Un modèle pour prédire la complexité lexicale et graduer les mots

Núria Gala[1] Thomas François[2] Delphine Bernhard[3] Cédrick Fairon[2]
(1) LIF-CNRS UMR 7279, Aix Marseille Université,
(2) CENTAL, Université Catholique de Louvain,
(3) LILPA, Université de Strasbourg
nuria.gala@lif.univ-mrs.fr, tfrancois@uclouvain.be, dbernhard@unistra.fr, cfairon@uclouvain.be

Résumé. Analyser la complexité lexicale est une tâche qui, depuis toujours, a principalement retenu l'attention de psycholinguistes et d'enseignants de langues. Plus récemment, cette problématique a fait l'objet d'un intérêt grandissant dans le domaine du traitement automatique des langues (TAL) et, en particulier, en simplification automatique de textes. L'objectif de cette tâche est d'identifier des termes et des structures difficiles à comprendre par un public cible et de proposer des outils de simplification automatisée de ces contenus. Cet article aborde la question lexicale en identifiant un ensemble de prédicteurs de la complexité lexicale et en évaluant leur efficacité via une analyse corrélationnelle. Les meilleures de ces variables ont été intégrées dans un modèle capable de prédire la difficulté lexicale dans un contexte d'apprentissage du français.

Abstract. Analysing lexical complexity is a task that has mainly attracted the attention of psycholinguists and language teachers. More recently, this issue has seen a growing interest in the field of Natural Language Processing (NLP) and, in particular, that of automatic text simplification. The aim of this task is to identify words and structures which may be difficult to understand by a target audience and provide automated tools to simplify these contents. This article focuses on the lexical issue by identifying a set of predictors of the lexical complexity whose efficiency are assessed with a correlational analysis. The best of those variables are integrated into a model able to predict the difficulty of words for learners of French.

Mots-clés : complexité lexicale, analyse morphologique, mots gradués, ressources lexicales.

Keywords: lexical complexity, morphological analysis, graded words, lexical resources.

1 Introduction

La complexité lexicale n'est pas une notion qui puisse être définie dans l'absolu. En effet, un terme est perçu différemment en fonction du public qui y est confronté (apprenants de langue maternelle, apprenants de langue seconde, personnes avec une difficulté ou une pathologie liée au langage, etc.), d'où le terme de 'difficulté' (complexité subjective, (Blache, 2011)). De même, s'appuyer sur le seul critère de la fréquence pour appréhender la complexité du lexique semble réducteur : bien que ce critère se soit avéré très efficace dans la littérature (voir section 2), cette variable ne peut seule expliquer l'ensemble des problèmes rencontrés par différentes catégories de lecteurs. La notion de 'complexité' est, ainsi, multidimensionnelle (vitesse d'accès au lexique mental, compréhension, mémorisation, prononciation, activation du sens, orthographe, etc.), difficilement saisissable à partir de critères uniquement statistiques et très liée aux caractéristiques du public envisagé.

Dans le cadre de cet article, nous visons un public d'apprenants du français langue maternelle (L1) ou de français langue étrangère (FLE). En tenant compte de plusieurs ressources existantes, nous avons identifié un ensemble de variables intra-lexicales et statistiques que nous avons intégrées dans un modèle statistique cherchant à prédire le degré de complexité de mots dont la difficulté a été annotée par ailleurs. Notre hypothèse est que la combinaison de plusieurs variables intralexicales fines, associées à des informations statistiques, peut donner des indications plus précises sur le degré de complexité d'un mot. Dans ce sens, après un état de l'art introductif à la section 2, nous présentons la méthodologie et les ressources que nous avons utilisées pour identifier des variables susceptibles de caractériser la complexité lexicale (section 3). Dans un deuxième temps, nous présentons ces variables et nous discutons de leur impact à la section 4. À la section 5, nous décrivons le modèle de difficulté intégrant ces prédicteurs et nous analysons les résultats obtenus. Enfin, nous concluons l'article par une discussion sur notre approche et les résultats obtenus à la section 6, avant de proposer quelques futures améliorations à la section 7.

2 État de l'art

Analyser la complexité lexicale est une tâche qui, depuis toujours, a principalement intéressé les psycholinguistes et les pédagogues. En effet, de nombreux travaux sont décrits dans la littérature et se basent, par exemple, sur des tâches telles que la décision lexicale, la catégorisation sémantique, etc. pour explorer les propriétés du lexique. Ainsi, l'un des critères majeurs pour considérer qu'un mot est simple ou complexe est celui de la fréquence : de nombreux travaux démontrent la corrélation étroite entre la haute fréquence d'un terme et le fait que celui-ci soit perçu comme plus 'simple' (Howes & Solomon, 1951; Monsell, 1991). C'est d'ailleurs le critère que plusieurs auteurs avaient utilisé dans la première moitié du 20e siècle pour construire les premières ressources de lexique 'simplifié', par exemple la liste de Thorndike (1921), le *Teachers' Book of Words*, qui reprend les 20 000 mots les plus courants de la langue anglaise assortis de leur fréquence d'usage, ou encore le *Français fondamental* de Gougenheim (1958) qui comprend 1 500 mots usuels pour l'apprentissage du français, aussi bien en tant que langue étrangère que maternelle. La liste de Thorndike reste une référence dans le domaine de la lisibilité (avant l'apparition des listes obtenues par traitement informatisé). Elle s'avère un instrument de mesure objectif de la difficulté lexicale des textes et ce malgré quelques faiblesses, comme la mauvaise estimation des fréquences des mots appelés *disponibles* (mots avec fréquence variée selon les corpus mais usuels et utiles [1]).

D'autres critères avancés dans la littérature pour identifier des mots 'simples' concernent plutôt la familiarité d'un terme (Gernsbacher, 1984) ou encore son âge d'acquisition (Brysbaert *et al.*, 2000). La familiarité lexicale a été utilisée pour la constitution d'une liste de mots simples par Dale (1931). Dans l'expérience menée par Dale et ses collègues, la mesure de familiarité a été définie comme suit : dans une liste de 10 000 mots, n'ont été retenus que les termes connus par au moins 80% des élèves de quatrième primaire (CM1), ce qui a réduit la liste à 3 000 mots. Le nombre de voisins orthographiques (nombre d'unités de même longueur ne se différenciant que par une seule lettre) a aussi été envisagé par Coltheart *et al.* (1977) comme une mesure discriminante de la difficulté d'accès au lexique mental, même si les résultats dans des tâches de décision lexicale semblent varier selon les langues. Enfin, la longueur (en nombre de syllabes et/ou caractères) apparaît aussi comme un facteur déterminant dans la façon de percevoir les unités lexicales, en particulier parce qu'un mot plus long augmente la probabilité de fixer la fovéa (zone de la rétine où la vision des détails est la plus précise) sur un point de position non optimal, ce qui engendre une perte de temps à la lecture (Vitu *et al.*, 1990). Plus récemment, Schreuder & Baayen (1997) démontrent que le nombre de morphèmes et la taille de la famille morphologique jouent un rôle dans la décision lexicale visuelle (reconnaissance de mots parmi une série de mots et non-mots). Laufer (1997), pour sa part, identifie une série de facteurs linguistiques influençant l'acquisition du lexique, parmi lesquels la familiarité des phonèmes, la régularité dans la prononciation, la cohérence graphème-phonème, la transparence morphologique ou la polysémie. Potentiellement, ces facteurs contribuent tous à la façon dont les mots sont perçus.

