Bleu, contusion, ecchymose : tri automatique de synonymes en fonction de leur difficulté de lecture et compréhension

Thomas François[1] Mokhtar B. Billami[2] Núria Gala[2] Delphine Bernhard[3]
(1) Chargé de recherche FNRS, CENTAL, IL&C, UCLouvain
(2) LIF-CNRS UMR 7279, Aix Marseille Université
(3) LiLPa - EA 1339, Université de Strasbourg
thomas.francois@uclouvain.be, mokhtar.billami@lif.univ-mrs.fr,
nuria.gala@lif.univ-mrs.fr, dbernhard@unistra.fr

RÉSUMÉ

La lisibilité d'un texte dépend fortement de la difficulté des unités lexicales qui le composent. La simplification lexicale vise ainsi à remplacer les termes complexes par des équivalents sémantiques plus simples à comprendre : par exemple, BLEU ('résultat d'un choc') est plus simple que CONTUSION ou ECCHYMOSE. Il est pour cela nécessaire de disposer de ressources qui listent des synonymes pour des sens donnés et les trient par ordre de difficulté. Cet article décrit une méthode pour constituer une ressource de ce type pour le français. Les listes de synonymes sont extraites de BabelNet et de JeuxDeMots, puis triées grâce à un algorithme statistique d'ordonnancement. Les résultats du tri sont évalués par rapport à 36 listes de synonymes ordonnées manuellement par quarante annotateurs.

ABSTRACT

Automatic ranking of synonyms according to their reading and comprehension difficulty

The readability of a text strongly depends on the individual difficulty of its lexical units. Lexical simplification consists in replacing complex terms by semantic equivalents which are easier to understand : for instance the French *bleu* (BRUISE, 'damage resulting from a blow') is easier than *contusion* or *ecchymose* (ECCHYMOSIS). For this task it is thus necessary to have resources that list synonyms for given senses and sort them according to their difficulty. This article describes a method for building such a resource for French. Synonym lists are extracted from BabelNet and JeuxDeMots, then sorted using a statistical ranking algorithm. The results of the ranking are evaluated against 36 lists of synonyms manually ordered by forty annotators.

MOTS-CLÉS : lisibilité, annotation sémantique, synonymes, prédiction de la difficulté lexicale, tri en niveaux de difficulté.

KEYWORDS: readability, semantic anotation, synonyms, word difficulty prediction, difficulty ranking.

1 Introduction

Identifier le niveau de lisibilité d'un texte a suscité de l'intérêt depuis longtemps dans le secteur de l'éducation, que ce soit pour des lecteurs en langue maternelle (L1) ou en langue étrangère ou seconde (L2). En effet, une telle connaissance permet de mieux associer textes et lecteurs et augmente les bénéfices de la pratique de la lecture. La technique classique pour évaluer la lisibilité d'un

texte consiste à utiliser une formule de lisibilité, calculée à partir d'un ensemble de caractéristiques textuelles supposées influencer le processus de lecture. L'une des formules de lisibilité les plus populaires est celle de Flesch (1948) et elle se caractérise par l'emploi de la régression linéaire et d'un nombre très restreint de prédicteurs (un de type lexical et un de type syntaxique).

Les modèles de lisibilité de cette nature sont intéressants pour des tâches pédagogiques de type « recherche d'information », dans lesquelles l'objectif est de trouver des textes adaptés à un lecteur donné (cf. Newbold *et al.*, 2010). De ce fait, le domaine de la lisibilité a connu des avancées importantes sous l'impulsion du traitement automatique des langues (TAL) [1]. Toutefois, ces modèles présentent une faiblesse : ils sont axés sur la génération d'un score global de lisibilité et sont donc de peu de secours lorsqu'il s'agit d'adapter un texte trop complexe. Une alternative consiste à chercher à identifier dans un texte les passages les plus difficiles. C'est l'un des objectifs du domaine de la simplification lexicale [2] : une phase d'identification des unités à simplifier est nécessaire afin de procéder à leur remplacement par un équivalent plus simple.

En lien avec cette dernière problématique, nous nous intéressons à la difficulté lexicale, qui s'est souvent révélée être l'un des meilleurs indices de la lisibilité textuelle. Déterminer le niveau de difficulté des mots d'un texte peut servir à estimer plus globalement sa lisibilité afin de le simplifier. De plus, la prédiction de la difficulté lexicale comme tâche en soi a fait l'objet d'un intérêt croissant depuis quelques années dans la communauté TAL (voir Section 2 pour des détails). Elle a généralement été abordée comme un problème de classification (Gala *et al.*, 2014; Shardlow, 2013) ou, plus rarement, comme un problème d'ordonnancement (Jauhar & Specia, 2012).

Dans cet article, nous nous inspirons de cette approche par ordonnancement afin de proposer une ressource lexicale pour le français où les synonymes sont (1) désambiguïsés, c'est-à-dire rassemblés par sens et (2) triés en fonction de leur difficulté. La notion de difficulté doit ici être comprise comme une valeur qui situe l'unité lexicale en question sur une échelle de complexité de lecture et de compréhension par rapport à des termes sémantiquement équivalents, par exemple : BLEU ('résultat d'un choc') par rapport à CONTUSION ou ECCHYMOSE.

Pour atteindre ce double objectif, après avoir présenté plus en détails la problématique de la prédiction de la difficulté du lexique à la section 2, nous décrivons à la section 3 la façon dont nous avons constitué cette ressource de synonymes, avant de détailler, à la section 4, le modèle statistique utilisé pour ordonner les mots de la ressource (il repose sur la prise en compte combinée d'un large ensemble de variables linguistiques et psycholinguistiques). Enfin, les performances de ce modèle sont évaluées et discutées à la section 5, à l'aide d'un jeu de données de référence obtenu grâce à une campagne d'annotation.

