Analyse des noms agentifs dans les espaces vectoriels distributionnels

Marine Wauquier

Analyse des noms agentifs dans les espaces vectoriels distributionnels

Marine Wauquier[1]
(1) CLLE, CNRS & Université de Toulouse, 5 Allées Antonio Machado, 31058 Toulouse, France
marine.wauquier@univ-tlse2.fr

RÉSUMÉ

Notre étude s'inscrit dans le cadre d'une thèse ayant pour but d'exploiter les modèles distributionnels pour décrire sémantiquement des classes de mots définies selon des critères morphologiques. Nous utilisons des indices morphologiques et formels fournis par une base lexicale pour cibler les noms agentifs déverbaux construits par suffixation en *-eur*. Nous montrons qu'il est possible de constituer un représentant prototypique de la classe sémantique des noms agentifs en *-eur* dans les modèles distributionnels. L'étude de ce représentant met en évidence que l'information sémantique véhiculée par le suffixe varie en fonction du corpus d'étude et du degré de lexicalisation des dérivés.

ABSTRACT

Analysis of agent nouns in vector space models.

This experiment is part of a PhD thesis, the purpose of which is the semantic study through vector space models of word classes defined according to morphological criteria. We make use of morphological and formal clues given by a lexical database in order to target deverbal agent nouns derived with the french suffix *-eur*. We show that it is possible to build a prototypical representative of the agent noun semantic class in VSMs. The analysis of this representative highlights that the semantic instruction conveyed by the suffix varies depending on the corpus studied and the degree of lexicalization of the derivatives.

MOTS-CLÉS : Sémantique distributionnelle, sémantique lexicale, morphologie, linguistique de corpus.

KEYWORDS: Distributional semantics, lexical semantics, morphology, corpora linguistics.

1 Introduction

Dans le cadre de notre thèse, nous nous intéressons à la confrontation des modèles distributionnels à des catégories sémantiques établies selon des critères linguistiques, et en particulier morphologiques (noms d'agent, noms d'action, etc.) : notre objectif est d'utiliser les espaces vectoriels pour explorer la sémantique de ces catégories. Les outils distributionnels sont arrivés à maturité (Fabre & Lenci, 2015) et sont plébiscités pour leurs performances dans diverses tâches linguistiques (Kulkarni *et al.*, 2015; Verhoeven *et al.*, 2012). Mais les travaux se proposant de comparer sémantiquement des mots construits morphologiquement à l'aide d'outils d'analyse distributionnelle automatique sont à notre connaissance encore peu nombreux. Des travaux se sont par exemple penchsés sur l'étude des dérivés masculins et féminins en allemand, et ont montré que le genre des dérivés a un impact sur la distance distributionnelle (Zeller *et al.*, 2014). D'autres travaux ont quant à eux cherché à comparer

des procédés de nominalisation concurrents en allemand et ont montré qu'ils étaient distincts sur le plan distributionnel (Varvara *et al.*, 2016). Enfin, des travaux ont pour leur part exploité des indices distributionnels pour l'entraînement de classifieurs automatiques visant à distinguer les lectures événementielles des noms d'action polysémiques anglais en *-ment* (Lapesa *et al.*, a paraître).

Dans l'expérimentation que nous présentons plus spécifiquement dans cet article, nous nous intéressons aux noms agentifs déverbaux en *-eur* comme *chanteur* et *détecteur*, respectivement dérivés de *chanter* et *détecter*. Nous utilisons une ressource morphologique dérivationnelle, Lexeur, pour cibler ces déverbaux. Nous accédons à l'information sémantique véhiculée par ces déverbaux dans les corpus, et nous évaluons l'influence de deux paramètres, le corpus et le degré de lexicalisation (ou au contraire de nouveauté) du mot dérivé, sur la représentation distributionnelle de cette classe sémantique. Nous passons pour cela par une représentation prototypique que nous calculons à partir des membres de la classe sémantique. Nous montrons que le représentant prototypique de la classe des noms déverbaux en *-eur* que nous construisons véhicule un sens agentif, dans une acception relativement large, qui varie en fonction du corpus choisi et du niveau de lexicalisation de la classe considérée.

Nous détaillons dans un premier temps le dispositif expérimental que nous utilisons dans le cadre de cette étude (section 2). Nous présentons ensuite la démarche que nous avons suivie pour représenter sur le plan distributionnel la classe sémantique des agentifs déverbaux en *-eur* et les premières observations que l'on en tire, en nous penchant notamment sur la variation du premier paramètre que représentent les corpus (section 3). Nous évaluons ensuite l'influence de la lexicalisation en reprenant l'expérimentation avec de nouveaux paramètres (section 4). Enfin, nous discutons de ces résultats et nous évoquons les pistes de travail qu'ils ébauchent (section 5).

2 Dispositif expérimental

Nous faisons le choix d'utiliser conjointement un outil de sémantique distributionnelle, Word2Vec (Mikolov *et al.*, 2013), et une ressource dérivationnelle, Lexeur. L'utilisation de Lexeur permet de croiser des connaissances expertes validées par des linguistes avec les représentations sémantiques fournies par Word2Vec. Cette ressource nous permet en effet d'exploiter le critère morphologique comme indicateur de la classe sémantique, information absente de la représentation vectorielle, et de contrôler la nature des unités lexicales que nous considérons dans la première partie de notre expérimentation.

