Différenciation sémantique de dérivés morphologiques à l'aide de critères distributionnels

Marine Wauquier[1], Cécile Fabre[1], et *Nabil* Hathout[1]

[1]CLLE, CNRS & Université de Toulouse, 5 Allées Antonio Machado, 31058 Toulouse Cedex 9, France
marine.wauquier@univ-tlse2, cecile.fabre@univ-tlse2.fr, nabil.hatout@univ-tlse2.fr

Résumé. Dans ce travail, nous examinons sur le plan distributionnel le sens de dérivés morphologiques, et plus précisément des noms d'agent déverbaux en *-eur*, *-euse* et *-rice*, et des noms d'action déverbaux en *-age*, *-ion* et *-ment*. Nous utilisons une approche distributionnelle automatisée et un lexique dérivationnel. Nous proposons une représentation de l'information distributionnelle permettant d'examiner le sens prototypique des dérivés et l'instruction sémantique prototypique des suffixes. Nous montrons notamment que la différence entre les suffixes *-eur*, *-euse* et *-rice* ne relève pas seulement du genre et que les dérivés en *-age*, *-ion* et *-ment* présentent des profils spécifiques sur le plan distributionnel.

Abstract. Contributions of distributional semantics for the semantic study of morphologically derived words. In this paper, we examine on a distributional level the meaning of morphologically derived words. We take a closer look at deverbal agent nouns formed with the French suffixes *-eur*, *-euse* and *-rice*, and nominalisations formed with the French suffixes *-age*, *-ion* and *-ment*. We combine a distributional approach and the use of a linguistic resource. We provide a representation of distributional information that allows us to examine the prototypical meaning of derivatives and the prototypical semantic instruction of suffixes. In particular we show that the distinction between the suffixes *-eur*, *-euse* and *-rice* is not limited to the gender. Moreover, we show that the suffixes *-age*, *-ion* and *-ment* show distributional specificities.

1 Introduction

L'hypothèse distributionnelle, proposée par Harris (1954), Firth (1957) ou Miller et Charles (1991), stipule que la proximité sémantique entre mots peut être assimilée à leur degré de proximité distributionnelle. Ces principes ont été traduits en modèles computationnels dans lesquels les mots sont représentés sous la forme de vecteurs de contextes (Sahlgren, 2008, Baroni et Lenci, 2010). La proximité de deux vecteurs est alors une indice de la proximité

sémantique des mots représentés. La nature mathématique de cette représentation du sens permet d'envisager des opérations sur les vecteurs résultants, pour simuler des opérations sémantiques (compositionnalité, désambiguïsation, analogie, etc.) (Baroni et al. 2014).

Ces méthodes automatiques fondées sur une approche distributionnelle du sens connaissent aujourd'hui un succès important en traitement automatique des langues (Fabre et Lenci, 2015). La linguistique commence à se les approprier, afin de tirer parti de la possibilité de mettre en œuvre à très large échelle, sur de vastes corpus, l'hypothèse harrissienne, sur des questions aussi diverses que l'évolution du sens des mots (Kulkarni et al., 2015), le figement (Baroni et Zamparelli, 2010, Verhoeven et al., 2012) ou la mise au jour de classes sémantiques (Schulte Im Walde, 2006).

À la lumière des résultats produits par ces modèles d'analyse distributionnelle, notre objectif est de tirer parti de la possibilité d'étudier à grande échelle sur de gros corpus variés les propriétés distributionnelles des mots construits. Nous voulons ainsi réexaminer certaines questions qui intéressent la morphologie, comme la différenciation sémantique des suffixes au sein des familles dérivationnelles. Nous envisageons la sémantique distributionnelle comme un outil permettant de déployer une approche extensive de la morphologie, par la prise en compte d'un très grand nombre de contextes pour définir les profils distributionnels de lexèmes ou de familles de lexèmes. Nous utilisons dans cette étude une ressource linguistique, Lexeur, constituée manuellement et regroupant des noms d'agent en *-eur* et une partie de leurs familles dérivationnelles. Nous examinons en particulier les noms d'agent déverbaux en *-eur*, *-euse* et *-rice*, et les nominalisations processives en *-age*, *-ion* et *-ment* pour mettre au jour les contrastes entre les suffixes *-euse* et *-rice* et entre les suffixes *-age*, *-ion* et *-ment*.

Nous dressons tout d'abord un état des lieux des critères habituellement utilisés pour aborder la question de la différenciation sémantique des suffixes. Nous présentons ensuite notre dispositif expérimental et les premiers résultats de l'étude, fondés sur une représentation distributionnelle dont le niveau de généralité et d'abstraction va croissant, partant des familles de lexèmes (section 4) et abordant le profil distributionnel des suffixes eux-mêmes (section 5).

2 Différenciation sémantique des dérivés morphologiques

Nous présentons dans un premier temps quelques généralités sur les suffixes qui font l'objet de notre étude et ce que l'on sait de leurs rapports sémantiques avec leur base verbale.

2.1 La nominalisation processive en -age, -ion et -ment

La nominalisation est un procédé dérivationnel permettant de créer des noms d'action à partir de verbes. Ce procédé implique des opérations catégorielle (V → N) et formelle (ajout d'un affixe), mais théoriquement pas d'opération sémantique (Roché, 2009)1. Le verbe et le nom d'action seraient donc sémantiquement maximalement proches, puisque ce dernier dénoterait simplement sous une autre forme syntaxique la situation dynamique décrite par le verbe.

La proximité sémantique entre le verbe et ses dérivés a fait l'objet de nombreux travaux, qui ont principalement porté sur deux types de critères : la préservation de la structure argumentale du verbe (Grimshaw, 1990) et l'héritage de propriétés sémantiques, en particulier aspectuelles, du verbe par le nom (Haas et al., 2008). Ces travaux se fondent

1 L'idée est ancienne et remonte au moins à Chomsky (1970).

généralement sur l'application de tests d'acceptabilité, éventuellement complétés par des procédures d'annotation de corpus (Balvet et al., 2011). De nombreux procédés sont disponibles pour créer des noms d'action, comme les suffixes *-ure*, *-ité*, *-ance* et *-ence* ou encore la conversion. Mais les suffixations en *-age*, *-ion* et *-ment* sont de loin les plus productives (Fradin, 2014). Toutes les trois peuvent former des déverbaux et divers critères ont été proposés pour expliquer le choix d'un suffixe plutôt qu'un autre. La transitivité du verbe fait notamment partie des critères syntaxiques évoqués (Dubois, 1962 ; Fradin, 2014). Fradin (2014) a aussi souligné la nécessité d'une base savante pour le suffixe *-ion* et d'une base populaire pour les suffixes *-age* et *-ment*. Sur le plan sémantique, les auteurs se sont penchés tantôt sur la nature sémantique des arguments du verbe, qu'il s'agisse du sujet (Martin, 2010) ou de l'objet (Fradin, 2014), tantôt sur la télicité du verbe (Martin, 2010). L'action dénotée par le nom déverbal a aussi été considérée du point de vue de la longueur de sa chaîne évenementielle ou de son incrémentialité (Martin, 2010), mais aussi de son domaine ontologique (Dubois, 1962 ; Martin, 2010).