2 Travaux récents et problématique

La problématique de la prédiction automatisée de la difficulté du lexique est une tâche intéressante, non seulement en vue d'applications pédagogiques, mais aussi parce qu'elle constitue une approche holistique de questions théoriques largement explorées en psycholinguistique. Elle pose certains défis d'envergure, en particulier celui de sa mesure. En effet, s'il est évident que la complexité lexicale est liée à diverses caractéristiques du lexique (fréquence, longueur des mots, polysémie, etc.), elle

1. Pour une synthèse sur ces questions, consulter les travaux de Collins-Thompson (2014) ou de François (2015).
2. Consulter Siddharthan (2014) pour plus de détails sur ce domaine.

est également dépendante des caractéristiques de l'individu qui perçoit les mots (ex. : expertise en lecture, couverture du vocabulaire, niveau du développement conceptuel, connaissance du domaine traité, etc.), ce qui la rend difficile à capturer.

Pour l'instant, la tâche a surtout été abordée du point de vue de la complexité des mots. Ainsi, Jauhar & Specia (2012) proposent, dans le cadre de SemEval 2012, un système ordonnant des synonymes dans le but de sélectionner le meilleur candidat pour une simplification lexicale. Pour ce faire, ils se basent sur des fréquences lexicales, le nombre de syllabes, un modèle N-gramme, un modèle LSA, mais également sur des variables psycholinguistiques (imageabilité, âge d'acquisition, niveau de concrétude et familiarité). Ils obtiennnent un κ [3] de $0,496$ entre les prédictions de leur modèle et le classement de référence.

Plus récemment, la tâche de prédiction lexicale a été abordée comme un problème de classification. Il ne s'agit plus d'ordonner un ensemble de synonymes en fonction de leur difficulté, mais d'attribuer aux mots un niveau de difficulté, en référence à une échelle de difficulté. Shardlow (2013) propose ainsi un modèle de classification par séparateurs à vastes marges (SVM) basé sur quelques variables (fréquence, nombre de lettres, nombre de syllabes, nombre de sens, nombre de synonymes, etc.) qui vise à détecter les mots anglais complexes, en vue d'une tâche de substitution lexicale. Gala *et al.* (2014) enrichissent le jeu des variables linguistiques et psycholinguistiques considérées pour ce problème et développent un modèle SVM pour le français qui classe 62% des mots correctement parmi trois niveaux de difficulté. Pour l'espagnol, Baeza-Yates *et al.* (2015) cherchent à prédire la difficulté des mots à la lecture pour des enfants dyslexiques. Ils développent des variables qui visent à capturer des patrons orthographiques reconnus comme difficiles pour les dyslexiques. Leur modèle classe correctement 72,3% des mots parmi 2 niveaux (facile et complexe). Ces diverses tentatives sont confrontées au même problème, à savoir des performances assez moyennes et un faible gain par rapport à une baseline uniquement basée sur la fréquence lexicale.

Une alternative à la classification automatique consiste à construire un lexique gradué à partir d'un corpus de textes dont le niveau de difficulté est connu. Lété *et al.* (2004) ont ainsi proposé Manulex, une ressource qui décrit les distributions des mots du français sur trois niveaux du primaire (CP, CE1 et un niveau allant du CE2 à la CM2). François *et al.* (2014) ont appliqué la même technique sur des textes destinés à des apprenants de langue étrangère, mettant au point FLElex, qui classe les mots selon l'échelle du Cadre européen commun de référence pour les langues (CECR) (Conseil de l'Europe, 2001). Kidwell *et al.* (2009) ont élaboré une méthode statistique plus complexe qui donne automatiquement, sur un corpus de textes pédagogiques, une estimation de l'âge d'acquisition des mots. Enfin, Brooke *et al.* (2012) ont produit un lexique gradué à l'aide d'une méthode inspirée de la conception automatique de lexiques de polarité (Turney & Littman, 2003).

À la croisée de ces deux dernières approches, nous avons proposé dans Gala *et al.* (2013) une ressource graduée de synonymes appelée ReSyf. Nous avons repris le réseau de synonymes de JeuxDeMots (Lafourcade, 2007) et nous avons attribué à chaque mot l'un des trois niveaux de difficulté de Manulex. Pour graduer les mots absents de Manulex, un modèle de classification a été employé (Gala *et al.*, 2014). Dans une version postérieure de la ressource, un travail de désambiguïsation des synonymes par sens a été effectué (Gala *et al.*, 2015). ReSyf constitue, ainsi, un premier pas vers un lexique gradué de synonymes, utile notamment pour la substitution lexicale. En l'état actuel, cette ressource comporte cependant quelques défauts. Tout d'abord, le recours à l'échelle à trois niveaux de Manulex limite la finesse de discrimination des synonymes. Pour reprendre notre exemple, si BLEU se voit attribuer la classe 1 et ressort comme le synonyme le plus simple, CONTUSION et ECCHYMOSE

3. Les auteurs utilisent une variante du κ pour une tâche d'ordonnancement, qui est présentée dans Specia *et al.* (2012).

appartiennent tous les deux au niveau 3, sans qu'aucune distinction ne soit faite entre ces deux termes. Un second problème est que, pour une entrée donnée, ReSyf dispose d'une granularité trop fine de sens. Par exemple, pour SOURIS, il existe de nombreux sens, parmi lesquels 'espèce de petit rongeur', 'genre de rongeur' et 'rongeur'. Un tel niveau de précision dans la désambiguïsation sémantique ne nous semble pas souhaitable pour la ressource.

Dans cet article, afin de proposer une nouvelle version de ReSyf qui surmonte ces deux faiblesses, nous avons développé une méthode d'ordonnancement automatique de synonymes. Pour ce faire, nous avons tout d'abord constitué une liste de synonymes à partir de JeuxDeMots, mais aussi de BabelNet (Navigli & Ponzetto, 2012), où sont clairement distingués, pour une entrée donnée, les synonymes correspondant à ses principales acceptions (*cf.* section 3). Dans un second temps, nous avons entraîné un modèle statistique capable de trier une liste de synonymes du plus simple au plus compliqué en se basant sur un ensemble de variables linguistiques et psycholinguistiques (*cf.* section 4). Enfin, nous avons voulu confronter les performances de cet algorithme de tri par rapport à des jugements humains concernant la relative difficulté de synonymes (*cf.* section 5).