2.1 Ressource lexicale

Lexeur [1] regroupe 5 974 noms agentifs en *-eur* ainsi qu'une partie de leur famille dérivationnelle. Grâce à une procédure d'annotation manuelle, chaque nom agentif en *-eur* a été associé à son équivalent féminin en *-euse* et/ou en *-rice*, à sa base (*Vb* verbale ou *Nb* nominale) et à une liste de noms processifs identifiés. Ces noms sont issus du *Trésor de la Langue Française* (désormais *TLFi*), et complétés par des attestations issues du web. Chaque lexème est étiqueté morphosyntaxiquement. Nous utilisons le terme de noms agentifs dans une acception relativement large, puisque Lexeur regroupe indistinctement des noms d'agent (*abatteur*) et des noms d'instrument (*abatteuse*). De la

1. Cette ressource a été constituée par l'équipe CLLE-ERSS(Hathout & Fabre, 2002). En attendant sa mise en ligne sur le site REDAC, elle sera envoyée à toute personne sur simple demande à nabil.hathout@univ-tlse2.fr.

même façon, Lexeur regroupe sans distinction des noms d'action à l'interprétation événementielle (*abattage*), résultative (*abattis*), ou encore stative (*abattement*). Les lexèmes intégrés à la ressource présentent en outre divers degrés de polysémie (*sculpture* est à la fois une activité et un résultat). Toutes les familles ne sont pas homogènes. Le tableau 1 illustre la variété des entrées de Lexeur.

Nom d'agent masculin	**Nom d'agent féminin**	**Base**	**Cat.**	**Autres dérivés**
abatteur/Ncms	abatteuse/Ncfs	abattre/Vmn	Vb	abat/Ncms abattement/Ncms abatture/Ncfs abattage/Ncms abattis/Ncms
endoscopeur/Ncms	endoscopeuse/Ncfs	Ø	Ø	endoscopie/Ncfs
sculpteur/Ncms	sculpteuse/Ncfs sculptrice/Ncfs	sculpter/Vmn	Vb	sculpture/Ncfs ; sculptage/Ncms
wheeleur/Ncms	wheeleuse/Ncfs	wheel/Ncms	Nb	Ø

TABLE 1 – Extrait de Lexeur

L'ensemble des entrées comporte un nom agentif féminin en *-euse* et/ou en *-rice*, mais l'on peut noter que le suffixe *-euse* est surreprésenté (4 542 contre 1 514). 22% des familles sont dépourvues de noms processifs, à l'image de *wheeleur*. Pour les autres, on dénombre en moyenne 1,47 nom d'action par famille, le nombre de noms d'action par entrée variant entre 1 et 8. Nous nous concentrons dans cette étude sur les familles dont la base est verbale, ce qui représente 78% des entrées de Lexeur.

2.2 Corpus

Nos observations sont issues pour l'essentiel du corpus *Wikipédia*, constitué de la version française de 2013 de l'encyclopédie en ligne du même nom. Il compte environ 255 millions de mots. Du fait de son caractère encyclopédique, ce corpus couvre des domaines très divers, du plus général au plus technique, et présente donc un vocabulaire varié. Il est de fait particulièrement adapté à notre ressource qui regroupe des lexèmes à la technicité tout aussi variable (*chanteur*, *extérorécepteur*).

Ces observations sont confrontées à deux corpus aux genres textuels et à la taille variables, *LM10* et *frWaC*, pour tester la stabilité de nos observations. Le corpus *LM10* est constitué des articles du journal *Le Monde* publiés entre 1991 et 2000. Il contient environ 200 millions de mots. Le corpus *frWaC* est quant à lui composé de pages de sites internet appartenant au domaine .fr, pour un total de 1,3 milliard de mots. Nous avons choisi ce dernier corpus pour sa taille et pour les usages plus récents qu'il propose.

Les trois corpus ont été au préalable lemmatisés grâce à l'analyseur syntaxique Talismane (Urieli, 2013). Lorsque Talismane ne parvient pas à identifier le lemme d'une forme donnée, la forme est conservée. Cela explique la présence de certaines formes fléchies dans les résultats que nous présentons dans la suite de cette étude.

2.3 Espace vectoriel

Nous construisons les représentations distributionnelles des mots à l'aide de Word2Vec (Mikolov *et al.*, 2013). Nous nous basons dans cette étude sur l'examen de cooccurrences lexicales dans une fenêtre de 5 mots, sans prise en compte des relations syntaxiques. Word2Vec calcule le score de proximité de deux vecteurs, compris entre 0 (proximité nulle) et 1 (proximité maximale), sur la base du cosinus des vecteurs. Nous faisons le choix d'utiliser les paramètres par défaut [2] de Word2Vec, à savoir l'architecture CBOW, l'algorithme d'entraînement Negative Sampling, un seuil minimum de fréquence de 5, un seuil de sous-échantillonnage des mots fréquents de 10^{-3}, une taille de fenêtre maximale de 5, et comme nombre de dimensions des vecteurs 100. Nous construisons une matrice par corpus. Du fait des paramètres de l'outil, les mots issus de Lexeur et représentés dans les modèles vont donc varier en fonction du corpus. Le tableau 2 donne le nombre de noms agentifs présents dans les modèles.

	Wikipédia	***LM10***	***frWaC***
Masculin	1 334	1 147	1 444
Féminin	329	220	475

TABLE 2 – Nombre de noms agentifs issus de Lexeur et pris en compte en fonction du suffixe de constitution et du corpus

2.4 Construction d'un vecteur moyen

Notre objectif est de représenter de façon prototypique la classe des noms agentifs dans un corpus donné. Nous considérons ici la notion de prototype dans l'idée d'une catégorisation graduelle (Kleiber, 1990) : l'appartenance d'un élément à la catégorie est estimée selon son degré de ressemblance avec l'élément prototypique.