Ces critères sont assez variés mais ne proposent pas une vision d'ensemble, à grande échelle. Ils poussent par ailleurs à s'interroger sur l'équivalence sémantique de ces trois suffixations.

2.2 La suffixation en *-eur*, *-euse* et *-rice*L'étude du sens des noms d'agent déverbaux inclut notamment l'examen des cas de concurrence suffixale, pour expliquer qu'une forme prévale sur une autre dans le cas de paires suffixales comme *-ee* (*attendee* 'participant') et *-er* (*attender* 'participant') en anglais (Heyvaert, 2011) ou *-iste* (*chimiste*) et *-ien* (*physicien*) en français (Lignon, 2007). Dans notre étude, du fait des données disponibles dans la base Lexeur (cf. section 3.1), nous faisons le choix de nous concentrer sur les noms d'agent féminins en *-euse* et *-rice*. Peu d'études ont à notre connaissance été menées sur la comparaison sémantique des formes masculines et féminines, d'une part, et des formes féminines entre elles d'autre part, sinon dans une approche psycholinguistique ou sociolinguistique.

Nous passons par le suffixe *-eur* pour comparer les suffixations en *-euse* et *-rice*. Ces trois suffixes forment des noms d'agent (*acheteur*) ou d'instrument (*distributeur*) à partir de verbes (*acheter*, *distribuer*), et plus rarement de noms (*camion* → *camionneur*). Un nom d'agent désigne une entité animée qui réalise l'action décrite par le verbe de façon intentionnelle. Un nom d'instrument désigne quant à lui l'artefact prototypiquement utilisé pour réaliser l'action que le verbe décrit (Huyghe et Tribout, 2015).

La distinction sémantique entre le suffixe *-eur* et les suffixes *-euse* et *-rice* a connu une évolution diachronique. Le suffixe *-eur* servait historiquement à désigner l'agent et les suffixes *-euse* et *-rice* l'outil ou l'instrument à partir de la même base, à l'image de *moissonneur* et *moissonneuse* (Dubois, 1962). Cette différence se serait cependant effacée à mesure de l'utilisation croissante de machines et de l'automatisation du travail (Dubois, 1962), mais aucune étude diachronique n'a, à notre connaissance, confirmé cette hypothèse.

Les suffixes masculin et féminins diffèrent concernant le genre référentiel (le genre de la personne dénotée) du nom d'agent qu'ils forment. À l'image des suffixes *-trice* en italien (*lavatrice* 'lave-linge'), *-in* en allemand (*Autorin* 'auteur femme') ou *-ess* en anglais (*huntress* 'chasseuse'), les suffixes *-euse* et *-rice* indiquent le genre féminin de la personne dénotée. Des travaux soulignent la présence d'une valeur sémantique supplémentaire du féminin liée aux attentes et aux valeurs culturelles, à l'image de *mister* 'monsieur' et *mistress* 'maîtresse' en anglais (Marcato et Thüne, 2002 ; Hellinger, 2001). Ces travaux sont cependant encore peu nombreux, se concentrant généralement plus sur des aspects formels que sémantiques (Schafroth, 2001).

Lorsqu'ils co-existent, les suffixes féminins ne sont pas non plus strictement équivalents. Les suffixes *-euse* et *-rice* sont porteurs, à différents degrés, de connotations

sociolinguistiques (Dawes 2003), lorsqu'ils ne sont pas utilisés pour désigner la femme de l'agent (Le Draoulec et Péry-Woodley, 2016), à l'image de *ambassadrice*. Le suffixe *-rice* est notamment jugé plus noble et plus valorisant que le suffixe *-euse*, jugé dépréciatif (Houdebine-Gravaud, 1998 ; Dawes, 2003 ; Lenoble-Pinson, 2008). Cette tendance se retrouve dans d'autres langues, romanes comme germaniques : le suffixe français *-esse* et ses équivalents italien *-essa*, roumain *-esa* et allemand *-ess* sont, eux aussi, fortement connotés (Dawes, 2003 ; Marcato et Thüne, 2002 ; Meurice, 2001 ; Bußmann et Hellinger, 2003). Ces connotations sont d'ordre sexuel ou dépréciatif. Des formes non connotées existent alors en parallèle, comme les suffixes italien *-trice* et allemand *-in*.

2.3 Contributions de l'étude Les travaux que nous venons d'évoquer se fondent principalement sur l'application de tests d'acceptabilité, selon une approche empirique à partir d'un nombre nécessairement limité de cas. Nous nous proposons d'appliquer un outil d'analyse distributionnelle automatique pour éclairer la différenciation sémantique des dérivés en *-eur*, *-euse*, *-rice* d'une part, et en *-age*, *-ion* et *-ment* d'autre part. Les travaux exploitant l'analyse distributionnelle automatique pour comparer les suffixes et les dérivés qu'ils forment sont encore peu nombreux. Nous citerons par exemple Zeller et al (2014) qui montrent que la différence de genre référentiel se traduit par une distance distributionnelle variable entre les noms d'agent masculin et féminin. Varvara et al (2016) ont pour leur part différencié sur le plan distributionnel deux procédés de nominalisation processive concurrents de l'allemand. Lapesa et al (2017) utilisent quant à eux des indices distributionnels pour entraîner des classifieurs automatiques à identifier les lectures événementielles des noms d'action anglais en *-ment*.

Dans la suite de ce travail, nous examinons l'hypothèse selon laquelle, sur le plan sémantique, les dérivés en *-euse* et les dérivés en *-rice* sont uniquement les équivalents féminins des dérivés en *-eur* correspondants. Une deuxième hypothèse est que les suffixes *-age*, *-ion* et *-ment* ne comportent pas non plus de différences sur le plan distributionnel.