Les répercussions de ces travaux sont d'abord théoriques, aidant, par exemple, à comprendre l'organisation du lexique mental et comment il se distribue dans les différentes zones du cerveau. D'un point de vue plus pratique, certaines de ces études ont cependant débouché sur la construction de listes utilisées pour l'enseignement des langues. Plus récemment, la question de l'évaluation de la difficulté lexicale a fait l'objet d'un intérêt grandissant dans le domaine du traitement automatique des langues (TAL) et, en particulier, en simplification automatique de textes. Dans ce domaine, le but reste d'identifier des termes et des structures difficiles à comprendre par un public cible et de proposer des outils de simplification automatisée de ces contenus. Bien que la plupart des travaux en simplification de textes se focalisent sur des aspects syntaxiques (par exemple (Chandrasekar *et al.*, 1996)), certains auteurs ont mis en oeuvre des systèmes qui visent le traitement du lexique. Dans ce cas, différents aspects doivent être pris en compte : (i) la détection des mots ou termes complexes à remplacer, (ii) l'identification de substituts et (iii) l'adéquation au contexte. Ces trois aspects ne sont pas toujours pris en compte de manière conjointe. Sous sa forme la plus simple, la substitution lexicale se fait en fonction de la fréquence des synonymes extraits d'une ressource comme WordNet, sans prise en compte du contexte (Carroll *et al.*, 1998). Récemment, des travaux ont fait appel à des corpus comparables comme Wikipedia et sa version simplifiée pour l'anglais (*Simple English Wikipedia*) pour acquérir des ressources utiles pour la simplification lexicale : ainsi, Biran *et al.* (2011) proposent une mesure de la complexité d'un mot qui est fonction de sa fréquence dans les deux versions de Wikipedia et de sa longueur. D'une manière générale, les critères utilisés pour sélectionner le meilleur substitut restent relativement simples. Pour la tâche de simplification lexicale organisée lors de la campagne SemEval 2012 (Specia *et al.*, 2012), la *baseline* correspondant à une simple mesure de fréquence dans un gros corpus n'a été battue que par un seul système. Ce résultat rend compte de la difficulté de la tâche : même si les travaux en psycholinguistique ont mis en évidence des facteurs complexes, leur intégration dans des systèmes automatisés n'est pas encore résolue.

1. Par exemple "fourchette", "coude", etc.

3 Exploitation de ressources existantes

Pour réaliser les différentes expériences présentées dans cet article, nous avons eu recours à un ensemble de ressources qui ont été utilisées en vue de deux objectifs : certaines ressources lexicales ont servi de liste de référence pour l'apprentissage du modèle, tandis que les autres ressources ont été employées pour récupérer diverses informations linguistiques utilisées dans nos variables.

3.1 Ressources d'apprentissage

Pour entraîner un modèle statistique capable de prédire ou comparer la difficulté de mots, il importe de disposer d'un nombre suffisant de mots dont la difficulté est connue et exprimée en fonction d'une unité pratique. En psycholinguistique, il est commun d'associer le temps de réponse nécessaire pour réaliser une tâche associée à un mot à la difficulté de ce mot (Ferrand, 2007). Cependant, cette approche nécessite de disposer d'un nombre important de sujets et de moyens, en particulier lorsqu'on envisage un large vocabulaire. C'est pourquoi nous avons choisi de constituer notre ressource d'entraînement d'une autre façon : en nous basant sur l'association des mots à des niveaux scolaires déterminés, calculés sur la base de l'apparition de ces mots dans des manuels scolaires. Notre hypothèse est qu'un mot facile apparaîtra en général plus tôt dans les manuels scolaires qu'un mot plus complexe. Par chance, il existe deux ressources pour le français qui recensent les mots utilisés dans des manuels scolaires de différents niveaux, à savoir Manulex (Lété *et al.*, 2004) et FLELex (François *et al.*, 2014).

Manulex [2] a été créée à partir de 54 manuels scolaires (pour un total de 1,9 millions d'occurrences). Il décrit la distribution d'unités lexicales en fonction de leur apparition dans des manuels qui ont été classés en trois niveaux : (1) la première année de primaire ou CP (6 ans), (2) la deuxième année ou CE1 (7 ans) et (3) une catégorie qui regroupe les trois années suivantes (CE2-CM2, 8-11 ans). Ce choix se justifie en termes de volume d'acquisition de vocabulaire : au CP, se construit le lexique de l'enfant sur la base de la médiation phonologique ; au CE1, se construit le lexique orthographique par automatisation progressive de la reconnaissance du mot écrit et au cycle 3, le stock lexical se consolide et s'enrichit par exposition répétée à l'écrit [3]. La ressource, librement disponible, totalise 23 812 lemmes, mais nous n'avons conservés que les mots lexicaux (noms, adjectifs, adverbes et verbes), ce qui réduit le nombre de lemmes à 19 038. Il faut aussi signaler que les fréquences associées à chaque mot de la ressource ne correspondent pas aux valeurs absolues observées dans les manuels, mais à des valeurs adaptées en fonction d'un indice de dispersion qui augmente l'importance des termes en fonction du nombre de documents dans lesquels ils sont apparus. La Figure 1 présente un exemple d'entrées issues de Manulex.

lemme	POS	Fréq. N1	Fréq N2	Fréq. N3
pomme	N	724	306	224
vieillard	N	-	13	68
patriarche	N	-	-	1
cambrioleur	N	2	-	33

TABLE 1 – Exemple d'entrées de Manulex

FLELex, quant à lui, a été obtenu à l'aide d'une méthodologie similaire, mais sur la base d'un corpus de 28 manuels de français langue étrangère (FLE) et de 29 livres simplifiés également destinés à des lecteurs en FLE. Ces ouvrages étaient classés selon l'échelle de difficulté proposée par le cadre européen commun de référence pour les langues (Conseil de l'Europe, 2001) ou CECR, qui définit six niveaux de maîtrise communicationnelle : A1 (niveau introductif ou de survie) ; A2 (niveau intermédiaire) ; B1 (niveau seuil) ; B2 (niveau avancé ou utilisateur indépendant) ; C1 (niveau autonome ou de compétence opérationnelle effective) et C2 (maîtrise). La ressource totalise 14 053 lemmes lexicaux et 183 lemmes grammaticaux, dont les fréquences ont été estimées sur 777 835 occurrences.