Nous abordons la notion de dérivé prototypique par le biais du dérivé moyen, dont nous définissons la représentation comme étant la moyenne des représentations des mots formés à partir de ce suffixe. Nous traduisons cela sur le plan vectoriel par la construction d'un barycentre, approche notamment utilisée dans des travaux sur la prédication verbale (Kintsch, 2001). Cela revient donc à calculer le vecteur $\overrightarrow{SUFF}$ du dérivé prototypique d'un suffixe *suff* comme la moyenne des vecteurs $\overrightarrow{Nsuff_i}$ des mots porteurs du suffixe tel qu'illustré en (1).

$$\overrightarrow{SUFF} = \frac{\sum_{i=1}^{n} \overrightarrow{Nsuff_i}}{n} \quad (1)$$

Nous ne prenons en compte que les vecteurs des mots présents dans Lexeur pour éviter de considérer des mots porteurs du suffixe correspondant mais n'appartenant pas à la catégorie sémantique visée (comme *fleur* pour *-eur*).

2. En réponse à la remarque d'un relecteur, nous discutons ce choix dans la section 5.

3 Représentation prototypique des noms agentifs en *-eur*

Le vecteur que nous construisons est abstrait car il agrège les informations des vecteurs qu'il moyenne. Il n'est pas la représentation d'un mot instancié dans le corpus. La question de son interprétation et de l'accès au sens qu'il véhicule se pose. À ce stade de la thèse, nous faisons le choix de passer par l'analyse des voisins distributionnels de ce vecteur théorique pour appréhender l'information sémantique qu'il véhicule. Nous favorisons une interprétation de nature qualitative, en nous limitant aux 50 premiers voisins. Pour le suffixe *-eur*, le 50 voisins les plus proches du dérivé prototypique calculé pour chacun de nos corpus sont présentés dans le tableau 3.

Wikipédia	réparateur - sèche-cheveux - soudeur - armurier - minuteur - wattman - conducteur - laborantin - machiniste - mécanicien - plombier - tournevis - stéthoscope - client - ventilateur - treuil - allumeur - mécano - coursier - déménageur - manomètre - aspirateur - soigneur - extincteur - vendeur - installateur - toiletteur - mélangeur - cric - ampèremètre - goniomètre - débogueur - technicien - ramasse-miettes - contacteur - descendeur - dépresseur - tune-o-matic - leurre - télérupteur - coupe-ongles - égoutier - microphone - juge-arbitre - opticien - nettoyeur - adaptateur - grappin - détecteur - ordinateur
LM10	ramoneur - bricoleur - toqué - alchimiste - chiot - nounours - moujik - magicien - ornithologue - matou - dragueur - bidouilleur - tâcheron - ludion - garagiste - fêlé - cinglé - comparse - imitateur - frelon - aventurier - coursier - barman - croque-mort - garnement - bouledogue - loubard - charretier - gandin - fripon - baroudeur - rouquin - coifeur - julot - boxeur - arnaqueur - malfrat - voyou - écuyer - prestidigitateur - moussaillon - cuisinier - sarret - puncheur - fêtard - camelot - afabulateur - braconnier - canari - garçon
frWaC	guindeau - perceur - surgrip - ferrailleur - rectifieur - multitours - coupe-papier - suiveur - basculeur - rilsan - servo-moteur - coupe-circuit - r-core - tromblon - palpeur - capsuleuses - accessoiriste - coupe-cigares - coutelas - marteau-pilon - talkie-walkie - rabot - cintreuses - pantographe - soudures - filin - emballeur - aspirateurs - petzl - montiss - cartillier - cintrier - batteur - mandrin - tuyautage - warbird - grignoteur - graisseur - perceuses - humbucker - elevateurs - gehennas - incorporateur - sebicape - dessouder - agitateur - encliquetables - avant-train - haute-pression - cymbales

TABLE 3 – 50 premiers voisins des dérivés prototypiques agentifs masculins en *-eur* dans les corpus *Wikipédia*, *LM10* et *frWaC*

La morphologie dérivationnelles se caractérise par une relation étroite entre la forme et le sens. Ces listes de voisins nous livrent des informations relatives à ces deux aspects.

Sur le plan sémantique, si l'on considère tout d'abord le corpus *Wikipédia*, on constate que les voisins du dérivé prototypique en *-eur* sont des noms de métier (*réparateur*, *armurier*), à raison de 44%, ou des noms d'instrument (*minuteur*, *coupe-ongles*), à raison de 46%. Dans le cas du corpus *LM10*, les 50 premiers voisins du dérivé prototypique en *-eur* sont majoritairement agentifs, à raison de 82% de noms de métiers (*ramoneur*, *charretier*) ou de noms d'humains (*garçon*, *comparse*). On retrouve par ailleurs 12% de noms d'animaux (*chiot*, *canari*). Enfin, parmi les 6% de voisins restants, on retrouve 2 noms propres, *sarret* et *camelot*, et un nom d'artefact ayant quasiment un statut d'instrument,

ludion. Pour ces 2 corpus, le dérivé prototypique semble donc bien capter le sens d'agentivité lié à la suffixation en *-eur*. On notera cependant que le passage d'un corpus encyclopédique à un corpus journalistique semble jouer sur la représentation de l'agentivité, perdant sa facette instrumentale.

Les résultats sont plus variés, et plus difficiles à interpréter, dans le cas du corpus *frWaC*. On est frappé par le caractère très spécifique du vocabulaire mis au jour : il s'agit de formes rares à la fréquence en moyenne plus faible (cf. tableau 5), pour certaines d'entre elles fléchies. On dénombre sur les 50 premiers voisins 52% d'instruments (*guindeau*, *coupe-cigares*), 16% d'agents (*ferrailleur*, *emballeur*) et 10% de noms à la double interprétation agentive et instrumentale (*batteur*). Parmi les 22% restants, on retrouve 4 entités nommées, dont 2 noms de marques d'outillage (*Petzl* et *Montiss*) et 2 noms propres (*Cartillier* et *Gehennas*). On constate par ailleurs la présence de 2 adjectifs (*encliquetables*, *haute-pression*), de 4 artefacts n'ayant pas clairement un statut d'instrument (*surgrip*, *soudures*) et d'un verbe (*dessouder*). On retrouve donc le sens d'agent et d'instrument que l'on avait déjà pour le corpus *Wikipédia*, mais de façon moins marquée.