Soulignons que nos questionnements et nos hypothèses sont directement issus de travaux de linguistique descriptive. Nous nous donnons essentiellement ici le moyen de vérifier à grande échelle la validité de ces hypothèses. Nos principales contributions sont : 1) l'utilisation de représentations sémantiques opérationnelles pouvant être comparées facilement ; 2) le traitement global d'ensembles de relations dérivationnelles qui évite d'avoir à travailler sur des petits échantillons d'exemples dont la représentativité n'est pas assurée.

3 Dispositif expérimental

Nous cherchons dans cette étude à utiliser les sources d'information en fonction de ce pourquoi elles ont été créées : nous souhaitons ainsi combiner ainsi l'efficacité des outils d'analyse distributionnelle automatique en termes d'analyse sémantique à des connaissances expertes validées par des linguistes. Word2Vec fournit à ce titre les représentations sémantiques et Lexeur les descriptions morphologiques.

3.1 Lexeur

Nous basons notre étude sur une ressource morphologique dérivationnelle, Lexeur, comportant 5974 noms d'agent en *-eur*. Cette ressource consacrée au recensement des noms en *-eur* et de leur famille dérivationnelle a été constituée au sein de l'équipe CLLE-ERSS (Hathout et Fabre, 2002). Les noms sont issus du *Trésor de la Langue Française*, complétés par des attestations issues du Web. Chaque nom en *-eur* a été associé, par une procédure d'annotation manuelle, à une partie de sa famille constructionnelle, composée de la base (verbale ou nominale), et d'une liste de tous les noms processifs identifiés. La ressource a

par la suite été complétée par l'ajout, pour chaque nom en *-eur*, de son ou ses équivalents féminins en *-euse* ou *-rice*, dans le cadre du projet Démonette (Hathout et Namer, 2014). Chaque lexème de la base est muni d'une étiquette morphosyntaxique. Cinq entrées de Lexeur sont illustrées dans le tableau 1.

Tableau 1. Extrait de Lexeur

Nom d'agent masc.	Nom d'agent fém.	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
fraudeur/Ncms	fraudeuse/Ncfs	frauder/Vmn--	Vb	fraude/Ncfs
sculpteur/Ncms	sculpteuse/Ncfs ; sculptrice/Ncfs	sculpter/Vmn--	Vb	sculpture/Ncfs ; sculptage/Ncms
wheeleur/Ncms	wheeleuse/Ncfs	wheel/Ncms	Nb	Ø

Ces exemples montrent la diversité des familles constructionnelles : certaines sont très fournies, comme dans le cas de *abatteur*, d'autres peuvent être lacunaires comme celle de *endoscopeur* (sans base verbale identifiée) ou de *wheeleur* (qui a seulement un dérivé agentif). 78 % des noms d'agent recensés sont construits à partir d'un verbe, 14 % à partir d'un nom, et 7 % n'ont pas de base associée (à l'image de *endoscopeur*). Tous les noms d'agent en *-eur* ont des équivalents féminins, mais les suffixes *-euse* et *-rice* n'apparaissent pas dans les mêmes proportions. On dénombre ainsi 3 fois plus d'agents féminins en *-euse* qu'en *-rice* (4542 contre 1514). À peine plus d'1 % des noms d'agent en *-eur* présentent les deux variantes, à l'image de *sculpteur* dans le tableau 1. Concernant les noms d'action, seules 78 % des familles constructionnelles contiennent au moins un nom d'action, tous suffixes confondus. Pour ces entrées-là, on dénombre en moyenne 1,47 nom d'action, le nombre de noms d'action par entrée variant entre 1 et 8.

Nous parlons par abus de langage de noms d'agent, mais Lexeur regroupe en réalité indistinctement des noms d'agent (*chanteur*) et des noms d'instrument (*transmetteur*). De la même façon, Lexeur regroupe sans distinction des noms d'action à l'interprétation événementielle (*abattage*), résultative (*sculpture*), ou encore stative (*abattement*). Enfin, les lexèmes intégrés à la ressource présentent divers degrés de polysémie, là aussi non renseignés (*construction* peut être une action, une activité ou un résultat).

3.2 Corpus

L'utilisation d'un système de calcul distributionnel automatique requiert l'analyse de corpus de grande taille. Nous avons opté pour l'utilisation de deux corpus de genre textuel distinct pour tester la stabilité des observations. Le premier est le corpus *Wikipédia,* issu de la version française de 2013 de l'encyclopédie en ligne. Il compte environ 255 millions de mots. Ce choix est guidé par le souhait de disposer d'un vocabulaire vaste et varié, relevant de domaines hétérogènes, à l'image de la diversité des lexèmes que nous étudions. Nous le comparons au corpus *LM10*, composé des articles du journal *Le Monde* publiés entre les années 1991 et 2000 et qui contient environ 200 millions de mots.

3.3 Word2Vec

La méthode distributionnelle a été automatisée dès les années 1990 (Grefenstette, 1994 ; Habert et Zweigenbaum, 2002). Dans ces modèles dits classiques, chaque dimension du vecteur représentant un mot enregistre son degré d'association avec l'ensemble des contextes considérés dans le corpus d'analyse. Une réduction de dimensions est généralement réalisée pour rendre le vecteur plus dense. Récemment, des outils basés sur des réseaux de neurones, comme Word2Vec (Mikolov et al., 2013) ou fastText (Bojanowski et al., 2016), ont été développés et se sont popularisés du fait de leurs performances, de leur efficacité en termes de coût de traitement et de leur facilité d'utilisation. Ces outils exploitent des modèles dits prédictifs qui, sur la base d'un apprentissage non supervisé, sont entraînés à prédire les mots susceptibles d'apparaître dans un contexte donné.

Nous utilisons Word2Vec pour construire les représentations distributionnelles des mots. Word2Vec fournit une représentation vectorielle du sens des mots d'un corpus et exploite cette représentation à l'aide de différents modules permettant de déterminer les voisins distributionnels des mots, de calculer le score de proximité distributionnelle entre plusieurs mots ou de proposer des solutions à des équations analogiques. Le score de proximité entre deux mots, calculé à partir du cosinus des vecteurs, varie de 0 (proximité nulle) à 1 (proximité maximale pour deux formes dont les représentations distributionnelles sont identiques). Ces outils sont relativement simples d'utilisation, mais leur efficacité a pour prix une opacité des traitements intermédiaires. Contrairement aux méthodes classiques, où chaque dimension d'un vecteur est identifiable, la condensation de l'information distributionnelle en quelques centaines de dimensions rend ces dernières non directement interprétables.