Dans les deux cas, et comme le montre la Table 1, le problème de ces ressources par rapport à notre propre objectif de recherche est qu'elles offrent la distribution des fréquences de chaque mot par niveau, mais n'associent pas strictement un mot à un niveau donné. C'est pourquoi nous avons dû transformer ces distributions en un niveau. Trois techniques ont été testées pour ce faire. La plus simple d'entre elles consiste à attribuer à un mot le premier niveau où il a été observé dans le corpus. Ainsi pour Manulex, *pomme* et *cambrioleur* se voient attribuer le niveau 1, tandis que *patriarche* est associé au

2. http://www.manulex.org
3. http://leadserv.u-bourgogne.fr/bases/manulex/manulexbase/indexFR.htm

niveau 3. On comprend aisément que cette façon de faire, qui assimile la distribution de *pomme* à celle de *cambrioleur* n'est pas optimale. C'est pourquoi nous avons également considéré chaque distribution comme une série statistique (ex. pour *pomme*, une série constituée de 724 un, de 306 deux et de 224 trois, les chiffres 'un' à 'trois' renvoyant respectivement aux trois niveaux scolaires, voir Table 1) et pris comme valeur représentative soit son premier quartile, soit sa moyenne (ce qui donne alors une échelle continue, comprise entre 1 et 3 pour Manulex et 1 et 6 pour FLELex).

3.2 Ressources pour l'extraction de variables

Pour l'extraction de variables, nous avons mobilisé plusieurs ressources contenant différentes informations linguistiques. Leur usage particulier au sein de nos variables est décrit plus en détail dans la section 4, cette section présentant ces ressources de façon plus générale.

La première d'entre elles est Lexique 3 [4] (New *et al.*, 2001). il s'agit d'une ressource librement accessible qui contient un grand nombre d'informations linguistiques (transcription phonétique, structure syllabique, flexion, etc.) et statistiques (nombre de phonèmes, de syllabes, de morphèmes, fréquence dans des corpus de livres et de sous-titre de films, etc.). Elle contient 142 728 mots correspondant à 47 342 lemmes.

Polymots [5] (Gala & Rey, 2008) est un lexique morphologique. Les mots ont été segmentés morphologiquement en bases et affixes, des informations sur les familles morphologiques et sur des unités de sens associées sont également disponibles. La version 3 contient 19 510 lemmes et 2 364 familles. La segmentation morphologique et le regroupement en familles ont été effectués manuellement, ce qui a comme répercussion une couverture assez faible (par rapport à d'autres ressources comme Lexique3, par exemple). Ainsi, l'intersection entre Manulex (restreint aux mots lexicaux) et Polymots est de 55,75 %, c'est-à-dire qu'il y a 10 614 mots communs entre les deux ressources [6]. De ce fait, nous avons décidé d'utiliser une méthode par apprentissage non supervisé pour l'obtention des variables morphologiques. Les lemmes de Polymots, tout comme ceux de Morphalou 2.0 (Romary *et al.*, 2004), nous ont servi à enrichir notre corpus d'apprentissage pour l'analyse morphologique.

Nous avons également utilisé un corpus issu d'enregistrements de patients atteints de la maladie de Parkinson (2 271 formes pour 373 lemmes). Il s'agit d'une vingtaine d'enregistrements (correspondant à une tâche de description d'une image de la vie quotidienne) de patients en état "off", c'est-à-dire, sans médicaments qui pourraient inhiber les effets de la maladie sur la parole. Nous nous sommes intéressés à ce type de parole pathologique car la maladie de Parkinson, bien qu'elle soit plus connue pour des symptômes moteurs (tremblements, rigidité musculaire, etc.), entraîne également des difficultés au niveau de la parole (Pinto *et al.*, 2010). Par conséquent, nous postulons que ce type de parole pathologique peut être représentative d'une langue plus simple et donc d'un lexique plus simple. La classification que nous proposons des structures syllabiques (variable 8, section 4) est issue d'observations faites sur ce corpus. Les données de ce corpus ont aussi servi à enrichir le corpus d'apprentissage pour l'extraction des informations morphologiques.

Enfin, pour l'obtention de variables sémantiques, nous avons utilisé les réseaux lexicaux JeuxDeMots [7] (Lafourcade, 2007) et BabelNet (Navigli & Ponzetto, 2010). JeuxdeMots contient à ce jour 314 494 termes, dont 136 421 ont au moins une relation de type idée associée (synonymie, hyperonymie, etc.). Des 19 037 lemmes de Manulex, 6 068 sont étiquetés comme polysémiques (31,2%). Nous avons utilisé cette ressource pour extraire des synonymes pour les mots de Manulex. Quant à BabelNet, il s'agit d'un réseau multilingue construit à partir de WordNet et Wikipédia. Nous avons utilisé les informations sur les 23 242 lemmes du français, en particulier le nombre de synsets associés.

4 Analyse de variables pour caractériser la complexité lexicale

4.1 Typologie

Dans cette section nous introduisons un ensemble de variables présentées ci-dessous. Nous mettons l'accent (en gras) sur les variables morphologiques et sémantiques, qui constituent les deux apports principaux de cet article par rapport à

4. http://www.lexique.org/
5. http://polymots.lif.univ-mrs.fr
6. À la base, Polymots ne contient pas des mots composés ('mainmise'), ni des noms avec tiret ('amour-propre'), des mots originaires d'autres langues ('mortadelle'), des mots grammaticaux ('tellement') ou encore des mots techniques ('dyoxide'). La création manuelle ainsi que l'écart de ces mots justifient sa faible couverture par rapport à Manulex
7. http://www.jeuxdemots.org/

des approches proches dans ce domaine (Gala *et al.*, 2013). C'est pourquoi, la section 4.2 est consacrée à la description détaillée de l'implémentation des variables morphologiques, les autres étant directement décrites dans la liste ci-dessous.

