*dentiste, brosse, altitude, blocage, diagnostic, avant-bras, glucose, fourrure, dépression, artères, carnassier, ankylose, contraception, cavité, électrode, compression, autorité, bile, lactation, monocyte, ecchymose, angoisse, diphtérie, allergène, caséine, bronche, entorse, accident, grossesse, aine, cheville, aversion, alvéole, côlon, fesse, carcinome*) groupés selon leur catégorie syntaxique et leurs fréquences.

Parmi les clusters avec les mots non compréhensibles nous avons par exemple :

– des substances chimiques (*glutamine-scyllo-inosose, héparosane-N-sulfate-glucuronate, cynurénine-oxoglutarate, désoxycytidine-diphosphate, désoxythymidine-monophosphate, biotine-acétyl-CoA-carboxylase, alkylglycérone-phosphate, apolipoprotéine-lipoprotéine, diméthylallyltransférase, dihydroxyisovalérate, phosphoribosylaminoimidazolecarboxamide, désoxyadénosine-diphosphate*) groupés selon leur catégorie syntaxique, les types de caractères contenus et les fréquences ;



(a) A1 (b) A2 (c) A3

(d) A4 (e) A5

Fig. 4 – Évolution du pourcentage de mots en *-tomie* dans chaque catégorie.

– des emprunts (*felis, punctum, Saprolegnia, pigmentosum, framboesia, equuli, dissimilis, frutescens, materia, mégarectum, diminutus, flavivirus,glauca, ghost, marinus, dolorosi, kansasii, immitis, folliclis, musculi*) groupés selon leur catégorie syntaxique, les chaînes finales et les fréquences ;
– des noms propres (*Dee, Fiedler, Muehrcke, Keshan, Liebow, Thogoto, Churg, Lyme, Luschka, Blessig, Rokitansky-Aschoff, Gerlier, Koerber-Salus-Elschnig, Corino, Danlos, Sténon, Pouteau-Colles, Fiessinger-Leroy-Reiter*) groupés selon leur catégorie syntaxique.

À l'intérieur de ces clusters dédiés à la catégorie *3 (Je ne peux pas comprendre le mot)*, on n'observe pas la progression des annotateurs ni une meilleure maîtrise de cette terminologie de leur part.

Il existe cependant des clusters plus mitigés, qui comportent autant les mots de la catégorie *1 (Je peux comprendre le mot)* que de la catégorie *3 (Je ne peux pas comprendre le mot)*, de même que des hésitations. Ces clusters contiennent par exemple :

– des substances chimiques et des aliments (*créatinine, antitussif, acétyl-CoA-carboxylase, céphalosporine, dopamine, créatine, aminophylline, centaurée, chaudronniers, ambroisie, aubergine, carotte, cerisier, béribéri, gentamicine, antidépresseur, antiacide, bauxite, bromure, anguille, dioxyde*) groupés selon leurs chaînes finales, axes sémantiques et fréquences ;
– des fonctions de l'organisme, des maladies et des procédures médicales (*paraparésie, négligence, névralgie, extrasystole, myéloblaste, nucléotide, syncope, psychose, persistance, immunodéficience, spasticité, putrescine, mortalité, léchage, orgasme, réplication, précocité, puberté, séminome, raideur, malaise, spermicide, sénescence, pseudopuberté*) groupés selon les fréquences, les chaînes finales et la catégorie syntaxique ;
– des adjectifs plus spécialisés relatifs aux organes et aux maladies (*périprostatique, sous-tentoriel, péribuccale, condylienne, hémolysant, paratyphoïde, modulant, présphénoïde, médio-tarsienne, intracortical, fibrosante, monohydraté, sous-diaphragmatique, rétracté, polyploïde, nécrosant, prévésical, trachéo-bronchique, prédisposant, myoïde, infiltrant, rémittente, péribuccaux*) groupés

selon la catégorie syntaxique et les fréquences.
C'est surtout au sein de ces derniers clusters, qui recouvrent en partie les figures 3 et 4, que l'évolution de la compréhension peut être présente et observable. Par exemple, nous avons un cluster avec des procédures médicales en *-tomie*, dont les mots, au fur et à mesure de l'annotation sont moins annotés dans la catégorie *3 (Je ne peux pas comprendre le mot)* et plus dans la catégorie *2 (Je ne suis pas sûr)*, voire dans la catégorie *1 (Je peux comprendre le mot)*.