Sur le plan formel, on constate que 56% des voisins du dérivé moyen ne sont pas suffixés en *-eur* dans le corpus *Wikipédia*. La tendance se confirme dans les corpus *LM10* et *frWaC* puisque respectivement 78% et 50% des voisins ne sont pas porteurs du suffixe *-eur*. On retrouve ainsi par exemple des dérivés suffixés en *-mètre*, *-ier*, ou encore *-ien*, trois suffixes qui forment néanmoins eux aussi des noms d'agent ou d'instrument.

Nous donnons à titre de comparaison dans le tableau 4 les 50 premiers voisins des vecteurs construits de la même façon mais à partir des déverbaux issus de la colonne « Nom d'agent féminin » de Lexeur (cf. tableau 1).

L'analyse des 50 premiers voisins du dérivé agentif féminin prototypique en fonction des corpus montre que les trois vecteurs captent une nouvelle fois une notion d'agentivité. En effet, les voisins obtenus pour les trois corpus sont des noms de métier (*coiffeuse*, *stripteaseuse*), des mots rattachés à des sujets culturellement associés à la femme (*stiletto*, *manucure*) ou des patronymes de femmes (*Herzigova*, *Sorokina*). Cette agentivité incorpore ici des informations relatives au genre féminin, notamment une connotation négative (Wauquier *et al.*, 2018), et est bien distincte de l'agentivité observée pour les agentifs en *-eur*. On observe par ailleurs de nouveau une différenciation entre le corpus journalistique et les deux autres corpus. Les voisins obtenus pour les corpus *Wikipédia* et *frWaC* sont pour beaucoup des entités nommées, quand le corpus *LM10* en est exempt. L'analyse des voisins obtenus pour *frWaC* est une nouvelle fois rendue plus difficile du fait de la faible fréquence des formes liées au féminin, plus nette encore dans ce corpus (cf. tableau 5).

Ces résultats montrent une certaine hétérogénéité de la classe sémantique des noms agentifs en *-eur*. On y distingue ainsi plusieurs sous-classes déjà connues (Huyghe & Tribout, 2015), comme les noms de métiers, les noms d'instruments, ou encore des noms référentiels, mais aux comportements bien distincts sur le plan distributionnel.

4 Impact de la lexicalisation

La ressource Lexeur nous a été utile dans cette première étape pour garantir la validité des unités lexicales observées. À l'aune de sa couverture assez large, nous considérons dans la suite de cette étude que l'absence d'un nom agentif en *-eur* dans la ressource Lexeur est un indice du caractère néologique de cet agentif. Mais cette ressource a de fait l'inconvénient de limiter le vocabulaire

Wikipédia	herzigova - coiffeuse - venhard - naymark - manucure - vericel - sorokina - trulle - cover-girl - gitane - séménoff - chammah - comédienne - estragnat - yma - stroyberg - réju - tallier - soubrette - alycia - montalant - minouche - dartonne - ménine - metmer - rembauville - jitka - catzéflis - prepon - denarnaud - marie-olivier - tainsy - cuisinière - chauffeuse - anicée - serveuse - stripteaseuse - kajmak - laury - ballerine - barmaid - lunchlady - pierens -laparé - servantie - mammamia - stresi - irma - elfride - vendell
LM10	duègne - rousse - jolie - gitane - pulpeux - vamp - ravissant - bacchante - chatte - diablesse - boulotte - mignonne - allumeuse - madone - rockeuse - danseuse - parisienne - nymphomane - débutante - brune - mégère - lhamo - almée - ingénue - soubrette - véro - blonde - mamelue - pimbêche - adorable - femme-objet - femme-oiseau - garce - pétulant - servante-maîtresse - servante - dévergond - antillaise - trémière - courtisane - arnaqueuse - donzelle - nastassia - diva - guenon - chasseresse - junon - demi-mondaine - rieuse - belle-de-nuit
frWaC	kiraly-picot ouria - roitfeld - joustra - pezeril - extrado - montiss - punkette - gommettes - poupee - lainer - capsuleuses - spigarelli - pomagalski - bouvrain - prucnal - diffused - klinge - biboud - naomie - vitteaut - existais - corbery - raszewski - perleuse - taraud - baetz - planard - trouvain - wathier - gouttiere - loutch - robart - yomoshi - danner - cheriie - melodick - devraigne - plaignaud - stiletto - dheran - nallamoutou - poupée - morganne - turbulette - impertinante - marieras - chipette - carpenito - burada

TABLE 4 – 50 premiers voisins des dérivés prototypiques agentifs féminins en *-euse* et *-rice* dans les corpus *Wikipédia*, *LM10* et *frWaC*

	Wikipédia	***LM10***	***frWaC***
Masculin	814	451	163
Féminin	72	142	18

TABLE 5 – Fréquence moyenne des voisins des dérivés prototypiques en fonction du suffixe de constitution et du corpus

observable, en focalisant l'analyse sur les formes installées dans le lexique de la langue, et dites lexicalisées, puisque issues pour la majeure partie du *TLFi*. Leur sens a donc potentiellement évolué depuis leur formation, du fait de l'histoire de chaque mot. Nous pouvons notamment citer comme exemple le nom *échangeur*, dérivé du verbe *échanger*, qui a maintenant davantage le sens de système autoroutier que de personne ou d'instrument permettant d'échanger quelque chose. Or, le suffixe *-eur* est productif, et il produit encore des noms d'agents et d'instruments (Dubois, 1962; Dubois & Dubois-Charlier, 1999; Aronoff & Lindsay, 2014).