Le calcul distributionnel est basé dans cette étude sur l'examen de cooccurrences lexicales dans une fenêtre contextuelle donnée, sans prise en compte des relations syntaxiques. Nous construisons une matrice par corpus. Les mêmes paramètres par défaut sont utilisés pour les deux matrices. Word2Vec utilise par défaut 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. Les corpus ont été au préalable lemmatisés.

4 Sens lexical prototypique

Notre objectif est de construire une représentation de l'information sémantique prototypiquement associée à un suffixe donné. Puisque cette abstraction n'est pas instanciée dans le corpus, nous ne pouvons pas en calculer la représentation vectorielle comme pour n'importe quel mot. Nous ne disposons pour cela que des vecteurs des lexèmes construits par ce suffixe.

4.1 Représentation prototypique des dérivés

Nous utilisons ici une notion de dérivé prototypique2 dont nous définissons le sens comme étant la moyenne des sens des mots formés à partir de ce suffixe3. Le vecteur **SUFF** du dérivé prototypique d'un suffixe *suff* est ainsi calculé comme la moyenne des vecteurs **Nsuff**$_{\mathbf{i}}$ des mots porteurs du suffixe tel qu'indiqué en (1).

2 Nous employons la notion de prototype en regard de l'idée d'une catégorisation graduelle (Kleiber, 1990). Nous cherchons ici à décrire le dérivé qui instancierait le plus de traits caractéristiques d'une catégorie sémantique dérivationnelle donnée.

3 Nous nous inspirons pour cela du travail de Kintsch (2001) sur les prédicats.

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

Pour constituer la représentation prototypique du dérivé d'un suffixe donné, nous additionnons l'ensemble des vecteurs de la série dérivationnelle correspondante (dans notre exemple, tous les vecteurs de noms d'agent en *-eur*) et nous divisons ce vecteur global par le nombre de vecteurs qui ont été additionnés. Nous ne prenons en compte que les vecteurs des mots porteurs d'un suffixe donné et présents dans Lexeur (tableau 2), pour éviter de considérer des mots porteurs de la chaîne de caractères correspondante mais non porteurs de l'instruction sémantique visée (comme *fleur* pour *-eur*).

Tableau 2. Nombre de mots pris en compte pour le calcul des vecteurs prototypiques

	-eur	***-euse***	***-rice***	***-age***	***-ion***	***-ment***
Wikipédia	1 334	239	90	707	1 635	592
LM10	1 147	155	65	563	1 507	561

Une fois ce vecteur abstrait construit, nous étudions l'information sémantique qu'il véhicule. Pour cela, nous choisissons d'observer les 50 voisins distributionnels les plus proches de ce vecteur.

Nous pouvons ainsi vérifier les hypothèses que nous avons formulées : est-ce que la différence principale, sur le plan distributionnel, entre le dérivé prototypique en *-eur* d'une part et les dérivés prototypiques en *-euse* et *-rice* d'autre part, relève du genre sexuel du référent ? De même, les dérivés prototypiques des suffixes *-age*, *-ion* et *-ment* sont-ils similaires sur le plan distributionnel dans la mesure où tous les trois suffixes ont en théorie la même instruction sémantique ?

4.2 Dérivés en *-eur*, *-euse* et *-rice*

Pour accéder à l'instruction sémantique de nos vecteurs construits, nous observons leurs 50 premiers voisins distributionnels.

4.2.1 Une féminisation à deux vitesses

Le tableau 3 présente les 50 premiers voisins du vecteur du dérivé prototypique en *-eur* pour le corpus *Wikipédia*. Du fait du procédé de création de ce vecteur abstrait, on pouvait s'attendre à trouver dans le voisinage de celui une majorité de noms d'agent en *-eur*, à savoir les mots ayant servi à sa création. Or on constate que ce n'est pas le cas, puisque 56 % des voisins du dérivé moyen ne sont pas suffixés en *-eur*. On retrouve ainsi par exemple des dérivés suffixés en *-mètre*, *-ier*, ou encore *-ien*. Un des voisins en *-eur* (*débogueur*) est absent de Lexeur et n'a donc pas été pris en compte dans la création de cette représentation prototypique. Si tous les voisins du dérivé prototypique en *-eur* donnés dans le tableau 3 sont des noms d'agent (*soudeur*) ou d'instrument (*minuteur*), tous ne sont pas déverbaux (*client*, *stéthoscope*). Cela confirme que le vecteur construit à partir des noms d'agent en *-eur* véhicule le sens à la fois agentif et instrumental associé à la suffixation en *-eur* mais qui ne lui est pas exclusif.

Tableau 3. 50 premiers voisins du vecteur moyen des dérivés en *-eur* dans le corpus *Wikipédia*

-eur	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

Le tableau 4 présente les 50 premiers voisins des vecteurs des dérivés prototypiques en *-rice* et *-euse*. Là encore, on ne compte respectivement que 16 % et 10 % de voisins porteurs des suffixes *-rice* et *-euse*, dont les mots *co-fondatrice* et *stripteaseuse*, absents de Lexeur. On retrouve pour le reste notamment les suffixes *-ette*, *-ière* ou encore *-ienne*. On ne trouve pas de noms d'instruments parmi les voisins du dérivé moyen en *-rice*, et près de 40 % de noms de métier. *A contrario*, les 50 premiers voisins du dérivé moyen en *-euse* comptent seulement 14 % de noms de métier, et 8 % de noms d'instrument (*chauffeuse*, *chocolatière, cuisinière*). Dans les deux cas, les noms de métier sont de genre référentiel féminin, à l'exception du nom épicène *trapéziste*. Les noms d'instrument sont quant à eux de genre grammatical féminin, du fait des suffixes qui les construisent. Les 60 % de voisins restants du dérivé moyen en *-rice* sont des prénoms ou patronymes associés à des personnalités féminines (réelles ou fictives), à l'exception de A*nska* et S*lávka* qui sont respectivement des noms de rivière et d'astéroïde. Les autres voisins du dérivé prototypique en *-euse* sont plus variés. On trouve notamment des mots relatifs à la cuisine (*trulle*, *basquaise*) ou à la mode (*salopette*, *poulaine*), des domaines culturellement associés à la femme. On trouve surtout des mots désignant ou qualifiant la femme (*minouche*, *mariée*), dont un certain nombre très fortement connotés péjorativement (*diablesse*, *souillon*, *cochonne*). Le dérivé prototypique en *-euse* est donc davantage connoté que le dérivé prototypique en *-rice*, dont le sens agentif semble assez bien défini. Les dérivés en *-euse* et *rice* ne sont donc pas strictement équivalents sur le plan distributionnel, et nous retrouvons dans les voisinages distributionnels les différences identifiées dans les études descriptives de ces suffixations (cf. section 2.2).

