Confrontation des procédés dérivationnels et des catégories sémantiques dans les modèles distributionnels

Marine Wauquier

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 2 - Jean Jaurès

Présentée et soutenue par

Marine WAUQUIER

Le 4 décembre 2020

Confrontation des procédés dérivationnels et des catégories sémantiques dans les modèles distributionnels

---

Ecole doctorale : **CLESCO - Comportement, Langage, Education, Socialisation, Cognition**

Spécialité : **Sciences du langage**

Unité de recherche :
CLLE - Unité Cognition, Langues, Langage, Ergonomie

Thèse dirigée par
Nabil HATHOUT

Jury

M. Olivier Bonami, Rapporteur
M. Ingo Plag, Rapporteur
Mme Cécile Fabre, Examinatrice
Mme Fiammetta Namer, Examinatrice
M. Laurent Prévot, Examinateur
M. Nabil Hathout, Directeur de thèse

En mémoire de ma mère

Remerciements

Alors que l'expérience qu'est le doctorat touche à sa fin, mes pensées se tournent vers toutes les personnes qui ont rendu cette aventure possible, ou qui l'ont rendue plus enrichissante encore.

Au premier rang des ces personnes se trouvent Nabil Hathout et Cécile Fabre, que je ne remercierai jamais assez de m'avoir donné cette chance de faire de la recherche, et de m'avoir soutenue (ou devrais-je dire supportée ?) depuis maintenant 5 ans. La confiance qu'ils m'ont accordée, la richesse des conseils qu'ils m'ont prodigués, et la bienveillance dont ils ont fait preuve, entre autres, m'ont été précieuses, et m'ont permis de m'épanouir tant professionnellement qu'humainement. Je n'aurais pu rêver d'un meilleur encadrement, et je leur suis à ce titre infinimment reconnaissante.

Je tiens aussi à remercier Olivier Bonami, Ingo Plag, Fiammetta Namer, et Laurent Prévost d'avoir accepté de faire partie de mon jury de thèse, et d'accorder à mon travail leur temps et leur expertise.

Cette expérience n'aurait pas été la même sans la participation de Richard Huyghe, qui a eu la folle mais riche idée de proposer à la jeune doctorante que j'étais une collaboration. Sa foi en mon travail et la richesse de nos échanges ont grandement contribué à façonner cette expérience. Mille mercis, donc.

Je remercie aussi pour l'émulation scientifique et humaine les membres de (feu) l'axe Cartel, et plus particulièrement Ludovic Tanguy et Mai Ho-Dac pour leur accompagnement depuis le master, ainsi que Franck Sajous et Basilio Calderone pour le concours technique et statistique. Plus largement, je remercie le laboratoire CLLE, qui m'a fourni un environnement de travail des plus propices, et l'ensemble des collègues chercheurs et enseignants-chercheurs avec qui j'ai eu l'occasion d'échanger et de travailler, et parmi eux Michel Roché, Josette Rebeyrolle, Fabio Montermini, Juliette Thuilier, et j'en passe. J'aurais aussi aimé remercier l'ensemble des jeunes chercheurs, chercheurs et enseignants-chercheurs dont j'ai croisé la route depuis mes premiers pas en Sciences du langage, mais une thèse ne serait sans doute pas suffisante pour tous les citer.

Aussi riche cette expérience soit-elle, ma sanité d'esprit n'aurait pas été

celle qu'elle est désormais sans la solicitude, la compassion et l'amitié de mes collègues doctorants et/ou jeunes docteurs, de France et de Navarre, des Sciences du langage et de Games studies, aux premiers rangs desquels Julie R., avec qui j'ai eu la joie de découvrir l'envers du décor de la vie de laboratoire et tant d'autres choses encore. J'ai évidemment une pensée pour Lison, Natasha, Marc-Philippe et Chiara pour les échanges philosophico-intellectuels par tableau interposé (Thesaurus Regex nous en soit témoin), ainsi que pour Nataly, Julie H., Filip, Bénédicte, Silvia, Camilla, Karla, Daniele, Yizhe, Lena, et tous les autres, pour les soirées raclettes, les moments passés en conférence, et tous ces petits instants de répit qui redonnent un peu d'humanité et de souffle à cette période si particulière. Enfin, merci à Siegfried, Paul-Antoine, qui malgré la distance – scientifique et/ou géographique – m'ont aussi aidée à traverser tout ça.

Tout cela n'aurait évidemment pas été possible sans le soutien inconditionnel de ma famille, qui n'a eu de cesse de croire en moi toutes ces années durant, y compris lorsque je n'y croyais plus. Je leur dois tout, et plus encore. Merci aussi à mes amis pour leur affection, et particulièrement Noëlle et Roxane, Elise et Lucie (à nos annuelles retrouvailles au bar interne), Léna, parmi tant d'autres. Je ne le ferai évidemment pas car ça n'aurait pas de sens, mais si elles savaient lire, j'aurais remercié mon chat Mabel et mon chien Ozy, qui ont su rendre plus supportables les moments compliqués.

Enfin, à celles et ceux que je n'ai pas mentionnés ici, par manque de place ou de ressources cognitives à l'instant où je rédige ces humbles lignes, merci.