Nous faisons l'hypothèse que les formes les plus récentes obtenues par la suffixation en *-eur* sont sémantiquement plus transparentes vis-à-vis de l'instruction sémantique de leur verbe de base, à l'image de *bloggeur* et *blogger*. Nous souhaitons donc créer une représentation de ces agentifs néologiques, afin d'évaluer l'impact de la lexicalisation sur la classe sémantique des noms agentifs. Nous reproduisons pour cela la même expérience mais à partir de noms agentifs déverbaux suffixés en *-eur* qui ne sont pas présents dans Lexeur.

4.1 Repérage des noms agentifs en *-eur* néologiques

4.1.1 Extraction automatique de paires potentielles

Nous extrayons dans un premier temps des paires (N*eur*, V). Nous cherchons des paires, et pas simplement des noms en *-eur*, pour garantir que le nom en *-eur* est bien un nom déverbal. La procédure est la suivante : dans un premier temps, nous récupérons l'ensemble des 4 675 paires (N*eur*, V) présentes dans Lexeur. Pour chaque paire, un programme apprend la règle de transformation formelle liant le dérivé à sa base (Tanguy & Hathout, 2007). Dans un second temps, le programme parcourt l'ensemble des mots du vocabulaire du modèle distributionnel considéré. Le programme traite chaque mot finissant par la chaîne *-eur* comme un dérivé potentiel. Il lui attribue alors une base potentielle, à partir des règles qu'il a apprises précédemment. Un même dérivé peut éventuellement se voir attribuer plusieurs bases potentielles, à l'image de *superordinateur* qui est associé à *superordonner*, *superordiner* et *superordinater*. Nous nous retrouvons à ce stade avec 6 152 paires potentielles pour le corpus *Wikipédia*, contre 3 677 pour le corpus *LM10* et 10 665 pour le corpus *frWaC*. Dans un dernier temps, le programme élimine les paires dont la base potentielle est absente du modèle, soit parce qu'elle n'est pas assez fréquente, soit parce qu'elle n'existe pas. Nous obtenons 218 paires pour le corpus *Wikipédia*, contre 87 pour le corpus *LM10* et 726 pour le corpus *frWaC*.

4.1.2 Vérification manuelle des paires

Une étape de vérification manuelle s'ensuit pour ne conserver que les paires sémantiquement et formellement valides. Nous considérons comme valide une paire dont le premier élément est bien un nom agentif (agent ou instrument) en *-eur* dérivé d'un verbe, et dont le second élément est bien le verbe duquel est dérivé le nom agentif considéré, à l'image de *bloggeur* et *blogger*. Nous faisons donc le choix d'exclure les paires erronées selon les critères suivants :

— La base ne correspond pas à une forme verbale dans le corpus. Cela exclut des paires comme (*seigneur*, *seigner*), où *Seigner* est un nom propre.
— Le dérivé n'a pas d'emploi agentif dans le corpus. Cela exclut des paires comme (*sueur*, *suer*).
— Il s'agit d'une variante orthographique d'une paire présente dans Lexeur. Cela exclut les paires (*co-producteur*, *co-produire*) ou (*realisateur*, *realiser*) puisque les paires (*coproducteur*, *coproduire*) et (*réalisateur*, *réaliser*) sont déjà présentes dans Lexeur. Nous n'excluons cependant pas les paires comme (*coanimateur*, *coanimer*), malgré la présence dans Lexeur de la paire (*animateur*, *animer*), puisque la préfixation pourrait ici se traduire par une variation sémantique en corpus.
— Le verbe et son dérivé ne sont pas liés sémantiquement. Cela exclut des paires comme (*primeur*, *primer*).

À l'issue de cette vérification manuelle, nous obtenons 81 paires pour le corpus *Wikipédia*, 27 pour le corpus *LM10* et 169 pour le corpus *frWaC*.

4.2 Comparaison du comportement distributionnel

Nous avons d'abord comparé sur le plan distributionnel le comportement des paires (N*eur*, V) issues de Lexeur à celles issues du corpus. Puisque le nom agentif pour les paires néologiques est considéré

comme régulier, nous faisons l'hypothèse que le verbe et son déverbal en *-eur* seront sémantiquement plus proches que ne le seraient un verbe et son déverbal lexicalisé. Cela devrait donc se traduire, sur le plan distributionnel, par un score de proximité plus important pour les paires issues du corpus. Nous faisons par ailleurs l'hypothèse que cela s'appliquera à une majorité des paires considérées. Cela se traduirait par une dispersion moindre des paires et donc un écart type plus faible.

Nous faisons la moyenne des scores de proximité fournis par Word2Vec pour les paires issues de Lexeur et pour celles issues du corpus et nous comparons ces scores de proximité moyens. Nous calculons aussi l'écart type pour évaluer la dispersion de nos paires.

	frWaC		***Wikipédia***		***LM10***	
	Lexicalisation	Néologie	Lexicalisation	Néologie	Lexicalisation	Néologie
Proximité	0.293	**0.307**	0.271	**0.324**	0.262	**0.346**
Écart type	**0.171**	0.185	**0.165**	0.197	**0.163**	0.181

TABLE 6 – Score de proximité moyen et écart type entre le verbe et son nom agentif en *-eur* en fonction du corpus et du type de noms agentifs

Les résultats regroupés dans le tableau 6 montrent que le score de proximité entre le verbe et son déverbal en *-eur* est en moyenne plus élevé pour les paires à caractère néologique que pour les paires lexicalisées. Cela va donc dans le sens de notre hypothèse initiale d'une plus grande proximité liée à une plus grande transparence du dérivé. Le calcul du t-test montre que la différence est significative pour les corpus *Wikipédia* et *LM10* (p-value de 0.01 et 0.02), mais qu'elle ne l'est pas pour le corpus *frWaC* (p-value = 0.3). Par ailleurs, nous constatons que l'écart type est plus élevé pour les paires issues du corpus que pour les paires issues de Lexeur. Cela signifie que les verbes sont en moyenne plus proches de leur dérivé néologique, mais avec une dispersion plus importante.