Tableau 4. 50 premiers voisins des vecteurs moyens des dérivés en *-euse* et *-rice* dans le corpus *Wikipédia*

-rice	professeure - co-fondatrice - cofondatrice - herzigova - directrice - pharmacienne - traductrice - chercheure - saint-lucienne - venhard - éducatrice - co-directrice - fondatrice - laury - conférencière - comédienne - blogueuse - gogean – desmarais-rondeau - assistante - mageina - musicienne - vyghen - ingénieure - gérante - mammamia - spaziani - anska - bouhenni - slávka - joano - séménoff - herzigová - shrier - dartonne - warmus - présentatrice - bourgeois-leclerc - tonietti - otternaud – directrice-adjointe - guirous - saller - sculptrice - tshiteya - naymark - écrivaine - rajskub - pomfresh - fadeïeva
-euse	gitane - trulle - chauffeuse - manucure - soubrette - trapéziste - coiffeuse - chocolatière - yma - cuisinière - minouche - salopette - allumeuse - barancey - herzigova - souillon - diablesse - cochonne - vericel - serveuse - sorokina - stroyberg - naymark - rivale - corré - venhard - sarbel - kajmak - râblure - fédora - montalant - poulaine - stripteaseuse - catzéflis - mini-jupe - rosine - mariée - ptereleotris - tallier - irma - suffel - cover-girl - épicière - marie-olivier - javotte - kerny - basquaise - emilienne - estragnat - tigresse

4.2.2 Variations entre corpus

Nous nous sommes interrogés sur l'impact du corpus dans la représentation de ces dérivés prototypiques, et notamment sur le dérivé prototypique en *-euse*. En effet, Wagner et al (2015) ont montré que le lexique utilisé dans les pages Wikipédia dédiées à des personnalités féminines différait de celui utilisé pour les personnalités masculines. Les femmes y sont ainsi davantage caractérisées relativement à leur vie amoureuse que les hommes. Les mêmes analyses réalisées sur le corpus *LM10* (cf. annexe 1) confortent pourtant les observations réalisées sur le corpus *Wikipédia*.

Les 50 premiers voisins du dérivé prototypique en *-eur* dans le corpus *LM10* sont là encore majoritairement agentifs (*ramoneur*, *charretier*). Les noms d'instruments disparaissent au profit de noms d'humains (*garçon*, *comparse, rouquin*, *fripon*) et de noms d'animaux (*chiot*, *canari*). L'agentivité du dérivé prototypique est donc ici à prendre dans un sens plus large d'entité animée masculine. Parmi les 50 premiers voisins du dérivé prototypique de *-rice*, on retrouve de nouveau des noms de métiers (*fondatrice*, *papesse*) et des noms de personnalités féminines (*Solange*). On constate l'apparition de quelques noms communs (*réalisations*, *différentialistes*) et gentilés (*québecquoise*). Le sens du dérivé prototypique de *-rice* semble globalement relever ici aussi des notions d'agentivité et de féminin. Quant au dérivé prototypique de *-euse*, on retrouve parmi ses voisins dans le corpus *LM10* des noms de métier (*duègne*, *lavandière*) et les termes connotés *diablesse* ou *allumeuse* présents pour le corpus *Wikipédia.* Les noms d'instruments disparaissent, et laissent la place exclusivement à des adjectifs ou des noms qualifiant et désignant les femmes, pour la plupart fortement connotés comme *nymphomane* ou *blondeur*.

Ces observations indiquent qu'il y a également une distinction sur le plan distributionnel entre les noms d'agent masculins et féminins d'une part, et entre les dérivés en *-euse* et *-rice* d'autre part dans le corpus *LM10*. Cela suggère que ces différences ne sont pas strictement liées au genre textuel du corpus, mais potentiellement propres aux suffixations étudiées.

4.3 Dérivés en *-age*, *-ion* et *-ment*Les 50 voisins des dérivés prototypiques des suffixes *-age*, *-ion* et *-ment* dans le corpus *Wikipédia* présentés dans le tableau 6 se caractérisent par une forte unité morphologique. En effet, entre 74 % et 88 % des voisins des dérivés prototypiques en *-age*, *-ion* et *-ment* sont porteurs du suffixe représenté. Rappelons que dans le cas des dérivés en *-eur*, *-euse*, et *-rice*, le rapprochement sur la base du suffixe était bien moins important (entre 10 % et 46 %) et les voisins semblaient donc principalement unis sur le plan sémantique. La présence fortement majoritaire du suffixe dans les voisins de chaque dérivé en *-age*, *-ion* et *-ment* suggère que la contribution sémantique du suffixe est suffisamment marquée pour séparer les noms d'action. En d'autres termes, les suffixations en *-age*, *-ment* et *-ion* ne semblent pas neutres sur le plan distributionnel, et l'on se serait attendu à trouver des dérivés construits ayant les trois exposants parmi les voisins des trois dérivés prototypiques. Cela se maintient, dans une moindre mesure, dans le corpus *LM10*, le nombre de voisins porteurs du suffixe du dérivé prototypique duquel ils sont rapprochés oscillant entre 56 % (pour *-age*), 82 % (pour *-ment*) et 84 % (pour *-ion*).

Nous constatons par ailleurs que les voisins des trois dérivés sont exclusivement des noms d'action. Une quantification précise des voisins en fonction de leur interprétation processive reste à faire. On constate par ailleurs que les voisins du dérivé prototypique en *-age* désignent principalement des opérations concrètes ou techniques (*usinage*, *polissage*). *A contrario*, les voisins du dérivé prototypique en *-ion* désignent des notions plus abstraites et sous-spécifiées, soumises à une polysémie relativement importante (*activation*, *assimilation*). Les voisins du dérivé prototypiques en *-ment* sont parfois techniques (*colmatage*), parfois plus abstraits (*ajustement*). Signalons que la mise en évidence d'une différenciation selon la dimension spécificité/généralité est nouvelle et constitue l'une des contributions de ce travail.