Le contenu des clusters et nos observations indiquent que :
– étant donné un ensemble de descripteurs appropriés, il est possible de créer de manière non supervisée des clusters de mots selon la facilité de leur compréhension dans les non-spécialistes ;
– au sein de certains clusters, il est possible d'observer l'évolution des annotateurs dans leur maîtrise de certains types de mots techniques. Par exemple, cet effet peut être observé avec les mots signifiant les maladies ou les procédures. Cependant, pour d'autres types de mots (substances, emprunts, noms propres) aucune évolution ni amélioration de compréhension n'est observable.

6 Conclusion et perspectives

Nous faisons l'hypothèse que les régularités linguistiques, lorsqu'elles apparaissent suffisamment souvent, peuvent aider dans le décodage et la compréhension de mots techniques et inconnus. Notre expérience est effectuée avec les mots techniques du domaine médical. Un ensemble de presque 30 000 mots est annoté par cinq annotateurs. Les catégories possibles sont *Je peux comprendre le mot*, *Je ne suis pas sûr* et *Je ne peux pas comprendre le mot*. Dans l'ensemble de chaque annotateur, les mots sont ordonnés de manière aléatoire et différente. Nous effectuons une analyse au niveau de l'ensemble de 30 000 mots et au niveau de deux suffixes (*-ite* et *-tomie*). Notre travail indique que plusieurs annotateurs montrent des signes d'un meilleur décodage des mots techniques, ce qui va dans le même sens que les travaux existants (Lüttmann *et al.*, 2011). Cet effet est observé sur l'ensemble des mots ou pour les deux suffixes présentés. En revanche, pour d'autres annotateurs et ensembles de mots, le décodage des mots n'est pas possible : le pourcentage de mots inconnus augmente progressivement. La perception et compréhension de mots semblent donc suivre différents schémas selon la sensibilité linguistique des locuteurs.

Ces observations ont été croisées avec les clusters générés sur la base de descripteurs linguistiques et extra-linguistiques. D'une part, le contenu des clusters peut être mis en relation avec les catégories d'annotation manuelle et donc être relatif à la compréhension des mots : plusieurs clusters comportent une majorité de mots provenant des catégories *1 (Je peux comprendre le mot)* et *3 (Je ne peux pas comprendre le mot)*. D'autre part, au sein de certains clusters, en particulier lorsqu'ils contiennent des groupes de mots homogènes (maladies, procédures...) nous pouvons voir une amélioration de la part des annotateurs quant à leur compréhension. Ces deux résultats sont très intéressants et confirment l'hypothèse de notre travail, selon laquelle les régularités linguistiques peuvent aider à décoder les mots techniques et inconnus. De plus, les descripteurs appropriés peuvent également permettre de distinguer la facilité de compréhension de manière non supervisée. En revanche, d'autres clusters comportent des mots très spécifiques (substances chimiques, noms propres, emprunts...) et très peu fréquents, pour lesquels une amélioration de compréhension n'est pas observée.