4.1.1 Critères orthographiques

1. *Nombre de lettres* : nombre de caractères alphabétiques dans un mot ;
2. *Nombre de phonèmes* : pour calculer le nombre de phonèmes dans un mot, un système mixte a été mis en place. Pour les mots présents dans Lexique3, nous avons simplement récupéré l'information issue de cette ressource. Pour les mots absents de Lexique3, nous avons généré leur représentation phonétique au vol via *eSpeak* [8] ;
3. *Nombre de syllabes* : comme pour le nombre de phonèmes, le nombre de syllabes d'un mot a soit été récupéré directement dans Lexique3, quand l'information était disponible, soit a été calculé automatiquement en deux étapes. Tout d'abord, la forme phonétique a été générée (comme au point précédent), avant d'y appliquer l'outil de syllabification de Pallier (1999) ;
4. *Voisinage orthographique* : les informations concernant le nombre ou la fréquence des voisins orthographiques [9] proviennent également de Lexique 3 et nous les avons déclinées en 3 variables : (4a) nombre de voisins, (4b) fréquence cumulée de tous les voisins, (4c) nombre des voisins les plus fréquents ;
5. *Cohérence phonème-graphie* : le nombre de phonèmes et de lettres dans un mot ont été comparés sur la base de la classification suivante : 0 pour l'absence de différence (c'est-à-dire, une transparence parfaite), par exemple *abruti* [abRyti] ; 1 pour une différence de 1 ou 2 caractères, par exemple *abriter* [abRite] ; 2 pour une différence supérieure à 2 caractères, par exemple dans *lentement* [l@t-m@] [10] ;
6. *Patrons orthographiques* : 5 variables ont été définies autour de la présence de graphèmes complexes dans les mots, à savoir (6a) des voyelles orales (par ex. 'au' [o]), (6b) des voyelles nasales (par ex ; 'in' [ɛ̃]), (6c) des doubles consonnes (par ex. 'pp'), (6d) des doubles voyelles (par ex. 'éé'), (6e) ou encore des digrammes (par ex. 'ch' [ʃ]) ;
7. *Structure syllabique* : trois niveaux de complexité pour les structures syllabiques présentes dans les mots ont été définis sur la base des fréquences de ces structures dans le corpus de parole « simple » Parkinson : (7a) les structures les plus fréquentes [11] (CYV, V, CVC, CV), (7b) les structures relativement fréquentes (CCVC, VCC, VC, YV, CVY, CYVC, CVCC, CCV), (7c) et les structures peu fréquentes (combinaisons de plusieurs consonnes, par exemple CCCVC) ;

4.1.2 Critères morphologiques

8. ***Nombre de morphèmes*** : nombre total de préfixes, suffixes et de bases dans le mot ;
9. ***Fréquence minimale des affixes (préfixes et suffixes)*** : nombre de mots différents (types) dans lesquels apparaît le préfixe / suffixe le moins fréquent ;
10. ***Fréquence moyenne des affixes (préfixes et suffixes)*** : moyenne des fréquences absolues des préfixes / suffixes ;
11. ***Préfixation*** : attestation ou non de la présence de préfixes ;
12. ***Suffixation*** : attestation ou non de la présence de suffixes ;
13. ***Composition*** : attestation ou non de la présence de deux bases ou plus ;
14. ***Taille de la famille morphologique*** : voir section 4.2 ;

4.1.3 Critères sémantiques

15. *Polysémie selon JeuxdeMots* : booléen indiquant si le mot est polysémique ou non ;
16. *Polysémie selon BabelNet* : nombre de synsets répertoriés dans BabelNet ;

8. http ://espeak.sourceforge.net
9. Les voisins orthographiques regroupent l'ensemble des mots de même longueur ne se différenciant que par une seule lettre tels que, pour SAGE, les mots MAGE, SALE, etc.).
10. La transcription est celle de Lexique 3 qui utilise l'alphabet SAMPA (*Speech Assessment Methods Phonetic Alphabet*).
11. La notation utilisée est la suivante : C pour consonne, V pour voyelle, Y pour les glides [j], [w] et [ɥ].

4.1.4 Critères statistiques

17. *Fréquence dans Lexique3* : logarithme des fréquences extraites de Lexique3 (calculées à partir d'un corpus de sous-titres de films). Pour traiter les mots absents de la ressource, nous avons appliqué un algorithme de lissage par Good-Turing (Gale & Sampson, 1995) afin d'attribuer une log-probabilité très petite par défaut à ces termes hors vocabulaire ;
18. *Présence/absence dans la liste de Gougenheim* : pour chaque mot, un booléen indique s'il appartient ou non à la liste du *Français Fondamental* dans sa version longue (qui comprend 8 875 lemmes). Comme il est bien connu en lisibilité que la taille de la liste de mots simples utilisée comme variable influe sur la capacité de discrimination de celle-ci, nous avons expérimenté avec diverses tailles de liste, par tranche de 1 000 mots, de 1 000 à 8 875 mots.

4.2 Obtention des variables par analyse morphologique

Les variables morphologiques 8 à 14 ont été obtenues par analyse morphologique non supervisée, en utilisant les systèmes décrits dans (Bernhard, 2006) et (Bernhard, 2010). Pour ce faire, les lemmes issus de Morphalou 2.0, Manulex, Polymots, corpus Parkinson et FLELex ont été fusionnés et ont été associés à leur fréquence dans le corpus `2010-wiki-1M` [12] du portail Wortschatz (Quasthoff *et al.*, 2006), qui contient 1 million de phrases issues de Wikipédia. Cette liste de lemmes associés à leur fréquence sert à l'apprentissage non supervisé d'informations morphologiques.

Le premier système (Bernhard, 2006) permet le découpage des mots en segments morphémiques étiquetés : base (b), préfixe (p), suffixe (s) et élément de liaison (l), comme par exemple `im_p + pens_b + able_s + ment_s`. La première étape du système consiste à extraire une liste de préfixes et de suffixes sur la base des probabilités transitionnelles entre sous-chaînes observées dans le lexique. Cette étape est contrôlée par un paramètre N qui détermine la quantité de préfixes et suffixes extraits : dans nos expériences, nous avons fixé ce paramètre à 5 et 10, pour tester son influence sur les variables utilisées dans le modèle final. La deuxième étape permet d'extraire une liste de bases candidates en utilisant les affixes obtenus à l'étape 1. Les deux dernières étapes consistent à segmenter les mots et à identifier la meilleure segmentation possible.

Ce système a été utilisé pour définir la valeur des variables suivantes : nombre de morphèmes, préfixation (oui/non), suffixation (oui/non), est composé (oui/non), fréquence minimale des préfixes, fréquence minimale des suffixes, fréquence moyenne des préfixes, fréquence moyenne des suffixes. Un mot est considéré comme composé s'il contient deux bases dans la segmentation.

Les résultats de ce premier système ont également permis de déterminer la taille de la famille morphologique : tous les mots qui contiennent la même base sont regroupés dans la même famille. Ainsi, la famille d'« impensablement » contient tous les mots contenant le segment `pens_b`. Pour les mots composés qui appartiennent à plusieurs familles, seule la taille de la famille la plus petite a été prise en compte.