Tableau 6. 50 premiers voisins des vecteurs moyens des dérivés en *-age*, *-ion* et *-ment* dans le corpus *Wikipédia*

-age	démoulage - usinage - séchage - remplissage - perçage - soufflage - démontage - coulage - broyage - sablage - chargement - étirage - soudage - dégraissage - traçage - sciage - trempage - polissage – vissage - nettoyage - gonflage - lavage - pulvérisation - roulement - remontage - roulage - assemblage – réglage - meulage - recuit - dégagement - compostage - soudure - affûtage - salage - désinfection – foulage - cuivrage - enrobage - vidange - refroidissement - étanchéité - clouage - décapage - rechargement - stockage - grattage - rinçage - dégivrage - brasage
-ion	activation - réévaluation - réduction - simplification - dégradation - détérioration - assimilation - acceptation - utilisation - modification - transformation - manipulation - application – évaluation - dilution - détermination - surcharge - stimulation - dilatation - généralisation - différenciation - altération - action - mutation - dispersion - complexification - dénaturation - homogénéisation - vérification - réaction - survenue - coupure - compréhension - fixation - intervention - limitation - appropriation - régénération - formulation - imputation - inhibition - prolifération - perception - définition - analyse - constatation - dissociation - actualisation - accumulation - mesure
-ment	enfoncement - durcissement - déplacement - blocage - dépassement - échauffement - élargissement - relâchement - abaissement - éparpillement - isolement - envahissement - affaissement – dégagement - effritement - ajustement - rejet - écoulement - dysfonctionnement - ralentissement - basculement - allongement - affaiblissement - accroissement - traitement - étirement - rétrécissement – équilibrage - endommagement - épuisement - emballement - encombrement - accumulation - absence - remplissage - relèvement - inconfort - tassement - ensablement - éloignement - renforcement - utilisation - étalement - engorgement - usure - redémarrage - lessivage - décollement - gonflement - colmatage

Ces résultats semblent montrer l'existence d'une sélection sémantique de la base spécifique à chaque suffixe qui reste encore à préciser. **5 Instruction sémantique prototypique**

Nous nous interrogeons sur la possibilité d'étendre la méthode à un niveau d'abstraction supplémentaire, et de représenter l'instruction sémantique des suffixes elle-même. Nous mettons à l'épreuve une méthode pour modéliser cette instruction sémantique en nous affranchissant de l'instruction sémantique de la base verbale présente dans les dérivés déverbaux.

5.1 Démarche

Pour modéliser cette instruction, nous nous inspirons des travaux de Bolukbasi et al (2016) et de Bonami (2017). Bonami fait le choix de représenter la dérivation agentive en *-eur* et la flexion à la 3e personne du singulier de l'imparfait sous la forme d'une soustraction afin de comparer la stabilité sémantique de ces deux procédés. Cette opération est à mettre en parallèle avec la théorie soutenue par Laca (2001) et Koonz-Garboden (2007) stipulant que la dérivation conserve ou ajoute du sens, mais n'amène jamais à une perte de sens. Nous pouvons dès lors envisager de représenter l'instruction sémantique d'un suffixe sous la forme d'un vecteur que l'on obtient lorsque l'on soustrait le vecteur de la base verbale au vecteur du dérivé (2).

Pour obtenir le vecteur **suff** représentant l'instruction sémantique prototypique d'un suffixe donné, nous calculons dans un premier temps tous les vecteurs $\mathbf{suff}_i$ pour toutes les instances du suffixe, en soustrayant à tous les vecteurs de noms dérivés le vecteur de leur

verbe associé (2), puis nous en faisons la moyenne selon la formule (3).

$$\boldsymbol{suff}_i = \boldsymbol{Nsuff}_i - \boldsymbol{V}_i \tag{2}$$

$$\boldsymbol{suff} = \frac{\sum_{i=1}^{n} suff_i}{n} \tag{3}$$

De la même façon que pour le dérivé prototypique, nous accédons au sens représenté par **suff** par le biais de ses voisins distributionnels. Nous calculons le score de proximité cosinus de notre vecteur avec l'ensemble des mots du modèle, et nous observons les 50 voisins pour lesquels ce score est le plus proche de 1. Les résultats bruts de cette opération de soustraction génèrent des voisins très peu fréquents et difficilement interprétables (cf. annexe 2), à l'image de *unaid* et *artiflex*. Les voisins ont donc été filtrés pour ne conserver que ceux dont la fréquence est supérieure ou égale à 100.

Signalons que seuls 1 235, 235 et 85 noms d'agent, respectivement en *-eur*, *-euse* et *-rice*, ont été pris en compte pour la constitution des vecteurs de suffixe. Cela est dû à la méthode de construction, puisque les formules (2) et (3) imposent que le modèle contienne simultanément la représentation du nom d'agent et celle du verbe correspondant.

5.2 Premières observations pour *-eur*, *-euse* et *-rice*

Nous faisons l'hypothèse que l'analyse des voisins des vecteurs de suffixes montrera la présence d'une instruction sémantique traduisant la notion d'entité humaine ou d'outil. Or, on ne trouve parmi les 50 premiers voisins du vecteur de *-eur* (cf. annexe 3) qu'un seul nom d'agent, *coproducteur* (deux, si l'on compte la forme anglaise *scientists*). On trouve en revanche beaucoup de prénoms, patronymes ou noms de compagnies (*Zacharias*, *Adamson*, *Koehler*). Un certain nombre de ces patronymes sont aussi utilisés comme toponymes (*Buckland*, *Needham*). Les voisins du suffixe *-euse* sont relativement similaires à ceux du suffixe *-eur*, et l'on notera qu'ils en partagent 12 (à l'image de *eda*, *Adamson* ou encore *rauch*). On constate une plus grande proportion de toponymes parmi les voisins du suffixe *-euse*, et que les patronymes ne sont pas exclusivement féminins. Les voisins du suffixe *-rice* sont quant à eux plus homogènes, puisque l'on obtient majoritairement des prénoms féminins (*Felicia*, *Kristina*, *Theresa*).

Malgré le filtrage, les voisins des instructions sémantiques prototypiques sont moins éloquents que ceux des dérivés prototypiques, pourtant obtenus sans filtrage. Certains voisins sont ainsi très difficilement interprétables (*socio, r-u*). On notera cependant que le contenu sémantique lié à l'action décrite par le verbe est ici absente, et que les trois suffixes sont majoritairement rapprochés d'entités animées, et tout particulièrement le suffixe *-rice*. D'une certaine manière, la dimension agentive semble ne pas pouvoir être séparée de la dimension processive incarnée par les bases verbales.