Nous avons plusieurs perspectives à ce travail : (1) effectuer le même type d'annotation, mais en fournissant la sémantique de certains ou de tous les composants (oralement ou dans un fichier fourni), mais il sera difficile de contrôler si cette information est effectivement exploitée par les annotateurs ;

(2) effectuer le même type d'annotation, mais en autorisant les annotateurs à consulter des sources d'informations externes (dictionnaires, exemples qui peuvent être trouvés en ligne...). Comme ce type d'approche demande plus de temps et de charge cognitive, il devra être effectué avec un ensemble de mots beaucoup plus petit ; (3) analyser l'évolution de la compréhension de mots, en prenant en compte un plus grand nombre de composants finaux ; (4) appuyer les observations par des tests statistiques de significativité ; (5) mieux croiser le contenu des clusters générés avec les différents descripteurs et propriétés des mots ; (6) exploiter les résultats de ce travail pour la formation et l'éducation des non-experts pour les aider dans la maîtrise de notions médicales qui sont importantes pour eux ; (7) exploiter les résultats de ce travail pour la simplification de textes techniques. Par exemple, les descripteurs des mots qui présentent des difficultés de compréhension peuvent décrire les classes de mots qui devraient être simplifiés.

Les ressources construites sont librement disponibles pour la recherche : `http://natalia.grabar.free.fr/rated-lexicon.html`.

Remerciements

Nous remercions les annotateurs pour la lourde tâche d'annotation de presque 30 000 mots. Nous remercions également les relecteurs anonymes pour leurs remarques et suggestions qui ont permis d'améliorer la qualité du travail et de sa présentation. Cette recherche a en partie bénéficié de l'aide des partenaires financeurs de l'IReSP dans le cadre de l'appel à projets général 2016 volet Services de Santé (GAGNAYRE-AAP16-HSR-6) et du projet ANR MIAM (ANR-16-CE23-0012).

Références

AMIOT D. & DAL G. (2008). La composition néoclassique en français et ordre des constituants. *La composition dans les langues*, p. 89–113.

BAUMANN J., EDWARDS E., BOLAND E., OLEJNIK S. & KAME'ENUI E. (2003). Vocabulary tricks : Effects of instruction in morphology and context on fifth-grade students' ability to derive and infer word meanings. *American Educational Research Journal*, **40**(2), 447–494.

BERTRAM R., KUPERMAN V., BAAYEN H. R. & HYÖNÄ J. (2011). The hyphen as a segmentation cue in triconstituent compound processing : It's getting better all the time. *Scandinavian Journal of Psychology*, **52**(6), 530–544.

BEYERSMANN E., COLTHEART M. & CASTLES A. (2012). Parallel processing of whole words and morphemes in visual word recognition. *The Quarterly Journal of Experimental Psychology*, **65**(9), 1798–1819.

BOWERS P. & KIRBY J. (2010). Effects of morphological instruction on vocabulary acquisition. *Reading and Writing*, **23**(5), 515–537.

BOZIC M., MARSLEN-WILSON W. D., STAMATAKIS E. A., DAVIS M. H. & TYLER L. K. (2007). Differentiating morphology, form, and meaning : Neural correlates of morphological complexity. *Journal of Cognitive Neuroscience*, **19**(9), 1464–1475.

CAIN K., TOWSE A. S. & KNIGHT R. S. (2009). The development of idiom comprehension : An investigation of semantic and contextual processing skills. *Journal of Experimental Child Psychology*, **102**(3), 280–298.

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

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

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

DEMPSTER A., LAIRD N. & RUBIN D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, **39**(1), 1–38.

DOHMES P., ZWITSERLOOD P. & BÖLTE J. (2004). The impact of semantic transparency of morphologically complex words on picture naming. *Brain and Language*, **90**(1-3), 203–212.

FELDMAN L. B. & SOLTANO E. G. (1999). Morphological priming : The role of prime duration, semantic transparency, and affix position. *Brain and Language*, **68**(1-2), 33–39.

FELDMAN L. B., SOLTANO E. G., PASTIZZO M. J. & FRANCIS S. E. (2004). What do graded effects of semantic transparency reveal about morphological processing ? *Brain and Language*, **90**(1-3), 17–30.

FISHER D. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, **2**(2), 139–172.

FLESCH R. (1948). A new readability yardstick. *Journ Appl Psychol*, **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.

FRISSON S., NISWANDER-KLEMENT E. & POLLATSEK A. (2008). The role of semantic transparency in the processing of english compound words. *Br J Psychol*, **99**(1), 87–107.