Les deux derniers systèmes, MorphoClust et MorphoNet (Bernhard, 2010) produisent uniquement des familles morphologiques et n'ont donc été utilisés que pour obtenir la taille de la famille. MorphoClust forme des familles par classification ascendante hiérarchique. La méthode MorphoNet est quant à elle fondée sur la détection de communautés dans des réseaux lexicaux. MorphoClust utilise la même première étape que le système de segmentation et est donc dépendant du même paramètre N, encore une fois fixé à 5 et 10. Des expérimentations conduites précédemment ont montré que MorphoClust obtient un meilleur rappel, tandis que MorphoNet est plus précis (Bernhard, 2010).

5 Intégration dans un modèle pour prédire la difficulté

Après avoir identifié un ensemble de 49 variables lexicales (les 18 mentionnées ci-dessus et leurs variantes, c'est-à-dire, par exemple, la proportion d'absents calculée sur la base de listes de taille différentes), nous avons effectué deux séries d'expériences à partir de nos deux listes de mots gradués (Manulex et FLELex). Dans un premier temps, l'efficacité de ces variables a été évaluée par une analyse bivariée, c'est-à-dire sans tenir compte de la multicolinéarité [13] présente au sein des données. L'objectif de cette étape est simplement de déterminer quels sont, parmi nos prédicteurs, ceux qui apportent le

12. Pour les mots absents du corpus, la fréquence a été fixée à 1.
13. Il s'agit de la redondance informationelle partielle entre certaines variables. Par exemple, les mots plus courts sont aussi souvent les plus polysémiques et les plus fréquents.

plus d'information sur la difficulté des mots. Dans un second temps, l'ensemble des variables significativement associées à la difficulté lexicale ont été combinées au sein d'un algorithme d'apprentissage automatisé (SVM).

5.1 Analyse des variables

Comme nous l'avons rappelé aux sections 1 et 4, les variables sélectionnées pour cette étude l'ont été sur la base de résultats en psycholinguistique attestant d'un lien de causalité entre ces caractéristiques lexicales et la difficulté des mots mesurées au moyen de multiples tâches. De plus, (Gala *et al.*, 2013) ont analysé le comportement d'un sous-ensemble de nos variables en lien avec la parole "pathologique" via le corpus Parkinson décrit à la section 3.2. Cette étude a comparé les caractéristiques des mots utilisés dans un langage supposé plus "simple" avec un lexique général (Lexique3). Il est apparu que les unités lexicales "simples" étaient effectivement plus courtes (6,3 lettres, 4,7 phonèmes, 1,96 syllabes contre 8,6 lettres, 6,8 phonèmes et 2,89 syllabes dans Lexique3). Les structures syllabiques des ces mots, étaient également plus simples. Cela est une première piste qui confirmerait, sur un large ensemble de données réelles, la relation existante entre certaines de nos variables et la difficulté des mots.

Pour obtenir un diagnostic plus précis de l'efficacité de chaque variable, une analyse corrélationnelle a été effectuée. Pour chaque variable, la force de sa relation avec la difficulté lexicale a été mesurée à l'aide du coefficient de corrélation de Spearman [14]. Nous avons répété cette analyse sur nos deux lexiques de référence : Manulex et FLELex. Comme ces ressources ne comprennent pas les mêmes mots et ceux-ci ne sont pas classés de la même façon, cela devrait augmenter la robustesse de nos conclusions quant à l'efficacité de nos variables comme prédicteurs de la difficulté lexicale en général.

Le résumé des analyses corrélationnelles est présenté à la Table 2. Les niveaux utilisés pour cette analyse sont ceux obtenus via la première des trois techniques de transformation de la distribution d'un mot en un niveau (voir section 3.1), à savoir le premier niveau où apparaît ce mot. L'analyse corrélationnelle a également été effectuée pour les autres techniques, mais les corrélations obtenues étaient inférieures. Par exemple, pour Manulex, la corrélation entre la log-fréquence des mots dans Lexique3 et la difficulté tombe à $\rho = -0,40$ pour la moyenne et à $-0,36$ pour le quartile. Il semble donc que la meilleure mesure de la difficulté d'un mot parmi celles que nous avons testées sur la base d'une liste de mots gradués, soit la première apparition d'un mot dans un niveau donné.

Variables	**Manulex (ρ)**	**FLELex (ρ)**
17 Freq. Lex3	-0,51	-0.53
18 AbsGoug (5000)	-0,41	-0.46
18 AbsGoug (4000)	-0,41	-0.47
02 Nb. phon	0,30	0,27
15 Polysémie	-0,29	-0.38
01 Nb. lettres	0,27	0,25
03 Nb. syllables	0,27	0,26
4a Nb. voisin	-0,25	-0,23
4b Voisin freqcum	-0,25	-0,23
16 Synset BabelNet	-0,20	-0,19
6b Voy. Nasale	0,08	0,07
14 Taille famille (morphoclust_10)	-0,08	-0,05
11 Prefix (seg_10)	0.07	0,06
08 Nb_morphs (seg_10)	0,06	0,08
06 Patrons ortho (a-d)	0,05	0,06
10 Moyenne sufx. (freq_seg_10)	-0,05	0,02

TABLE 2 – Sélection des meilleurs variables

Les résultats de cette analyse corrélationnelle semblent assez clairs et stables. En effet, les corrélations obtenues sur Manulex sont très proches de celles obtenues sur FLELex. S'il est vrai qu'une partie non négligeable de mots se retrouvent dans les deux listes (mêmes observations), ceux-ci n'ont pas nécessairement été annotés de la même manière. On note que les meilleures variables sont celles basées sur la fréquence (17) et sur la présence dans une liste de mots simples (18) de taille moyenne (entre 4 000 et 5 000 mots). Il est intéressant de noter que l'appartenance des mots à une liste de mots

14. Le choix de privilégier ce coefficient par rapport au coefficient de Pearson, plus commun, est essentiellement motivé par le fait que la relation entre les variables linguistiques et la difficulté n'est pas nécessairement linéaire, comme discuté dans François (2011). La corrélation de Spearman est plus adaptée pour capturer des relations monotones croissantes.

simples est plus discriminant en FLE que pour le français L1, de même qu'une liste plus courte est préférable pour le FLE, comme cela avait été montré par François (2011).

Une seconde constatation d'intérêt est l'efficacité et la robustesse des variables classiques telles que le nombre de lettres (01) et de syllabes (03). Moins utilisé dans la littérature, le nombre de phonèmes dans un mot (02) apparaît tout aussi efficace. Enfin, les informations relatives au statut polysémique des mots apportent également de l'information qui semble utile pour expliquer la difficulté du lexique, que ce soit via une information binaire sur le statut polysémique des mots issue de JeuxDeMots (15) ou via le nombre de synsets repris dans BabelNet (16). À notre connaissance, la question de la polysémie n'a pas encore été prise en compte en lisibilité et sa performance constitue une bonne surprise, d'autant que l'information véhiculée par cette variable ne recouvre que faiblement l'effet de fréquence [15].