3 Données et ressources

Cette section présente le processus de constitution de la ressource lexicale de synonymes désambiguïsés. Les listes de termes que nous avons utilisées proviennent du réseau sémantique BabelNet [4] et du dictionnaire Diko [5] (Lafourcade, 2011, 221-223), issu de JeuxDeMots. Nous détaillons, dans cette section, les expériences menées afin d'obtenir un niveau de raffinement des sens optimal. En effet, l'un des obstacles majeurs de la désambiguïsation sémantique est la granularité fine des inventaires de sens (Navigli, 2009). Par exemple, dans WordNet (Miller *et al.*, 1990), les distinctions entre sens sont parfois difficiles à effectuer pour les annotateurs humains (Edmonds & Kilgarriff, 2002). Notre objectif est dès lors d'obtenir une ressource de synonymes pour le français qui soit caractérisée par une granularité sémantique plus optimale, car cela facilite alors le processus de distinction des sens en contexte.

Pour la construction d'une telle ressource, nous avons appliqué deux méthodes : la première repose sur l'utilisation des sens issus de BabelNet, tandis que la deuxième utilise les raffinements sémantiques spécifiés dans JeuxDeMots. Chaque sens ou raffinement sémantique est associé à un ensemble de synonymes, que nous appellerons par la suite un vecteur de synonymes [6]. Les sections 3.1 et 3.2 décrivent respectivement le processus de traitement des données selon l'une et l'autre méthode. La section 3.3 décrit, quant à elle, les données de la ressource produite.

3.1 Construction de la liste de synonymes à partir des sens de BabelNet

Cette première approche consiste à extraire le réseau synonymique à partir des sens de BabelNet. Cette ressource a été construite de manière automatique en reliant WordNet avec plusieurs ressources lexicales et encyclopédiques (Wikipedia, Wikidata, OmegaWiki, Wiktionary, Open Multilingual WordNet) et elle comprend l'ajout de traductions automatiques entre plusieurs langues. Face à cette

4. Nous utilisons la version 2.5.1, `http://babelnet.org/download`

5. `http://www.jeuxdemots.org/diko.php`

6. Signalons que nous nous intéressons à la représentation des sens sans tenir compte de la présence des entités nommées. Nous considérons un sens comme étant un concept.

masse d'information provenant de BabelNet, nous sommes confrontés à deux problèmes majeurs : (1) le bruit, à savoir la présence de mots techniques et de mots provenant d'une langue étrangère ; (2) la granularité de sens qui est trop fine.

Le tableau 1 liste le nombre de sens pour le français pour chacune des catégories grammaticales ouvertes (noms, verbes, adjectifs et adverbes) d'après BabelNet et JeuxDeMots. Dans ce tableau, les traductions automatiques provenant de WordNet et Wikipédia ne sont pas prises en compte. On observe que la classe des noms de BabelNet est très largement majoritaire ($\simeq 97\%$) et est près de 35 fois plus large que la classe des noms dans JeuxDeMots. Le tableau 2 décrit le nombre de mots monosémiques (*monos*) et polysémiques (*polys*) selon BabelNet. La classe des noms reste toujours majoritaire que le mot soit ambigu ou non ($\approx 84\%$ des mots polysémiques sont des noms).

Catégorie	BabelNet	JeuxDeMots
Noms	622 132	18 030
Verbes	8 050	6 819
Adjectifs	7 576	4 860
Adverbes	1 634	180
Total	639 392	29 889

TABLE 1: Nombre de sens de BabelNet et JeuxDeMots servant à construire notre ressource

Catégorie	Mots monos	Mots polys
Noms	551 365	30 167
Verbes	2 280	2 878
Adjectifs	3 954	2 272
Adverbes	893	690
Total	558 492	36 007

TABLE 2: Données de BabelNet pour le français sans tenir compte des entités nommées et des traductions automatiques provenant des sens de WordNet et des articles de Wikipédia

Afin de réduire la liste des mots-synonymes proposés par BabelNet, d'une part, nous ne tenons pas compte des traductions, et d'autre part, nous utilisons un filtrage sur la base des lemmes présents dans JeuxDeMots. Nous avons une préférence pour JeuxDeMots du fait de sa nature de jeu associatif (donc annoté par des humains).

Pour réduire le nombre de sens par entrée, nous avons opté pour l'utilisation de NASARI [7] (*a Novel Approach to a Semantically–Aware Representation of Items*) (Camacho-Collados *et al.*, 2015), décrite ci-dessous, afin de ne garder que des sens bien distincts, c'est-à-dire, dont la similarité entre sens est faible. Cette première approche produit une ressource lexicale de mots-synonymes regroupés en plusieurs sens dont le vocabulaire provient de JeuxDeMots et l'organisation des sens provient de BabelNet.

NASARI : Il s'agit d'une approche permettant la modélisation de concepts *via* l'attribution d'une représentation sémantique des sens de mots tout en se basant sur un modèle sémantique distributionnel. Ce modèle représente les items lexicaux (mots ou sens) comme des vecteurs dans un espace sémantique. Le calcul des pondérations dans ces vecteurs repose sur l'utilisation de la spécificité lexicale (Lafon, 1980), une mesure statistique utilisée principalement pour l'extraction de termes. NASARI utilise les correspondances (sens BabelNet, article Wikipédia) : les entrées de NASARI représentent l'identifiant d'un sens de BabelNet possédant une correspondance dans WordNet et le titre d'un article de Wikipédia s'il en existe un.
NASARI ne propose des vecteurs sémantiques que pour les noms. Pour ReSyf, la polysémie n'est décrite que pour cette catégorie grammaticale la plus largement couverte par BabelNet. Pour les autres catégories, nous ne gardons que les mots monosémiques. Nous utilisons NASARI avec le type de représentation à base de mots pour le calcul de la similarité sémantique entre sens. La similarité

7. http://lcl.uniroma1.it/nasari

sur laquelle nous nous basons pour la comparaison des vecteurs est *Weighted Overlap* (WO) (Pilehvar *et al.*, 2013). Nous avons préféré d'utiliser la mesure WO au lieu du *cosinus* en raison de la petite dimension dont tiennent compte les vecteurs. La mesure *cosinus* a tendance à retourner des scores relativement faibles lorsque les dimensions sont petites, contrairement à la mesure WO qui n'est pas affectée par le nombre de dimensions.