6 Conclusion

Dans ce travail, nous avons mis au jour, sur le plan distributionnel, des régularités concernant les suffixations agentives en *-eur*, *-euse* et *-rice* d'une part, et processives en *-age*, *-ion* et *-ment* d'autre part. Nous avons ainsi proposé une caractérisation distributionnelle des dérivés protoypiques de ces suffixes. Nous avons dans un premier temps étudié les suffixations agentives en opposant le masculin et le féminin puis les

suffixes féminins *-euse* et *-rice*. En observant les voisins distributionnels des dérivés prototypiques, nous avons montré qu'ils n'étaient pas strictement équivalents, et plus spécifiquement que le suffixe *-euse* était porteur d'une connotation sexuée péjorative que l'on ne retrouve pas dans l'information distributionnelle associée au suffixe *-rice*. Les résultats sont par ailleurs similaires lorsque l'on passe d'un corpus encyclopédique à une corpus journalistique. Selon la même méthodologie, nous avons comparé dans un second temps les suffixations processive. Nous avons pu montrer que les dérivés prototypiques en *-age*, *-ion* et *-ment* se différenciaient fortement sur le plan distributionnel, chacun attirant des voisins formellement homogènes. Cette différenciation était aussi d'ordre sémantique puisque les dérivés en *-age* et *-ion* s'opposaient par le niveau de technicité et de spécification de leurs voisins.

Nous avons essayé dans un dernier temps de représenter sous la forme d'un vecteur l'instruction sémantique des suffixes *-eur*, *-euse* et *-rice*, mais l'analyse des voisins distributionnel de ces vecteurs a montré les limites de notre méthode. Notre dispositif expérimental ne permet pas en l'état de passer à un niveau d'abstraction supérieur.

Nous souhaitons poursuivre ce travail par une évaluation plus objective des voisins des dérivés prototypiques. Nous souhaitons par ailleurs compléter la caractérisation sémantique des suffixations en *-age*, *-ion* et *-ment* en confronter les résultats que nous obtenons grâce à notre approche distributionnelle automatique aux critères évoqués dans les sections 2.1 et 2.2.

Références

Balvet, A., Barque, L., Condette, M. H., Haas, P., Huyghe, R., Marin, R., & Merlo, A. (2011). La ressource Nomage. Confronter les attentes théoriques aux observations du comportement linguistique des nominalisations en corpus. *Traitement Automatique des Langues*, 52(3), 129-152.

Baroni, M. & Lenci, A. (2010). Distributional memory: A General Framework for Corpus-based Semantics. *Computational Linguistics, 36*, 673-721.

Baroni, M.; Bernardi, R. & Zamparelli, R. (2014). Frege in Space: A Program of Compositional Distributional Semantics. *LiLT (Linguistic Issues in Language Technology), 9,* 241-346.

Baroni, M. & Zamparelli, R. (2010). Nouns are Vectors, Adjectives are Matrices: Representing Adjective-noun Constructions in Semantic Space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP),* Cambridge,1183-1193.

Bojanowski, P.; Grave, E.; Joulin, A. & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606.*

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V. & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS),* Barcelone*,* 4349-4357.

Bonami, O. (2017). Predictability in Inflection and Word Formation. *ParadigMo 2017: First Workshop on Paradigmatic Word Formation Modeling.*

Bußmann, H., & Hellinger, M. (2003). Engendering Female Visibility in German. *Gender across languages: The linguistic representation of women and men*, *3*, 141-174.

Chomsky, N. (1970). Remarks on Nominalization. In*, R. Jakobs and P. Rosenbaum (Eds.), Readings in English Transformational Grammar*, 184-221.

Dawes, E. (2003). La féminisation des titres et fonctions dans la Francophonie : de la morphologie à l'idéologie. *Ethnologies, Association Canadienne d'Ethnologie et de Folklore,* 25, 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.

Fabre, C. & Lenci, A. (2015). Distributional Semantics Today – Introduction to the special issue. *Traitement Automatique des Langues*, 56(2), 7-20.

Firth, J. R. (1957). A Synopsis of Linguistic Theory 1930-1955. *Studies in linguistic analysis, Basil Blackwell*.

Fradin, B. (2014). La variante et le double. *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux,* 109-147.

Grefenstette, G. (1994). Corpus-Derived First, Second and Third-Order Word Affinities. In *Proceedings of the 6th International Congress on Lexicography (EURALEX*), Amsterdam, 279-290.

Grimshaw, J. (1990). *Argument Structure*. Cambridge, MA : MIT Press.

Haas, P., Huyghe, R., & Marin, R. (2008). Du verbe au nom : calques et décalages aspectuels. In *Congrès Mondial de Linguistique Française (CMLF)*, Paris, 2051-2065.

Habert, B. & Zweigenbaum, P. (2002). Contextual Acquisition of Information Categories: what has been done and what can be done automatically?. In *The Legacy of Zellig Harris: Language and information into the 21st century*, Bruce Nevin (resp.), Amsterdam, John Benjamins vol. 2. Mathematics and computability of language, Collection CILT n°229, 203–231.

Harris, Z. S. (1954). Distributional Structure. *Word,* 10, 146-162.

Hathout, N. et C. Fabre (2002). Constitution et exploitation de lexiques de formes déverbales. Communication aux Journées d'études sur les noms déverbaux.

Hathout, N. & Namer, F. (2014). Démonette, a French Derivational Morpho-semantic Network. *Linguistic Issues in Language Technology,* 11(5), 125-168.

Hellinger, M. (2001). English–Gender in a Global Language. *Gender Across Languages : The linguistic representation of men and women*, *1*, 105-113.

Heyvaert, L. (2011). Attenders or Attendees? Deverbal -ee and-er Variants in English. *Journal of Pragmatics*, *43*(1), 62-72.

Houdebine-Gravaud, A. M. (1998). L'imaginaire linguistique: questions au modèle et applications actuelles. *Limbaje şi comunicare*, 9-32.

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

Kintsch, W. (2001). Predication. *Cognitive science, 25*, 173-202.

Kleiber, G. (1990). *La sémantique du prototype: catégories et sens lexical*. Paris : Presses Universitaires de France.

Koontz-Garboden, A. (2007). *States, changes of state, and the Monotonicity Hypothesis*. (Phd thesis, Stanford University**).**

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),* Florence, 625-635.

Laca, B. (2001). Derivation. *Language Typology and Language Universals: An International Handbook,* (vol *2*, p1214-1227). Berlin, Boston: De Gruyter.

Lapesa, G.; Kawaletz, L.; Plag, I.; Andreou, M.; Kisselew, M. & Pado, S. (2017). Disambiguation of Newly Derived Nominalizations in Context: A Distributional Semantics Approach. Manuscrit soumis pour publication.

Le Draoulec, A., & Péry-Woodley, M.P. (2016). La femme de l'écrivain [Billet de blog]. Repéré à http://bling.hypotheses.org/1405

Lenoble-Pinson, M. (2008). Mettre au féminin les noms de métier : résistances culturelles et sociolinguistiques. *Le français aujourd'hui,* 73-79.