Enfin, d'autres variables sont également significativement corrélées à la difficulté, mais dans une moindre mesure. C'est le cas des informations concernant les voisins orthographiques (4a, 4b), les patrons orthographiques complexes (06), en particulier les voyelles nasales (6b), ou encore des informations morphologiques. A ce niveau, c'est surtout la taille de la famille (14), le nombre de morphèmes (08) et la présence d'un préfixe (11) qui apparaissent les plus utiles parmi l'ensemble de nos variables morphologiques. Signalons que la taille de la famille morphologique apparaissait déjà comme une variable significative pour (Schreuder & Baayen, 1997) dans une tâche de décision lexicale : plus la famille morphologique est grande, plus cela tend à induire un effet de facilitation. Toutefois, même si ces corrélations sont largement significatives (en terme de p-valeur) au vu du nombre élevé de données [16], on peut conclure que l'effet des variables morphologiques reste assez faible pour identifier la complexité d'un mot.

5.2 Un modèle pour prédire la difficulté du lexique

Au terme de l'analyse corrélationnelle, une sélection a été effectuée parmi nos 49 variables sur la base de deux critères. D'une part, les variables retenues pour la phase de modélisation devaient être significativement corrélées à la difficulté. D'autre part, lorsqu'un ensemble de prédicteurs constituaient des variantes d'une même information déterminée par un paramètre (par exemple, la taille de la liste de Gougenheim ou le paramètre N pour les variables morphologiques), seul celui présentant la corrélation la plus élevée a été retenu. Cela nous donne un total de 26 variables pour Manulex et de 24 variables pour FLELex.

Sur la base de ces deux ensembles de variables, deux modèles statistiques ont été entraînés, l'un en se basant sur les mots de Manulex et l'autre sur ceux de FLELex. Pour rappel, les annotations utilisées pour cet entraînement l'ont été au moyen de notre première méthode : le niveau de la première occurrence d'un mot. Cela nous donne pour Manulex : 5 863 lemmes pour le niveau CP, 4 023 lemmes pour le CE1 et 9 151 lemmes pour le cycle 3. En ce qui concerne FLELex, la répartition par niveau est la suivante : 4 142 lemmes pour le niveau A1, 2 735 pour A2, 4 002 pour B1, 1 312 pour B2, 1 672 pour C1 et seulement 501 lemmes spécifiques au niveau C2.

L'algorithme d'apprentissage automatisé utilisé est une machine à vecteurs de support ou SVM (Boser *et al.*, 1992). Lors d'expériences préliminaires, nous avons utilisé la librairie LibSVM (Chang & Lin, 2011) et avons comparé différents kernels (linéaire, RBF et polynomial). Les résultats étant relativement similaires, nous avons finalement privilégié le recours à LibLinear (Fan *et al.*, 2008), qui même si elle ne permet que d'utiliser un kernel linéaire, s'est révélée considérablement plus rapide sur nos données. Afin de limiter les effets de multicolinéarité et de surapprentissage au sein des données, nous avons centré et réduit les données et opté pour une méthode de régularisation "L2". Enfin, le méta-paramètre C (le coût) a été choisi par une exploration limitée de l'espace de valeurs entre 100 et $0,001$.

Les résultats obtenus par les deux modèles sont repris au Tableau 3. Leurs performances sont évaluées en terme d'exactitude des prédictions du modèle sur des données de test, une valeur qui a été estimée à l'aide d'un algorithme de validation croisée à cinq échantillons. Les résultats sont comparés à deux baselines : (1) le score qu'obtiendrait un modèle prédisant toujours la classe majoritaire dans les données et (2) un modèle qui ne se baserait que sur l'information de fréquence, qui est la caractéristique lexicale la plus communément associée à la difficulté des mots.

On peut noter que nos deux modèles se comportent nettement mieux qu'un modèle qui attribuerait toujours la classe majoritaire. Pour Manulex, le gain par rapport à cette méthode simpliste est de 15% d'exactitude, tandis qu'il est de 14% pour FLELex. Le gain est donc relativement similaire dans les deux cas, même si les performances absolues ne sont pas équivalentes eu égard à la différence de classes entre les deux listes. En effet, effectuer des prédictions parmi six classes constitue une tâche plus complexe que prédire parmi trois catégories. Par contre le gain de performance par rapport à

15. Pour Manulex, la corrélation entre le statut polysémique des mots d'après JeuxDeMots (15) et la log-fréquence (17) n'était que de $\rho = -0,19$.
16. Par exemple, le nombre de morphèmes présente un ρ de $0,06$, ce qui correspond encore à un $p < 0,001$.

Liste	Modèle	C	Exac.	Ecart-type
Manulex	Classe majoritaire	/	48%	/
	Baseline Fréq.	0,1	61%	0,4%
	Modèle	0,5	63%	0,7%
FLELex	Classe majoritaire	/	28,8%	/
	Baseline Fréq.	0,5	39%	0,8%
	Modèle	0,001	43%	0,5%

TABLE 3 – Performances des modèles sur les deux listes de mots

la seconde baseline, qui n'utilise que l'information fréquentielle, est faible : 2% pour Manulex et 4% pour FLELex. Ce résultat est relativement surprenant en regard du nombre de variables dans le modèle, mais semble en accord avec la situation observée dans le cadre de SemEval 2012 où, rappelons-le, un seul modèle avait réussi à battre la baseline fréquentielle.

6 Discussion

Afin de mieux appréhender les résultats obtenus, nous avons effectué une série d'expérimentations supplémentaires sur les données. Tout d'abord, cette tendance des performances à plafonner semble être partiellement liée au déséquilibre entre les différentes classes au sein de nos deux jeux de données. Nous avons donc effectué des expériences complémentaires afin de vérifier l'impact de la distribution des classes sur le résultat de l'apprentissage. Ces expériences ont été réalisées avec *Weka 3.7.10* [17] avec un modèle de régression logistique (*SimpleLogistic*) en utilisant une validation croisée à dix échantillons. Deux méthodes de ré-échantillonage ont été comparées. La première consiste à obtenir une distribution uniforme des classes, en réduisant le nombre d'instances de manière à obtenir un nombre d'instance par classes à peu près égal au nombre d'instances de la classe la moins représentée. La seconde réduit les données d'apprentissage au même nombre d'instances que la première, mais en conservant une distribution équivalente aux données initiales [18]. L'intérêt de ce second processus est de contrôler l'effet de la taille du jeu de données sur les performances, afin de mieux isoler l'effet lié au type de distribution.

Les résultats de ces expériences sont détaillés dans les Tables 4 et 5. Le modèle de régression logistique obtient quasiment les mêmes résultats que ceux obtenus avec SVM. Par ailleurs, les résultats obtenus pour les données échantillonnées avec la même distribution sont comparables à ceux obtenus sur les données complètes. Toutefois, le ré-échantillonnage avec une distribution uniforme des classes conduit à des résultats largement inférieurs, alors même que le nombre d'instances est identique à celui du ré-échantillonage avec la même distribution que les données initiales.