Filtrage des sens : Nous faisons d'abord un tri des sens du plus fort vers le plus faible. Le sens le plus fort est celui qui contient le plus grand nombre de connexions sémantiques dans le réseau. Une comparaison entre une paire de sens est effectuée. Si une similarité forte entre les deux sens existe, le plus fort est gardé et le plus faible est supprimé. Le seuil au-delà duquel une similarité est considérée comme forte est 0.5. La comparaison est effectuée par la suite sur une autre paire de sens et ainsi de suite jusqu'à l'obtention d'un ensemble de sens distincts. Nous avons choisi une suppression de sens afin de ne plus avoir une relation de parenté entre les sens d'un même mot. Nous n'avons pas pris la piste de regroupement de sens parce qu'en général les mots les plus techniques se trouvent dans les niveaux les plus profonds (sens possédant une faible connexion sémantique).

3.2 Construction de la liste de synonymes à partir des sens de JeuxDeMots

Cette deuxième approche se base, quant à elle, sur JeuxDeMots et tient compte des raffinements sémantiques, s'ils existent, présents dans la ressource. Comme cette dernière est en évolution constante et, qu'à l'heure actuelle, elle propose des synonymes pour les raffinements sémantiques, nous faisons une extraction directement des sens-synonymes. L'avantage de JeuxDeMots est qu'il permet d'avoir une représentation des différents sens d'un mot donné sous la forme d'un arbre (Lafourcade & Joubert, 2009), ce qui n'est pas le cas pour BabelNet. Cela nous permet ainsi d'identifier directement les sens les plus importants, situés au premier niveau de l'arbre. Par exemple, le mot BARRAGE possède 5 raffinements sémantiques : {'ouvrage d'art', 'tir de barrage', 'match de barrage', 'rocher', 'barrière'} ordonnés selon leur poids sémantique et le raffinement sémantique 'barrière 'possède lui-même un autre raffinement sémantique 'police' (Lafourcade, 2011, 125). Pour notre ressource, nous tenons compte seulement du premier niveau de raffinement sémantique lors de l'extraction des synonymes. Le tableau 1 présente le nombre de sens décrits dans JeuxDeMots. La classe des noms est la plus majoritaire ($\approx 60\%$).

Compte tenu de l'aspect associatif du réseau, et à ce jour, la relation de synonymie ne couvre pas tous les raffinements sémantiques. Nous utilisons deux stratégies différentes pour la prise en compte des synonymes : (1) nous nous référons seulement aux sens proposant des synonymes ; (2) pour les sens non couverts par la relation de synonymie, nous prenons les étiquettes des sens (le plus souvent des hyperonymes) comme synonymes. Pour l'exemple de BARRAGE, le sens 'tir de barrage' est le seul sens pour lequel on trouve des synonymes.

3.3 Données de la ressource lexicale

Les données de notre ressource sont obtenues suivant les deux méthodes décrites ci-dessus. La première tient compte des sens de BabelNet avec un vocabulaire de JeuxDeMots. La deuxième tient compte seulement de JeuxDeMots pour laquelle nous utilisons les deux stratégies présentées précédemment. Le tableau 3 décrit le nombre de mots retournés selon la méthode utilisée.

La ressource BabelNet∩JDM est celle décrite dans 3.1. La ressource JDM#Syns est celle décrite dans

POS	Ressources	Mots (sens=1)	Mots (sens>1)	Ensemble de mots
Noms	BabelNet∩JDM	17 017	4 309	**21 326**
	JDM#Syns	1 595	947	2 542
	JDM#Syns⊎Hypers	992	5 409	6 401
Verbes	BabelNet∩JDM	870	-	**870**
	JDM#Syns	558	276	834
	JDM#Syns⊎Hypers	1 512	1 982	3 494
Adjectifs	BabelNet∩JDM	1 377	-	**1 377**
	JDM#Syns	444	224	668
	JDM#Syns⊎Hypers	1 154	1 578	2 732
Adverbes	BabelNet∩JDM	395	-	**395**
	JDM#Syns	31	8	39
	JDM#Syns⊎Hypers	147	59	206
Total	BabelNet∩JDM	19 659	4 309	**23 968**
	JDM#Syns	2 628	1 455	4 083
	JDM#Syns⊎Hypers	3 805	9 028	12 833

TABLE 3: Description des données de la ressource construite à partir de BabelNet et JeuxDeMots

3.2 en tenant compte seulement des sens ayant des synonymes. La ressource JDM#Syns⊎Hypers est aussi décrite dans 3.2 en tenant compte de tous les raffinements sémantiques. La première méthode pour les mots ayant un seul sens couvre plus de mots que la deuxième méthode, sauf pour les verbes où JDM#Syns⊎Hypers est meilleure. Pour les mots ayant plus d'un sens, nous faisons une comparaison entre les deux méthodes seulement sur les noms vu la nature de NASARI. La méthode BabelNet∩JDM ne couvre pas plus de noms que JDM#Syns⊎Hypers par contre en terme de synonymes, nous obtenons des listes beaucoup plus importantes vu la nature de la méthode JDM#Syns⊎Hypers. Pour la totalité, nous nous retrouvons avec un nombre d'entrées beaucoup plus important (23 968) suivant la méthode BabelNet∩JDM contre 12 833 pour JDM#Syns⊎Hypers ou 4 083 pour JDM#Syns.