GALA N., FRANÇOIS T. & FAIRON C. (2013). Towards a french lexicon with difficulty measures : NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. In *eLEX-2013*.

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

GRABAR N., HAMON T. & AMIOT D. (2014). Automatic diagnosis of understanding of medical words. In *EACL PITR Workshop*, p. 11–20.

GRABAR N., KRIVINE S. & JAULENT M. (2007). Classification of health webpages as expert and non expert with a reduced set of cross-language features. In *AMIA*, p. 284–288.

GUILBERT L. (1971). De la formation des unités lexicales. In P. LAROUSSE, Ed., *Grand Larousse de la langue française*, p. IX–LXXXI.

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

HOLLE H., GUNTER T. C. & KOESTER D. (2010). The time course of lexical access in morphologically complex words. *Neuroreport*, **21**(5), 319–323.

IACOBINI C. (2003). Composizione con elementi neoclassici. In M. GROSSMANN & F. RAINER, Eds., *La formazione delle parole in italiano*, p. 69–96. Walter de Gruyter.

JAREMA G., BUSSON C., NIKOLOVA R., TSAPKINI K. & LIBBEN G. (1999). Processing compounds : A cross-linguistic study. *Brain and Language*, **68**(1-2), 362–369.

KOESTER D. & SCHILLER N. O. (2011). The functional neuroanatomy of morphology in language production. *NeuroImage*, **55**(2), 732–741.

KOHONEN T. (1989). *Self-Organization and Associative Memory*. Springer.

KOKKINAKIS D. & TOPOROWSKA GRONOSTAJ M. (2006). Comparing lay and professional language in cardiovascular disorders corpora. In A. PHAM T., JAMES COOK UNIVERSITY, Ed., *WSEAS Transactions on BIOLOGY and BIOMEDICINE*, p. 429–437.

KUO L. & ANDERSON R. (2006). Morphological awareness and learning to read : A cross-language perspective. *Educational Psychologist*, **41**(3), 161–180.

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

LIBBEN G., GIBSON M., YOON Y. B. & SANDRA D. (2003). Compound fracture : The role of semantic transparency and morphological headedness. *Brain and Language*, **84**(1), 50–64.

LÜTTMANN H., ZWITSERLOOD P. & BÖLTE J. (2011). Sharing morphemes without sharing meaning : Production and comprehension of german verbs in the context of morphological relatives. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, **65**(3), 173–191.

MCCALLUM A., NIGAM K. & UNGAR L. (2000). Efficient clustering of high dimensional data sets with application to reference matching. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 169–178.

MCCUTCHEN D., STULL S., HERRERA B. L., LOTAS S. & EVANS S. (2014). Putting words to work : Effects of morphological instruction on children's writing. *J Learn Disabil*, **47**(1), 1–23.

MEINZER M., LAHIRI A., FLAISCH T., HANNEMANN R. & EULITZ C. (2009). Opaque for the reader but transparent for the brain : Neural signatures of morphological complexity. *Neuropsychologia*, **47**(8-9), 1964–1971.

MILLER T., LEROY G., CHATTERJEE S., FAN J. & THOMS B. (2007). A classifier to evaluate language specificity of medical documents. In *HICSS*, p. 134–140.

NAMER F. (2000). FLEMM : un analyseur flexionnel du français à base de règles. *Traitement automatique des langues (TAL)*, **41**(2), 523–547.

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

POPRAT M., MARKÓ K. & HAHN U. (2006). A language classifier that automatically divides medical documents for experts and health care consumers. In *MIE 2006 - Proceedings of the XX International Congress of the European Federation for Medical Informatics*, p. 503–508, Maastricht.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *ICNMLP*, p. 44–49, Manchester, UK.

SCHONLAU M., MARTIN L., HAAS A., DEROSE K. & RUDD R. (2011). Patients' literacy skills : more than just reading ability. *J Health Commun*, **16**(10), 1046–54.

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

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

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.