Nous complétons notre observation par l'analyse de la distribution des paires (N*eur*, V) en fonction de leur score de proximité, selon leur degré de lexicalisation dans les trois corpus considérés (figure 1).

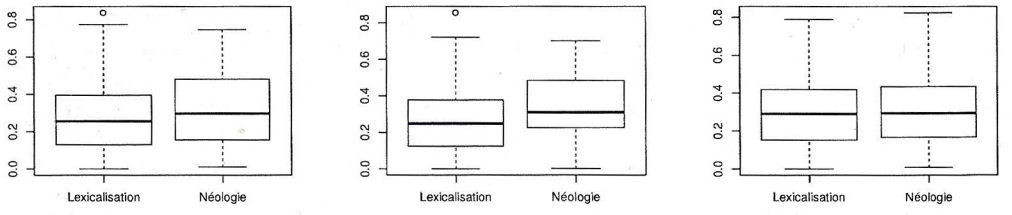

FIGURE 1 – Distribution des paires en fonction de leur score de proximité et du degré de lexicalisation du nom agentif. À gauche : corpus *Wikipédia*. Au centre : corpus *LM10*. À droite : corpus *frWaC*

On constate que la distribution des paires est similaire dans les corpus *Wikipédia* et *LM10*. On observe en effet un score de proximité médian et un score de proximité pour les interquartiles plus élevés pour les paires néologiques que pour les paires lexicalisées, et des extrêmes légèrement moins marqués. Ces observations vont une nouvelle fois dans le sens de notre hypothèse. Le corpus *frWaC* se caractérise quant à lui par une distribution quasi identique des paires quel que soit leur degré de lexicalisation. La seule variation observée concerne les valeurs extrêmes supérieures, qui semblent légèrement plus

élevées pour les paires néologiques. Ces résultats confirment le statut un peu à part du corpus *frWaC* dont nous analysons plus précisément les résultats en 4.3.

Une fois établies ces tendances générales, nous relevons quelques comportements particuliers dans l'ensemble des paires néologiques. Certaines présentent des scores de proximité très élevés (>0.7). C'est le cas de paires comme (*parseur*, *parser*) ou (*sampleur*, *sampler*) pour le corpus *frWaC*, (*parseur*, *parser*) et (*vocodeur*, *vocoder*) pour le corpus *Wikipédia*, ou (*rappeur*, *rapper*) pour le corpus *LM10*. Une analyse en corpus de ces différents lexèmes montre que les termes *parser*, *sampler*, *vocoder* ou *rapper* sont très majoritairement utilisés en tant que nom commun et non en tant que verbe, en lieu et place de leur équivalent en *-eur*. L'utilisation de ces anglicismes expliquent donc la forte proximité des formes de chaque paire.

On constate par ailleurs la présence de paires néologiques au score de proximité très faible (<0.05), là encore pour les trois corpus. On retrouve parmi elles des paires comme (*raveur*, *raver*) ou (*raideur*, *raider*) pour le corpus *frWaC*, (*desserviteur*, *desservir*) pour le corpus *Wikipédia* ou (*auditeur*, *auditer*) pour le corpus *LM10*. Une analyse des contextes d'apparition montre ainsi que *Raver* est principalement utilisé comme un nom propre. Enfin, les lexèmes les plus instanciés en corpus par les formes *raideur*, *desservir* ou *auditeur* ne correspondent pas aux familles sémantiques visées par les paires (N*eur*, V) auxquelles les formes appartiennent.

Dans les deux cas, que la proximité soit faible ou forte, cela met en exergue des paires problématiques. Un score de proximité en décalage par rapport à la tendance globale semble donc être un bon indice pour détecter les paires atypiques. On peut s'interroger sur la pertinence d'avoir conservé les paires comme (*parseur*, *parser*) évoquées précédemment. Ce choix est dû aux critères explicités dans la section 4.1.2 que nous avons appliqués de façon stricte, mais dont on voit ici les limites. Le fait d'avoir au moins une occurrence agentive ou verbale en corpus ne semble ainsi pas assez restrictif. De la même façon, le caractère néologique de certaines paires comme (*auditeur*, *auditer*) est à remettre en question. Ainsi, *auditeur* n'est pas un néologisme, et on le retrouve dans Lexeur dans deux familles, une liée au nom *audition*, et l'autre au nom *audit*. Si la paire (*auditeur*, *auditer*) n'est effectivement pas présente dans Lexeur, ce n'est pas dû à son caractère néologique, mais à un problème d'exhaustivité de la ressource.

En revanche, parmi les paires au comportement typique, on observe de nombreux cas de figure intéressants. On retrouve ainsi des paires comme (*uploadeur*, *uploader*), (*tchatteur*, *tchatter*), (*débogueur*, *déboguer*), (*bloggeur*, *blogger*) et (*multiplexeur*, *multiplexer*). Ces paires sont précisément celles qui nous intéressent puisqu'il s'agit bien ici de cas de dérivation récente d'agentifs à partir de verbe, où le sens du dérivé n'a pas encore évolué.

4.3 Observation des voisins des noms agentifs néologiques

Nous faisons le choix de nous concentrer plus précisément sur le corpus *frWaC* car c'est celui pour lequel nous avons le plus grand nombre de paires et donc de noms agentifs néologiques. Nous souhaitons par ailleurs nous pencher davantage sur les résultats *a priori* contre-intuitifs présentés en 4.2.