Lignon, S. (2007). Les noms de spécialistes en *-iste* et en *-ien* : le chimiste perturbé ou comment le physicien se réajuste. *Perturbations et Réajustements. Langue et langage*, 287-295.

Marcato, G., & Thüne, E. M. (2002). Gender and Female Visibility in Italian. *Gender Across Languages: The Linguistic Representation of Women and Men*, *2*, 187-217.

Martin, F. (2010). The Semantics of Eventive Suffixes in French. *The Semantics of Nominalizations across Languages and Frameworks.* Berlin, Mouton de Gruyter, 109-141

Meurice, F. (2001). Deconstructing Gender – The Case of Romanian. *Gender across Languages: The linguistic representation of men and women*, *1*, 229-252.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of International Conference on Learning Representations (ICLR),* Scottsdale.

Miller, G. A. & Charles, W. G. (1991). Contextual Correlates of Semantic Similarity. *Language and cognitive processes, Taylor & Francis*, 6, 1-28.

Roché, M. (2009). Pour une morphologie lexicale. *Mémoires de la Société de Linguistique de Paris*, n° 17, 65-87.

Sahlgren, M. (2008). The Distributional Hypothesis. *Italian Journal of Linguistics*, 20(1), 33-54.
Schafroth, E (2003). Gender in French - Structural Properties, Incongruences and Asymmetries. *Gender Across Languages: The Linguistic Representation of Women and Men*, 3, 87-117.
Schulte Im Walde, S. (2006). Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics, 32*, 159-194.
Varvara, R.; Lapesa, G. & Padó, S. (2016). Quantifying regularity in morphological processes: An ongoing study on nominalization in German. *ESSLLI DSALT Workshop: Distributional Semantics and Semantic Theory.*
Verhoeven, B.; Daelemans, W. & van Huyssteen, G. (2012). Classification of noun-noun compound semantics in Dutch and Afrikaans. In *Proceedings of the Twenty-Third Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)* Pretoria, 121-125.
Wagner, C.; Garcia, D.; Jadidi, M. & Strohmaier, M. (2015). It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *Proceedings of the 9th International AAAI Conference on Web and Social Media* (*ICWSM),* Oxford*,* 454-463.
Zeller, B. D., S. Padó, & J. Snajder (2014). Towards semantic validation of a derivational lexicon. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, Dublin, 1728–1739.

Annexes

Annexe 1. 50 premiers voisins des vecteurs moyens des dérivés en *-eur*, *-euse* et *-rice* dans le corpus *LM10*

-eur	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 - coiffeur - julot - boxeur - arnaqueur - malfrat - voyou - écuyer - prestidigitateur - moussaillon - cuisinier - sarret - puncheur - fêtard - camelot - affabulateur - braconnier - canari - garçon
-rice	fondatrice - directrice - boschi - présidente-fondatrice - québécoise - réalisations - agenin - professeure - lecallier - cofondatrice - gurrey - différentialistes - sullé - codet - lepoivre - hellekant - barbin - ambrosi - solange - bécue - experte - mahr - cibot - papesse - anny - setchouan - défricheuse - fouco - elzbieta - suppléante - codirecteur - bashkirtseff - barale - fossati - salvaire - gourfink - diaconesse - lucienne - mémoriale - lisette - pumain - woringer - wouters - busq - swenson - ballis - aubriot - beaurain - tinguy - lausseur
-euse	duègne - bacchante - gitane - ravissant - chatte - diablesse - pulpeux - vamp - jolie - boulotte - madone - mignonne - trogne - soubrette - donzelle - allumeuse - rousse - silhouetter - blonde - garce - adorable - mégère - brodeuse - blond - trémière - lavandière - nymphomane - pipe - rieuse - frimousse - nymphe - matrone - courtisane - voilette - ballerine - servante - femme-oiseau - midinette - naïade - espiègle - blondeur - almée - guenon - pomponner - femme-enfant - teigne - minauder - strip-teaseuse - pépée - citrouille

Annexe 2. 50 premiers voisins des vecteurs moyens du suffixe -eur dans le corpus Wikipédia avant filtrage par la fréquence

-eur	forsans - sasía - řezníček - mctabb - artiflex - donnaud - zinberg - monig - arabo - unaid - easynet - yachtman - xbt - beugniot - vasouy - oblata - guillevin - mejirov - daguerréotypiste - delisa - ruland - sébastien-joseph - gphs-rtg - chourer - wiliam - continuiste - karkamánis - gałecki - souilhe - bibeault - critiquement - nouvellement - lalliance - gromard - fivethirtyeight - rentilly - chevassus - sorabji - enayat-seraj - marisy - sicob - effets-spéciaux - izrael - gugliotta - hfr - plassat - kinnick - gvb - virae - étévé

Annexe 3. 50 premiers voisins des vecteurs moyens des suffixes *-eur*, *-euse* et *-rice* dans le corpus *Wikipédia* après filtrage par la fréquence

-eur	nouvellement - rauch - edson - fraichement - ridder - fraîchement - waddell - buckland - adamson - courvoisier - smits - lema - boe - eda - zacharias - rust - bateman - martyn - heyer - lanier - pipes - pugh - koehler - shuster - scientists - fisch - needham - hyman - straits - salter - ory - kow - keir - coproducteur - alzon - herd - tillie - perri - mast - matches - kling - cypher - giamatti - roché - theron - biggs - loudon - smet - edmundo - berenson
-rice	kem - felicia - allyson - kristina - skeeter - darlene - theresa - nouvellement - rauch - myrtle - kayla - mitzi - edson - fraichement - doreen - lexi - jayne - tutin - arnie - adamson - matches - stratton - lia - celeste - pickett - britt - zoë - izzy - beasley - tanja - allie - courtney - alana - robby - jennie - billings - vinny - angelika - masterson - margery - inez - koehler - mink - crabtree - eunice - r-u - trina - jordana - stacy - jana
-euse	nouvellement - eda - adamson - artistic - hoff - scientists - socio - ads - boe - rust - rauch - hyman - lema - evergreen - statistics - keir - heyer - strategies - manila - warsaw - tutin - systematics - christiansen - zweifel - krystyna - nominee - salerno - karolina - urb - blanda - hellenic - advisor - sdp - margery - ridder - jeffreys - asch - tanja - ory - projects - dsm - mast - araújo - trials - martyn - kovařík - millot - mcdonough - mikhailov - mcculloch