4 Méthode de ranking

Après avoir obtenu notre lexique, restait à en ordonner les vecteurs de synonymes en fonction de leur difficulté. Pour ce faire, nous utilisons un modèle d'ordonnancement, qui est régulièrement utilisé en recherche d'information pour trier les résultats d'une requête par ordre de pertinence. Ce type de modèle est bien documenté (Li, 2015) et se décline selon trois approches principales : *pointwise*, *pairwise* et *listwise*. Nous avons opté pour la seconde, et plus particulièrement, pour l'algorithme SVMRank (Herbrich *et al.*, 2000). L'entraînement de cet algorithme nécessite de disposer de données (ici, les mots) déjà triées ou associées à un niveau de difficulté, ainsi que de disposer de représentations de ces mots. Les caractéristiques linguistiques utilisées dans notre étude sont décrites dans la section 4.1. À partir de cette représentation vectorielle, on crée des paires d'entraînement en sélectionnant deux mots de difficulté différente et en fusionnant leurs vecteurs de caractéristiques (*cf.* section 4.2). La dernière étape consiste à optimiser le modèle et à choisir les meilleurs paramètres sur la base d'une évalution intrinsèque (*cf.* section 4.3).

4.1 Les données d'entraînement

Notre ressource vise avant tout des apprenants en milieu scolaire, nous avons donc opté pour la liste Manulex, décrite à la section 2, afin de disposer d'un ensemble de mots déjà annoté en fonction de leur difficulté. Manulex comprend 23 812 lemmes, mais nous n'avons conservé que les classes ouvertes (noms, adjectifs, adverbes et verbes), ce qui réduit le nombre de lemmes à 19 038. Chacun de ces lemmes est représenté sous la forme d'un vecteur de 69 caractéristiques, qui capturent diverses propriétés linguistiques et psycholinguistiques. Les variables, pour leur grande majorité, ont été proposées par Gala *et al.* (2014). Nous les rappellons brièvement ici :

— *Critères orthographiques* : le nombre de lettres par mot, le nombre de phonèmes par mot, le nombre de syllabes, des variables mesurant la densité et la fréquence du voisinage orthographique du mot cible, une mesure de la transparence entre phonèmes et graphèmes, la présence de certains patrons orthographiques plus complexes, et l'appartenance à une classe de structure syllabique plus ou moins fréquente.
— *Critères sémantiques* : une variable binaire indiquant si le mot est polysémique d'après JeuxDeMots et une variable discrète indiquant le nombre de synsets répertoriés dans BabelNet.
— *Critères fréquentiels* : logarithme de la fréquence du mot obtenue dans Lexique3 (New *et al.*, 2007) et présence du mot dans la liste de Gougenheim.
— *Variables morphologiques* : nombre de morphèmes, présence ou non de préfixes et de suffixes, fréquence minimale et moyenne des affixes, présence de plusieurs bases (cas des mots composés), taille de la famille morphologique. Nous avons également testé de nouvelles variables par rapport à Gala *et al.* (2014) : fréquence du mot le plus fréquent de la famille, fréquence moyenne des mots de la famille, fréquence cumulée dans la famille, sous forme de classe : petite, moyenne, grande, ou très grande.

4.2 Création des paires

La seconde étape a consisté en la préparation de paires d'entraînement, puisque nous avons adopté une approche *pairwise*. Pour deux mots w_i et w_j donnés, chacun associé à un niveau de difficulté (l_i ou l_j) et à un vecteur de caractéristiques ($\mathbf{v_i}$ ou $\mathbf{v_j}$), il s'agit de créer une paire $< w_i, w_j >$ associée à un nouveau vecteur $\mathbf{v_{ij}}$ issu de la combinaison des deux vecteurs $\mathbf{v_i}$ et $\mathbf{v_j}$. Il existe plusieurs méthodes pour ce faire, telles que la soustraction des deux vecteurs ($\mathbf{v_i} - \mathbf{v_j}$), leur rapport ($\frac{\mathbf{v_i}}{\mathbf{v_j}}$) ou leur concaténation ($\mathbf{v_i} \oplus \mathbf{v_j}$). Tanaka-Ishii *et al.* (2010) ayant montré que la soustraction produisait les meilleurs résultats pour une tâche d'ordonnancement de lisibilité de textes, nous avons également opté pour celle-ci.

En plus d'être associée à un nouveau vecteur de caractéristiques ($\mathbf{v_{ij}}$), notre paire doit aussi se voir attribuer un niveau unique (l_{ij}) obtenu en fonction des niveaux l_i et l_j des deux mots. Nous avons appliqué l'heuristique suivante : (1) Si $l_i > l_j$, alors $l_{ij} = 1$ et (2) si $l_i < l_j$, alors $l_{ij} = -1$. Autrement dit, si le niveau du premier mot est considéré comme supérieur à celui du second mot dans Manulex, on attribue la valeur 1 à la paire, tandis que c'est la valeur -1 qui est attribuée dans le cas inverse.

Une difficulté toutefois se pose lorsqu'on veut appliquer cette heuristique : Manulex décrit, pour chacun de ses 23 812 lemmes, une distribution de fréquence définie sur les trois niveaux. Il n'attribue donc pas à chaque mot un niveau de difficulté unique. C'est pourquoi, il a été nécessaire de transformer chaque distribution D en un niveau unique à l'aide d'une fonction $\phi(D)$. Pour définir cette fonction,

deux approches ont été testées. Dans la première, $\phi(D)$ renvoie simplement une valeur correspondant au premier des trois niveaux pour lequel la fréquence du mot n'est pas nulle (L). Nous appelerons le jeu d'entraînement qui en découle *Manulex-3N*. Cependant, comme notre algorithme de création de paires ignore les cas d'égalité (cad. quand $l_i - l_j = 0$), un grand nombre de données intéressantes sont dès lors ignorées. C'est pourquoi, nous avons également défini $\phi(D)$ pour qu'elle renvoie une valeur continue comprise entre 1 et 3, en utilisant la méthode décrite dans Gala *et al.* (2013) :

$$\phi(D) = L + e^{-r} \quad \text{où} \quad r = \frac{\sum_{l=1}^{L} U_l}{\sum_{l=L+1}^{3} U_l}$$

Le résultat de $\phi(D)$ est une valeur continue qui combine le premier niveau d'apparition du mot (L) à une quantité e^{-r} comprise entre 0 et 1. Cette quantité est définie en fonction du rapport entre la somme des effectifs des niveaux 1 à L et la somme des effectifs des niveaux $L+1$ à 3. Cette manière de faire permet de distinguer entre deux mots tels que POMME et CAMBRIOLEUR, qui apparaissent tout deux au niveau 1 ($L = 1$), mais 724 fois pour POMME contre 2 fois pour CAMBRIOLEUR. Il sera fait référence au jeu d'entraînement qui découle de cette seconde méthode comme *Manulex-Cont*.