De la même façon que dans la section 3, nous créons le vecteur représentant la moyenne des vecteurs des noms agentifs en *-eur* récupérés, et nous en analysons les 50 premiers voisins (tableau 7). Nous les comparons avec les 50 premiers voisins du dérivé prototypique lexicalisé dans le corpus *frWaC* reportés dans la dernière partie du tableau 3.

quenc - mytheather - toneport - expandeur - electret - comptact - microcontroleur - le-gnou - handsonic' - webeditor - genlock - go-to - attiny - oehlbach - o-system - shamallows - easybox - mickeyfreestyler - aldila - atmega - whirlwind - r-core - audiounit - coprocesseur - mini-navigateur - frw - enhancer - selector - seeprog - serato - realtek - twido - wago-i - testeur - hwmonitor - multiprogrammateur - speedo - winup - pod - crystalin - textorm - modul - beeprog - yokogawa - modutils - dragonfly - sniffeur - electromatic - hammerhead - encodeur

TABLE 7 – 50 premiers voisins du dérivé néologique prototypique en *-eur* dans le corpus *frWaC*

Rappelons que dans le cas des 50 premiers voisins obtenus à partir des noms agentifs présents dans Lexeur, on dénombrait seulement 10% d'entités nommées (*Petzl*, *Cartillier*) contre 88% de lexèmes qui n'étaient pas des entités nommées (*surgrip*, *ferrailleur*, *encliquetables*). Les 2% restants correspondaient à la forme *Rilsan*, qui peut être utilisé comme une entité nommée ou un nom commun. Ces voisins ne relèvaient pas du champ sémantique des nouvelles technologies, à quelques exceptions près (*Gehennas*), mais davantage du monde industriel ou technique (*servo-moteur*, *tuyautage*).

Le comportement diffère grandement pour les voisins du dérivé néologique prototypique dans *frWaC*. La première chose que l'on constate est que les 50 premiers voisins du dérivé prototypique en *-eur* formés à partir des noms agentifs absents de Lexeur sont peu fréquents (*oehlbach*, *attiny*), avec une fréquence moyenne de 54 contre 163 pour les voisins obtenus à l'aide des déverbaux issus de Lexeur (tableau 5). L'interprétation de ces voisins est d'autant plus compliquée qu'il s'agit majoritairement d'entités nommées (à raison de 68%), pour lesquelles il est difficile de capter une homogénéité sémantique. On retrouve notamment des marques (*Yokogawa*), des logiciels (*quEnc*) ou encore des pseudonymes (*MickeyFreeStyler*). On ne compte *a contrario* que 26% de lexèmes qui ne sont pas des entités nommées (*webeditor*, *crystalin*). Enfin, les 6% restants sont des formes qui sont utilisées en corpus tantôt en tant qu'entité nommée, tantôt en tant que nom commun. Notons qu'une grande partie de ces noms dénotent des objets, informationnels ou matériels, ou des instruments. Tous les voisins sont néanmoins liés par le champ sémantique auquel ils se rapportent, celui de l'informatique, de l'électronique et des nouvelles technologies. Ces résultats soulignent l'hétérogénéité du corpus *frWaC* mis en lumière par la figure 1. Il est en effet quantitativement plus important, mais son contenu, plus bruité, est plus difficile à exploiter.

5 Discussion

Lors de cette expérimentation, nous avons pu comparer sur le plan distributionnel des représentants de la classe sémantique des noms agentifs déverbaux en *-eur*, en évaluant l'impact du corpus et de leur niveau de lexicalisation. Les résultats préliminaires sont encourageants puisque nous parvenons à capter la notion d'agentivité au travers d'un représentant prototypique. Nous avons vu que cette représentation permettait d'accéder à certaines composantes du sens de la classe de mots observée, puisqu'elle intégrait l'information liée au genre dans le cas de l'analyse des noms agentifs féminins en *-euse* et *-rice*, et qu'elle permettait même d'accéder à des connaissances du monde liées à la connotation de ces noms agentifs féminins (Dawes, 2003; Schafroth, 2001). Nous avons par ailleurs constaté que les déverbaux suffixés en *-eur* ne constituent pas une classe sémantique homogène dans les modèles Word2Vec utilisés dans cette étude. On observe ainsi la présence de plusieurs catégories au sein de la classe des noms agentifs en *-eur*, à savoir les noms d'instruments et les noms d'agents, se

divisant elles-mêmes en sous-classes ayant des comportements distincts. Enfin, nous avons travaillé sur des noms agentifs néologiques, et nous avons montré qu'ils étaient distributionnellement plus proches de leur base que ne le sont les noms agentifs lexicalisés dans les corpus *Wikipédia* et *LM10*. Nous avons cependant constaté que ce n'était pas le cas dans le corpus *frWaC*, soulignant le caractère hétérogène de ce corpus mis en avant dans le reste de cette étude. Cela nous invite à nous interroger sur l'utilisation de corpus de taille importante, mais peu contrôlés et donc potentiellement très bruités, pour faire de l'analyse linguistique fine.

Nous envisageons à ce titre de poursuivre notre travail sur l'anglais. Les corpus y sont plus conséquents, et nous pourrions exploiter les informations sémantiques, morphologiques et formelles de certaines ressources à la couverture importante (Fellbaum & Miller, 2003). Outre la question du contrôle du corpus, une piste méthodologique ébauchée par cette étude concerne la démarche de sélection de noms agentifs en *-eur* néologiques. Nous avons vu les limites de notre définition de la néologie, mais nous avons fait le choix de conserver toutes les paires que nous avions constituées du fait de leur nombre déjà relativement limité. Cela nous invite néanmoins à reconsidérer nos critères de filtrage. Un premier moyen serait de nous servir de la distribution pour évincer les paires atypiques, ou du moins appeler à leur vérification manuelle. Un second moyen, complémentaire du premier, consisterait en une annotation morphosyntaxique des formes en corpus. Cela permettrait de ne conserver que les paires dont les deux membres ont pour acception principale le lexème voulu (verbe ou nom commun), et d'évincer de potentiels anglicismes. La projection d'un lexique anglais permettrait également de filtrer les formes erronées.