	Données complètes (19 037 instances)	Données échantillonnées (11 993 instances)	
		Même distribution	Distribution uniforme
Fréquence	61,2%	61,6%	51,2%
Attributs sélectionnés	62,7%	63,0%	53,7%
Tous les attributs	62,9%	63,0%	53,4%

TABLE 4 – Exactitude pour la régression logistique sur le corpus Manulex, avec et sans ré-échantillonage

	Données complètes (14 364 instances)	Données échantillonnées (2 872 instances)	
		Même distribution	Distribution uniforme
Fréquence	41,1%	40,8%	28,6%
Attributs sélectionnés	42,7%	42,4%	33,2%
Tous les attributs	43,1%	41,9%	32,6%

TABLE 5 – Exactitude pour la régression logistique sur le corpus FLELex, avec et sans ré-échantillonage

17. http://www.cs.waikato.ac.nz/ml/weka/index.html
18. Le ré-échantillonage a été réalisé avec le filtre *Resample* disponible dans Weka, avec un tirage sans remise.

Pour expliquer ces différences, il est nécessaire d'étudier les performances obtenues indépendamment pour chaque classe. Ces résultats, exprimés cette fois à l'aide de la F-mesure, sont présentés dans les Tables 6 et 7 pour le modèle utilisant les attributs sélectionnés. On constate à nouveau des F-mesures comparables sur les données complètes et sur les données ré-échantillonnées avec une distribution similaire. Dans les deux cas, les F-mesures sont nulles ou proches de zéro pour certaines classes, tandis qu'elles s'élèvent pour les classes les plus représentées dans les données d'apprentissage. Les résultats obtenus pour une distribution uniforme sont plus équilibrés, mais montrent également une tendance intéressante : les classes situées aux deux extrémités des échelles sont celles pour lesquelles les résultats sont les meilleurs. Ceci impliquerait que les mots très simples ou très complexes sont plus faciles à identifier que les mots situés au milieu de l'échelle. Une tendance similaire a déjà été signalée au niveau du texte dans des travaux précédents en lisibilité (François & Fairon, 2012).

	Données complètes (19 037 instances)	**Données échantillonnées (11 993 instances)**	
		Même distribution	**Distribution uniforme**
CP (5 863)	0,644	0,648	0,636
CE1 (4 023)	0,000	0,000	0,355
Cycle 3 (9 151)	0,731	0,734	0,588

TABLE 6 – F-mesure par classe avec les attributs sélectionnés pour le corpus Manulex, avec et sans ré-échantillonage.

	Données complètes (14 364 instances)	**Données échantillonnées (2 872 instances)**	
		Même distribution	**Distribution uniforme**
A1 (4 142)	0,628	0,631	0,572
A2 (2 735)	0,029	0,041	0,326
B1 (4 002)	0,492	0,493	0,123
B2 (1 312)	0,000	0,023	0,251
C1 (1 672)	0,089	0,098	0,298
C2 (501)	0,000	0,000	0,352

TABLE 7 – F-mesure par classe avec les attributs sélectionnés pour le corpus FLELex, avec et sans ré-échantillonage.

7 Conclusions

Nous avons présenté un modèle de la complexité lexicale reposant sur diverses variables intralexicales et statistiques. L'un des aspects innovants de notre approche est que le degré de complexité des mots a été déterminé à partir de la présence de ces mots dans des manuels destinés à des apprenants du français langue maternelle et étrangère. Cette façon de faire nous a permis de traiter un large ensemble de termes, à la différence d'approches plus classiques en psycholinguistique qui associent la difficulté des mots à la vitesse de réalisation de tâches telles que la décision lexicale et qui demandent dès lors de soumettre chaque mot à un ensemble de sujets.

Un second point intéressant de l'étude est la prise en compte combinée d'un large ensemble de variables, toutes justifiées par des résultats de travaux issus du domaine de la psycholinguistique. Parmi ces variables, certaines n'ont que très peu été étudiées en lien avec la prédiction de la difficulté des textes ou des mots. C'est le cas des variables morphologiques, des types de structure syllabiques, des patrons orthographiques ou encore du statut polysémique des mots (d'après JeuxDeMots et BabelNet). Toutes nos variables ont été analysées afin de déterminer celles qui présentent la plus forte corrélation avec la difficulté lexicale telle que définie dans nos deux ressources (Manulex et FLELex). Certaines de ces variables telles que la fréquence et le nombre de caractères sont bien connues et largement utilisées dans les modèles de lisibilité. Sans surprise, elles se sont révélées de bons prédicteurs, en particulier la fréquence, qui suffit pour classer correctement 61% des mots de Manulex et 39% de ceux de FLELex. Les variables morphologiques, l'une des principales voies d'exploration de cette article, ne présentent par contre qu'une faible relation avec la difficulté lexicale, ce qui pourrait être dû à leur acquisition de manière non supervisée. Enfin, nous avons testé deux variables sémantiques qui permettent d'identifier les mots polysémiques et qui présentent des corrélations significatives avec la difficulté du lexique.

Globalement, l'intégration de variables plus complexes que la simple fréquence n'apporte qu'une faible amélioration à notre modèle de prédiction de la complexité lexicale qui procède par apprentissage supervisé. Différentes explications

peuvent être avancées pour rendre compte de ce résultat : d'une part, les données d'apprentissage ne sont pas équilibrées, certaines classes étant bien moins représentées que d'autres dans nos données. Cependant, comme nous l'avons montré, le fait de rééquilibrer le jeu de données par ré-échantillonnage n'améliore pas les performances du modèle. Dès lors, ce qui pénalise le modèle et le pousse à se concentrer sur les classes les plus peuplées semble plutôt être son incapacité à discriminer entre les classes à l'aide des informations qui lui ont été fournies. Il faut donc postuler deux explications à cette situation : soit nos variables ne suffisent pas à expliquer la difficulté des mots, soit les données d'entraînement sont trop bruitées. Très probablement, ces deux phénomènes se combinent. D'une part, on observe que très peu de corrélations fort élevées au sein de notre ensemble de variables. D'autre part, la transformation des données d'origine vers un lexique classé selon des niveaux indépendants nécessite de faire des choix qui ne sont pas sans répercussion sur le modèle de classification. Sans compter que l'apparition d'un mot à un niveau précis d'un manuel reste dépendant des choix des éditeurs et de leur propre vision pédagogique.