Au terme de ce processus, il a été possible de créer les paires d'entraînement. Cependant, étant donné les plus de 19 000 mots de Manulex, le total des paires possibles dépassait les 360 millions de combinaisons. Nous avons donc opté pour un échantillonnage au hasard des paires, retenant 20 paires par mot. Cela donne un total de 238 728 paires pour *Manulex-3N* et de 291 263 paires pour *Manulex-Cont*.

4.3 Optimisation du modèle

La première étape de modélisation a été de sélectionner les meilleures variables parmi les 69. Pour ce faire, nous avons calculé la corrélation de Spearman entre ces variables et le niveau de difficulté des mots sur deux jeux de données : (1) sur les mots Manulex (avant la création des paires) et (2) sur les paires de *Manulex-3N*. Les corrélations les plus significatives sont reprises dans la table 4 :

Variables	**Manulex (ρ)**	**Paires (ρ)**
17 Freq. Lex3	-0,51	-0.57
18 AbsGoug (6000)	-0,41	-0.46
02 Nb. phon	0,30	0,35
15 Polysémie	-0,29	-0.33
01 Nb. lettres	0,27	0,32
03 Nb. syllables	0,27	0,32
4a Nb. voisins	-0,25	-0,23
15 Fréq. moyenne de la famille morpho.	-0,24	-0,27
15 Fréq. cumulée de la famille morpho.	-0,24	-0,27
15 Fréq. maximum de la famille morpho.	-0,24	-0,27
4b Voisin freqcum	-0,25	-0,23
16 Nombre de sens dans BabelNet	-0,20	-0,19

TABLE 4: Sélection des meilleures variables

Sans surprise, on retrouve des corrélations comparables à celles de Gala *et al.* (2014) pour les données de Manulex. Toutefois, nos nouvelles variables morphologiques basées sur la fréquence de la famille morphologique se démarquent des autres variables morphologiques par leur efficacité. Les corrélations

obtenues sur les paires sont quant à elles systématiquement supérieures à celles calculées sur les mots. Nous nous sommes donc basés sur les corrélations estimées sur les paires pour sélectionner un ensemble réduit de variables. Deux critères ont été utilisés à cette fin : (1) parmi les variantes d'une même variable (ex. longueurs différentes de la liste de Gougenheim), nous avons retenu celle qui avait la corrélation la plus élevée ; (2) parmi les variables significatives, seules celles ayant une corrélation supérieure à 0.09 ont été considérées, ce qui donne un ensemble de 21 variables.

Dans un second temps, nous avons entraîné des modèles SVM à noyau linéaire sur les deux jeux de données (*Manulex-3N* et *Manulex-Cont*). Pour chacun d'eux, une recherche par quadrillage a permis de sélectionner la meilleure valeur pour le méta-paramètre C. Ensuite, le nombre de paires bien classées (exactitude) a été estimée à l'aide d'une procédure par validation croisée à 10 plis. Les résultats obtenus par les deux modèles sont repris dans le tableau 5. Nous y rapportons également les performances des modèles intégrant l'ensemble des 69 variables, pour comparaison. On peut remarquer que le modèle à 21 variables obtient des performances comparables à celles du modèle à 69 variables sur les échantillons de tests, ce qui indique que notre heuristique de sélection de variables s'est révélée efficace.

Modèle	**C**	**21 var.**	**C**	**69 var.**
Manulex-3N	0,01	77,4%	0,01	77,8%
Manulex-Cont	0,01	72,4%	0,01	71,4%

TABLE 5: Exactitude des modèles de ranking

Par ailleurs, les modèles entraînés sur les niveaux définis sur la base de la première occurrence du mot (*Manulex-3N*) surpassent les modèles basés sur une approche continue des niveaux *Manulex-Cont*. Bien que cette définition de la fonction $\phi(D)$ soit intellectuellement peu satisfaisante, elle s'est révélée plus efficace et c'est donc le modèle qui a été retenu pour l'évaluation sur les données de test, décrite à la section suivante.

5 Évaluation

Cette dernière section évalue les performances du modèle sur un jeu de données différent et constitué par des évaluateurs humains. La section 5.1 décrit la façon dont ces données ont été collectées et évalue l'accord entre les juges. La section 5.2 rapporte, quant à elle, les performances du modèle d'ordonnancement sur ce jeu de données et discute les résultats obtenus.

5.1 Campagne d'annotation de synonymes en niveaux de difficulté

La synonymie est une relation lexicale sémantique d'équivalence entre signifiés. La synonymie exacte (ou absolue) étant rarissime, on considère comme synonymes deux unités lexicales ayant une « valeur sémantique suffisamment proche pour que l'une puisse être utilisée à la place de l'autre pour exprimer sensiblement la même chose. » (Polguère, 2002). Deux unités lexicales recouvrant (par inclusion ou intersection) la même notion sont donc des synonymes, par exemple BLEU et AZUR dans le sens 'couleur' ou BLEU, CONTUSION et ECCHYMOSE dans le sens 'résultat d'un choc'.

5.1.1 Annotation

Afin d'obtenir des données de référence pour évaluer notre modèle de tri, nous avons mené une campagne d'annotation dans laquelle nous avons demandé à des humains de classer des synonymes en fonction de leur difficulté de lecture et de compréhension.

Nous avons soumis à leur jugement quarante vecteurs de synonymes comportant chacun en moyenne 3,5 synonymes (pour un total de 150 unités lexicales) [8]. Chaque synonyme apparaissait hors-contexte et dans un ordre aléatoire. Pour chacun d'eux, il fallait lui attribuer un rang compris entre 1 et n, où n était le nombre de synonymes dans le vecteur. Les annotations ont été effectuées par quarante annotateurs (dont 28 francophones et 12 non-francophones ayant un niveau C1/C2 selon l'échelle du CECR).