Si nous avons pu vérifier la stabilité des observations à travers différents corpus, nous n'avons pas pris en compte d'autres paramètres liés au modèle. Nous avons fait le choix d'utiliser dans cette étude le modèle par défaut fourni par Word2Vec, en l'état, sans chercher dans un premier temps à intervenir sur ses hyperparamètres, comme le nombre de dimensions ou l'architecture. Nous souhaitons à ce stade de la thèse exploiter l'outil tel qu'il est dans le but d'orienter notre analyse linguistique. Malgré une analyse corpus par corpus et non globalisante des représentations prototypiques, nous devrions dans l'idéal moyenner les distances de plusieurs représentations distributionnelles (Antoniak & Mimno, 2018), mais nous n'avons pas pu réaliser ces expérimentations par manque de temps. Nous avons cependant reproduit la manipulation en fixant le nombre de dimensions des vecteurs à 300 sans que les résultats varient significativement. Dans cette optique, nous envisageons de comparer les différentes techniques distributionelles, et avons entamé la reprise de l'expérimentation avec l'outil fastText (Bojanowski *et al.*, 2016). Nous avons par ailleurs fait le choix dans cette étude de baser nos observations sur les 50 premiers voisins des vecteurs construits, mais nous envisageons à terme d'analyser de façon plus systématique les voisins des vecteurs construits en observant la répartition et la densité des noms d'agent dans ce voisinage.

Références

ANTONIAK M. & MIMNO D. (2018). Evaluating the Stability of Embedding-based Word Similarities. *Transactions of the Association for Computational Linguistics*, **6**, 107–119.

ARONOFF M. & LINDSAY M. (2014). Productivity, Blocking and Lexicalization. *The Oxford handbook of derivational morphology*, p. 67–83.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2016). Enriching Word Vectors with Subword Information. *arXiv preprint arXiv :1607.04606*.

DAWES E. (2003). La féminisation des titres et fonctions dans la francophonie : de la morphologie à l'idéologie. *Ethnologies*, **25**(2), 195–213.

DUBOIS J. (1962). *Étude sur la dérivation suffixale en français moderne et contemporain : essais d'interprétation des mouvements observés dans le domaine de la morphologie des mots construits*. Paris : Larousse.

DUBOIS J. & DUBOIS-CHARLIER F. (1999). *La dérivation suffixale en français*. Paris : Nathan.

FABRE C. & LENCI A. (2015). Distributional Semantics Today Introduction to the Special Issue. *Traitement Automatique des Langues*, **56**(2), 7–20.

FELLBAUM C. & MILLER G. A. (2003). Morphosemantic Links in Wordnet. *Traitement Automatique des Langues*, **44**(2), 69–80.

HATHOUT N. & FABRE C. (2002). Constitution et exploitation de lexiques de formes déverbales. *Communication aux Journées d'études sur les noms déverbaux. Silex, Université Lille*, **3**.

HUYGHE R. & TRIBOUT D. (2015). Noms d'agents et noms d'instruments : le cas des déverbaux en-eur. *Langue française*, (1), 99–112.

KINTSCH W. (2001). Predication. *Cognitive science*, **25**(2), 173–202.

KLEIBER G. (1990). *La sémantique du prototype : catégories et sens lexical*. PUF.

KULKARNI V., AL-RFOU R., PEROZZI B. & SKIENA S. (2015). Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th International Conference on World Wide Web (WWW)*, p. 625–635, Florence, Italy.

LAPESA G., KAWALETZ L., PLAG I., ANDREOU M., KISSELEW M. & PADO S. (à paraître). Disambiguation of Newly Derived Nominalizations in Context : A Distributional Semantics approach. (draft).

MIKOLOV T., CHAN K., CORRADO G. & DEAN J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of International Conference on Learning Representations (ICLR)*, Scottsdale, United States of America.

SCHAFROTH E. (2001). *Gender in French Structural Properties, Incongruences*, In *Gender Across Languages : The Linguistic Representation of Women and Men*, volume 3, p. 87–117. John Benjamins Publishing.

TANGUY L. & HATHOUT N. (2007). *Perl pour les linguistes*. TIC et Sciences Cognitives. Hermès sciences publications. Site d'accompagnement : http ://perl.linguistes.free.fr.

URIELI A. (2013). *Robust French Syntax Analysis : Reconciling Statistical Methods and Linguistic Knowledge in the Talismane Toolkit*. PhD thesis, Université de Toulouse II le Mirail.

VARVARA R., LAPESA G. & PADÓ S. (2016). Quantifying Regularity in Morphological Processes : An Ongoing Study on Nominalization in German. In *ESSLLI DSALT Workshop : Distributional Semantics and Semantic Theory*, Bolzano, Italy.

VERHOEVEN B., DAELEMANS W. & VAN HUYSSTEEN G. (2012). Classification of Noun-noun Compound Semantics in Dutch and Afrikaans. In *Proceedings of the 23rd Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, p. 121–125, Pretoria, South Africa.

WAUQUIER M., FABRE C. & HATHOUT N. (2018). Différenciation sémantique de dérivés morphologiques à l'aide de critères distributionnels. In *Congrès Mondial de Linguistique Française (CMLF)*, Mons, Belgique.

ZELLER B. D., PADÓ S. & SNAJDER J. (2014). Towards Semantic Validation of a Derivational Lexicon. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, p. 1728–1739, Dublin, Ireland.