Afin de poursuivre l'analyse et la prédiction de la complexité lexicale, nous envisageons plusieurs pistes. Tout d'abord, d'autres variables identifiées dans la littérature psycholinguistique pourraient encore faire l'objet d'expérimentations. Citons parmi celles-ci le caractère abstrait ou concret d'un mot, son voisinage phonologique, ou encore le niveau d'imageabilité. Ensuite, nous pourrions reproduire notre étude sur un lexique plus restreint, mais dont la difficulté aurait été contrôlée plus strictement, via des tests sur populations. L'objectif serait de déterminer dans quelle mesure le plafonnement des performances provient d'un manque informationnelle au sein du modèle ou du bruit dans le corpus. Enfin, une autre approche de la complexité lexicale est possible dans le cadre d'applications de simplification pour lesquelles il faut choisir le meilleur substitut. Il s'agirait de développer un modèle qui ne viserait plus à attribuer un niveau absolu à l'ensemble des mots d'un lexique, mais plutôt de comparer un ensemble restreint de mots sémantiquement liés (par exemple au sein d'un synset) afin de les trier en fonction de leur difficulté.

Références

BERNHARD D. (2006). Unsupervised morphological segmentation based on segment predictability and word segments alignment. In *Proceedings of the PASCAL Challenges Workshop on Unsupervised Segmentation of Words into Morphemes*, p. 19–23.

BERNHARD D. (2010). Apprentissage non supervisé de familles morphologiques : comparaison de méthodes et aspects multilingues. *Traitement Automatique des Langues*, **2**(51), pp. 11–39.

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

BLACHE P. (2011). A computational model for language complexity. In *1st Conference on Linguistics, Biology and Computational Science*, Tarragona, Spain.

BOSER B., GUYON I. & VAPNIK V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, p. 144–152.

BRYSBAERT M., LANGE M. & VAN WIJNENDAELE I. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition : Further evidence from the Dutch language. *European Journal of Cognitive Psychology*, **12**(1), 65–85.

CARROLL J., MINNEN G., CANNING Y., DEVLIN S. & TAIT J. (1998). Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*.

CHANDRASEKAR R., DORAN C. & SRINIVAS B. (1996). Motivations and methods for text simplification. In *16th conference on Computational linguistics*, p. 1041–1044.

CHANG C.-C. & LIN C.-J. (2011). Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, **2**(3), 27.

COLTHEART M., DAVELAAR E., JONASSON T. & BESNER D. (1977). Access to the internal lexicon. In *Attention and Performance VI*, p. 535–555, London : Academic Press.

CONSEIL DE L'EUROPE (2001). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Paris : Hatier.

DALE E. (1931). A comparison of two word lists. *Educational Research Bulletin*, **18**(10), 484–489.

FAN R.-E., CHANG K.-W., HSIEH C.-J., WANG X.-R. & LIN C.-J. (2008). Liblinear : A library for large linear classification. *The Journal of Machine Learning Research*, **9**, 1871–1874.

FERRAND L. (2007). *Psychologie cognitive de la lecture*. Bruxelles : De Boeck.

FRANÇOIS T. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère*. PhD thesis, Université Catholique de Louvain. Thesis Supervisors : Cédrick Fairon and Anne Catherine Simon.

FRANÇOIS T. & FAIRON C. (2012). An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP 2012)*, p. 466–477.

FRANÇOIS T., GALA N., WATRIN P. & FAIRON C. (2014). FLELex : a graded lexical resource for French foreign learners. In *Proceedings of International conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.

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 *E-lexicography in the 21st century : thinking outside the paper.*, Tallin, Estonia.

GALA N. & REY V. (2008). Polymots : une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques. In *TALN 2008, Conférence sur le Traitement Automatique des Langues Naturelles*, Avignon, France.

GALE W. & SAMPSON G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, **2**(3), 217–237.

GERNSBACHER M. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology : General*, **113**(2), 256–281.

GOUGENHEIM G. (1958). *Dictionnaire fondamental de la langue française*. Paris : Didier.

HOWES D. & SOLOMON R. (1951). Visual duration threshold as a function of word probability. *Journal of Experimental Psychology*, **41**(40), 1–4.

LAFOURCADE M. (2007). Making people play for Lexical Acquisition. In *Proc. SNLP 2007, 7th Symposium on Natural Language Processing.*, Pattaya, Thailande.

LAUFER B. (1997). *What's in a word that makes it hard or easy : Some intralexical factors that affect the learning of words.* Cambridge University Press.

LÉTÉ B., SPRENGER-CHAROLLES L. & COLÉ P. (2004). Manulex : A grade-level lexical database from French elementary-school readers. *Behavior Research Methods, Instruments and Computers*, **36**, 156–166.

MONSELL S. (1991). The nature and locus of word frequency effects in reading. In D. BESNER & G. HUMPHREYS, Eds., *Basic processes in reading : Visual word recognition*, p. 148–197. Hillsdale, NJ : Lawrence Erlbaum Associates Inc.

NAVIGLI R. & PONZETTO S. P. (2010). BabelNet : building a very large multilingual semantic network. In *48th annual meeting of the Association for Computational Llinguistics.*, p. 216–225, Uppsala, Suède.

NEW G. A., PALLIER C., FERRAND L. & MATOS R. (2001). Une base de données lexicales du français contemporain sur Internet : Lexique 3. *L'année psychologique*, **101**, 447–462.

PALLIER C. (1999). *Syllabation des représentations phonétiques de Brulex et de Lexique*. Rapport interne, Technical Report, update 2004. Lien : http ://www. pallier. org/ressources/syllabif/syllabation. pdf.

PINTO S., GHIO A., TESTON B. & VIALLET F. (2010). La dysarthrie au cours de la maladie de parkinson. histoire naturelle de ses composantes : dysphonie, dysprosodie et dysarthrie. *Revue Neurologique*, **166**(10), 800–810.

QUASTHOFF U., RICHTER M. & BIEMANN C. (2006). Corpus portal for search in monolingual corpora. In *Proceedings of the fifth international conference on Language Resources and Evaluation, LREC 2006*, p. 1799–1802, Genoa, Italy.

ROMARY L., SALMON-ALT S. & FRANCOPOULO G. (2004). Standards going concrete : from LMF to Morphalou. In *Workshop on Electronic Dictionaries, COLING, Conference on Computational Linguistics*, Geneva, Suisse.

SCHREUDER R. & BAAYEN H. (1997). How complex simplex words can be. *Journal of Memory and Language*, p. 118–139.

SPECIA L., JAUHAR S. K. & MIHALCEA R. (2012). Semeval-2012 task 1 : English lexical simplification. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), Montreal, Canada*.

THORNDIKE E. (1921). *The Teacher's Word Book*. New York : Teachers College.

VITU F., O'REGAN J. & MITTAU M. (1990). Optimal landing position in reading isolated words and continuous text. *Perception & Psychophysics*, **47**(6), 583–600.