5.1.2 Cohérence des annotations

Une fois les annotations collectées, nous les avons rassemblées afin d'obtenir une liste de référence. Lors de cette étape, nous avons écarté quatre vecteurs de synonymes de la liste initiale (pour un total de seize unités lexicales). Il s'agissait de cas où il y avait une égalité parfaite dans la graduation manuelle (vecteurs de deux synonymes : COUPE-VENT/ANORAK et RAPPEL/BIS), et de cas où plus de 30% des annotateurs considérait le terme comme une unité lexicale non pertinente dans la série (ex. CÉRÉBRAL dans la série OBSCUR/ÉSOTÉRIQUE), ou encore comme un terme inconnu (ex. IULE dans la série MILLEPATTES/MYRIAPODE).

Dans un second temps, la cohérence des annotations des 40 juges a été mesurée sur les 134 unités lexicales restantes, réparties en 36 sens distincts. Nous avons ainsi calculé, pour chaque vecteur de synonymes, l'alpha de Krippendorff (α) sur les 36 sens. Globalement, nous obtenons un accord inter-annotateur moyen de 0,4 (*cf.* lignes a. et d. de la table 6) [9]. Nous avons également évalué de façon distincte les vecteurs incluant 3 ou 5 synonymes (respectivement les lignes b./e. et c./f du tableau 6). Sans surprise, les résultats montrent que moins il y a de synonymes à annoter, plus l'accord inter-annotateur est élevé. Enfin, nous avons calculé l'α en tenant compte uniquement des locuteurs francophones (*cf.* lignes d. à f.) mais il semblerait que les différences provenant d'une maîtrise linguistique native du français *vs* une maîtrise L2 soient minimes.

	Nb sens	Annotateurs (Na)	Items (Ni)	Jugements (Na x Ni)	α
a.	36	40	139	5 560	**0,399**
b.	11	40	55	2 200	0,286
c.	10	40	30	1 200	**0,429**
d.	36	28	139	3 892	**0,412**
e.	11	28	55	1 540	0,358
f.	10	28	30	840	**0,419**

TABLE 6: Accord des 40 juges sur les 36 sens : α de Krippendorff.

8. La proportion des catégories grammaticales était : 53% noms, 23% verbes, 23% adjectifs, 1% adverbes.

9. Signalons que ce résultat, bien que non directement comparable, est situé dans le même ordre de grandeur que celui obtenu sur l'anglais, pour une tâche d'annotation comparable, dans SemEval 2012, à savoir un κ de 0,386 et 0,398 (Specia *et al.*, 2012).

5.2 Évaluation du modèle d'ordonnancement des synonymes

Lorsqu'on applique le modèle d'ordonnancement décrit à la section 4 sur ces données d'évaluation, les résultats nous apparaissent très satisfaisants. En effet, 83,33% des vecteurs sont triés de façon identique (BLEU, CONTUSION, ECCHYMOSE) ou à une distance d'un rang par rapport aux annotations des juges. Par exemple, le modèle trie MAIGRE, OSSEUX, SQUELETTIQUE, là où les annotateurs avaient majoritairement annoté MAIGRE, SQUELETTIQUE, OSSEUX. Seuls 16,67% des vecteurs comprennent une paire (ou plus) de mots inversés de plus d'un rang. Par exemple, BLEU, BIZUT, DÉBUTANT est l'ordonnancement prédit par le modèle, alors que les annotateurs avaient majoritairement proposé DÉBUTANT, BIZUT, BLEU.

En nombre de synonymes, 91,04% sont correctement triés ou inversés d'un rang et seuls 8,96% se sont vus attribuer des rangs d'une distance égale ou supérieure à deux par rapport à la référence. Parmi ces derniers, seuls 3 synonymes (2,24%) ont été classés avec une distance supérieure à deux. C'est le cas du vecteur DÉPOUILLER, APERCEVOIR, CONSTATER, DÉCELER, ANALYSER où le premier et le cinquième terme ont été intervertis par rapport aux annotations des juges, et du vecteur d'adjectifs MERVEILLEUX, FANTASTIQUE, FABULEUX, FORMIDABLE, SPLENDIDE où l'ordonnancement du modèle diffère pour les cinq éléments (les annotateurs avaient majoritairement proposé FABULEUX, FORMIDABLE, FANTASTIQUE, SPLENDIDE, MERVEILLEUX). Ce dernier cas illustre la difficulté de la tâche de tri pour des séries de synonymes que même les humains peinent à classer (α de Krippendorff = $0,04$) et où les termes présentent peu de différences de forme (nombre de syllabes, digraphes, etc.).

Parallèlement à cette analyse, nous avons utilisé deux mesures d'évaluation classiques pour ce type de tâche : (1) le κ de Cohen et (2) le rang réciproque moyen (MRR). Le κ est une mesure standard de l'accord inter-annotateur et nous avons utilisé ici sa version pondérée – car il s'agit de données ordinales – au moyen d'une fonction quadratique, qui pénalise davantage les inversions de rang plus importantes. Nous obtenons un κ de $0,63$, ce qui indique un accord fort (*substantial* d'après Artstein & Poesio (2008)) entre le modèle et les annotations humaines. Ce résultat apparaît, à première vue, meilleur que le κ obtenu par Jauhar & Specia (2012) sur l'anglais avec un modèle plus simple, bien que leur implémentation du kappa soit différente, puisqu'il s'agit de la version proposée par Callison-Burch *et al.* (2011). En ce qui concerne le MRR, notre modèle obtient le score de $0,84$ sur ce jeu d'évaluation. Ce résultat est très encourageant, car cela signifie que, dans la plupart des cas, notre modèle identifie correctement le synonyme considéré comme le plus simple par les annotateurs humains, ce qui est particulièrement utile pour la tâche de substitution lexicale.

6 Conclusion et perspectives

Dans cet article, nous avons présenté une méthode en deux temps pour constituer un lexique de synonymes désambiguïsés et triés du plus simple au plus difficile. D'une part, nous avons récupéré des données de ressources existantes et les avons nettoyées afin de conserver une granularité de sens optimale. D'autre part, nous avons proposé un modèle statistique qui trie des synonymes en fonction de leur difficulté, en se basant sur un ensemble de variables linguistiques et psycholinguistiques. Cet algorithme de tri obtient des résultats très satisfaisants sur des données annotées par des juges humains, ce qui ouvre la perspective d'intégrer notre ressource dans des applications de substitution lexicale et, de façon plus générale, dans des outils de simplification automatique de textes ou des logiciels d'entraînement ou d'assistance à la lecture [10].

10. La ressource finale, ainsi que la liste annotée par des juges, seront mises à disposition de la communauté en juin 2016.

Remerciements

Nous remercions les participants à la campagne d'annotation : étudiants et enseignants-chercheurs de l'université Grenoble Alpes, étudiants et enseignants-chercheurs d'Aix Marseille université. Nous remercions également Carlos Ramisch et Karën Fort pour leurs conseils avisés.

Références

ARTSTEIN R. & POESIO M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, **34**(4), 555–596.

BAEZA-YATES R., MAYO-CASADEMONT M. & RELLO L. (2015). Feasibility of word difficulty prediction. In *String Processing and Information Retrieval.*, p. 362–373. Springer.

BROOKE J., TSANG V., JACOB D., SHEIN F. & HIRST G. (2012). Building readability lexicons with unannotated corpora. In *Proceedings of the 1st Workshop on Predicting and Improving Text Readability for target reader populations*, p. 33–39 : Association for Computational Linguistics.

CALLISON-BURCH ., KOEHN P., MONZ C. & ZAIDAN O. F. (2011). Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, p. 22–64.

CAMACHO-COLLADOS J., PILEHVAR M. T. & NAVIGLI R. (2015). NASARI : a Novel Approach to a Semantically-Aware Representation of Items. In *Proceedings of the 2015 NAACL :HLT Conference*, p. 567–577, Denver, Colorado.

COLLINS-THOMPSON K. (2014). Computational assessment of text readability : A survey of current and future research. *International Journal of Applied Linguistics*, **165**(2), 97–135.

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

EDMONDS P. & KILGARRIFF A. (2002). Introduction to the special issue on evaluating word sense disambiguation systems. *Journal of Natural Language Engineering*, **8**(4), 279–291.

FLESCH R. (1948). A new readability yardstick. *Journal of Applied Psychology*, **31**(3), 221–233.

FRANÇOIS T. (2015). When readability meets computational linguistics : a new paradigm in readability. *Revue française de linguistique appliquée*, **20**(2), 79–97.

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

GALA N., BILLAMI M. B., FRANÇOIS T. & BERNHARD D. (2015). Graded lexicons : new resources for educational purposes and much more. In *22nd Computer-assisted language learning conference (EUROCALL-2015)*, p. 204–209, Padoue, Italie.

GALA N., FRANÇOIS T., BERNHARD D. & FAIRON C. (2014). Un modèle pour prédire la complexité lexicale et graduer les mots. In *Actes de TALN 2014*, Marseille.

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, Estonie.

HERBRICH R., GRAEPEL T. & OBERMAYER K. (2000). Large margin rank boundaries for ordinal regression. chapter 7, p. 115–132. Cambridge : MIT Press.

JAUHAR S. K. & SPECIA L. (2012). Uow-shef : Simplex–lexical simplicity ranking based on contextual and psycholinguistic features. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, p. 477–481.

KIDWELL P., LEBANON G. & COLLINS-THOMPSON K. (2009). Statistical estimation of word acquisition with application to readability prediction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 2, p. 900–909.

LAFON P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. volume 1 of *Mots*, p. 127–165.

LAFOURCADE M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07 : 7th International Symposium on NLP*, Pattaya, Chonburi, Thailande.

LAFOURCADE M. (2011). *Lexique et analyse sémantique de textes - structures, acquisitions, calculs, et jeux de mots.* Mémoire d'habilitation à diriger les recherches, Université Montpellier 2, LIRMM.

LAFOURCADE M. & JOUBERT A. (2009). Similitude entre les sens d'usage d'un terme dans un réseau lexical. *Traitement Automatique des Langues*, **50**(1), 179–200.

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.

LI H. (2015). *Learning to rank for information retrieval and natural language processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

MILLER G. A., BECKWITH R., FELLBAUM C., GROSS D. & MILLER K. (1990). Wordnet : An on-line lexical database. *International Journal of Lexicography*, **3**, 235–244.

NAVIGLI R. (2009). Word Sense Disambiguation : a survey. *ACM Computing Surveys*, **41**(2), 1–69.

NAVIGLI R. & PONZETTO S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**, 217–250.

NEW B., BRYSBAERT M., VERONIS J. & PALLIER C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, **28**(04), 661–677.

NEWBOLD N., MCLAUGHLIN H. & GILLAM L. (2010). Rank by readability : Document weighting for information retrieval. In *Advances in multidisciplinary retrieval*, p. 20–30. Springer.

PILEHVAR M. T., JURGENS D. & NAVIGLI R. (2013). Align, Disambiguate and Walk : a Unified Approach for Measuring Semantic Similarity. In *Proceedings of the 51st ACL Conference*, p. 1341–1351, Sofia, Bulgarie.

POLGUÈRE A. (2002). *Notions de base en lexicologie*. Montréal : Observatoire de Linguistique Sens-Texte, Université de Montréal.

SHARDLOW M. (2013). A comparison of techniques to automatically identify complex words. In *Proceedings of the ACL Student Research Workshop*, p. 103–109.

SIDDHARTHAN A. (2014). A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, **165**(2), 259–298.

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), Montréal, Canada*.

TANAKA-ISHII K., TEZUKA S. & TERADA H. (2010). Sorting texts by readability. *Computational Linguistics*, **36**(2), 203–227.

TURNEY P. D. & LITTMAN M. L. (2003). Measuring praise and criticism : Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, **21**(4), 315–346.