Résumé

La forme et le sens sont intimement liés en morphologie dérivationnelle, l'affixe d'un dérivé renseignant généralement sur la catégorie sémantique à laquelle le dérivé appartient. La relation entre affixes et catégories sémantiques n'est cependant pas exclusive, un affixe étant souvent associé à plusieurs catégories, et réciproquement. Les études abordent traditionnellement cette multiplicité des relations à l'aune de la rivalité entre les suffixes, c'est-à-dire leur capacité à former le même type de dérivés à partir d'un même type de base (*pavage* et *pavement* à partir de *paver*, *débiteuse* et *débitrice* à partir de *débiter*), avec l'objectif de faire émerger les spécificités de chaque suffixation expliquant leur coexistence en synchronie. Ces dernières années, les approches quantitatives se sont multipliées, visant notamment la modélisation de la concurrence comme une fonction de plusieurs facteurs, parmi lesquels des facteurs phonologiques, diachroniques, syntaxiques, ou encore sémantiques. Les facteurs sémantiques sont sans doute parmi les facteurs les plus difficiles à évaluer de façon empirique et statistique, et ont longtemps reposé sur une approche intuitive de la linguistique. Cette démarche, au travers de tests d'acceptabilité ou d'examen manuel en contexte, souffre cependant du faible degré de généralisation des observations, en l'absence d'un accès systématisé au sens. La sémantique distributionnelle se révèle depuis quelques années comme une des alternatives les plus populaires. Il s'agit d'une approche statistique du sens basée sur les usages en corpus, qui offre une représentation vectorielle du sens des mots. La quantification de la proximité sémantique des mots et la manipulation des représentations permises par les modèles distributionnels ouvrent de nouvelles perspectives sur l'analyse sémantique de la concurrence affixale. Nous mettons à profit dans cette thèse les espaces vectoriels distributionnels pour analyser des dérivés morphologiques au regard de ces relations many-to-many. Ce travail s'articule plus précisément autour de quatre grandes expériences construites autour de l'étude des noms d'agent et d'instrument déverbaux en *-eur*, *-euse* et *-rice* et des noms d'action déverbaux en *-age*, *-ion* et *-ment*. Dans un premier temps, nous quantifions la proximité sémantique entre membres de familles dérivationnelles à l'aide de

la proximité distributionnelle dans les espaces vectoriels, validant à grande échelle l'hypothèse d'une plus grande proximité du verbe et du nom d'action. Dans un second temps, nous corroborons les différences sémantiques entre les noms suffixés en *-eur*, *-euse* et *-rice* relativement à la caractérisation axiologique dépréciative de leur référent. Nous faisons émerger les spécificités de chaque suffixation par la comparaison de représentations unifiées calculées à partir de l'ensemble des membres de chaque classe. Dans un troisième temps, nous creusons les propriétés morphosémantiques des noms d'agent. En partant de la représentation unifiée des noms d'agent prototypiques en *-eur*, nous évaluons l'hétérogénéité morphologique et sémantique de la catégorie lexicale des noms d'agent. Enfin, dans un quatrième et dernier temps, nous explorons la différenciation sémantique des noms d'action en *-age*, *-ion* et *-ment*, que nous qualifions au regard de leur degré de technicité. Nous combinons des indices distributionnels et statistiques afin de modéliser cette différence de technicité. Au travers de ces quatre questions, cette thèse présente différents degrés d'adaptation des modèles distributionnels pour l'analyse linguistique, en tant qu'outil de validation et d'exploration. Nous proposons à ce titre une exploration méthodologique visant à illustrer le potentiel mais aussi les limites de l'utilisation des modèles distributionnels en linguistique.

Abstract

Form and meaning are closely related in derivational morphology, the affix of a derived word giving a cue on the semantic category to which the word belongs. The relationship between affixes and categories is not exclusive, an affix usually hinting at several categories, and reciprocally. Studies traditionally approach this many-to-many relationship through affix rivalry, that is their ability to derive the same type of words from similar base types (*pavage* and *pavement* 'paving' from *paver* 'pave', *débiteuse* 'female fast speaker', 'slicer' and *débitrice* 'female debtor' from *débiter* 'reel off', 'debit', 'produce'), and to coexist in synchrony. In recent years, quantitative approaches have grown popular. Their aim is to model affix rivalry as a multifactor function, including phonological, diachronic, syntactic or semantic factors. Among all these factors, semantic ones are arguably the most difficult to assess on a empirical and statistical basis, and they have long relied on intuition. This approach, along with the manual examination of contexts, suffers from a lack of generalization because they do not have a non-systemic access to meaning. Recently, distributional semantics established itself as one of the most popular alternatives, and can be described as a statistical approach of meaning based on corpus use which provides a vectorial representing word meaning. The ability of such models to quantify semantic proximity and to manipulate word representations provides new insights on the semantic analysis of affix rivalry. In this thesis, we use vector space models to analyze derived words in light of these many-to-many relationships. This work is structured around four main experiments built on the study of French deverbal agent and instrument nouns in *-eur*, *-euse* and *-rice*, and French deverbal action nouns in *-age*, *-ion* and *-ment*. First, we quantify the semantic proximity of lexemes belonging to derivational families based on their distributional proximity in vector space models, validating on a large scale the hypothesis that verbs and action nouns tend to be closer than other members in the families. Second, we confirm the semantic differences between *-eur*, *-euse* and *-rice* nouns induced by the depreciative axiological characteristics of their referents. We highlight the specificity of each suffixation through the comparison of unified

representations of classes of nouns. Third, we investigate the morphosemantic properties of agent nouns. Based on the unified representation of *-eur* prototypical agent nouns, we assess the semantic and morphological heterogeneity of the lexical category of agent nouns. Fourth, we examine the semantic differentiation of *-age*, *-ion* and *-ment* action nouns, which we approach based on their variable degree of technicality. We combine distributional and statistical clues to model the difference in technicality. Through these four axes, this thesis explores various degrees of adaptation of vector space models for linguistics research, as a validation and investigation tool. As such, we offer a methodological study aiming for the demonstration of the potential of vector space models and their limitations in linguistic studies.

Table des matières

Liste des tableaux

Liste des figures

Introduction

La forme et le sens sont intimement liés en morphologie dérivationnelle. L'affixe, lorsqu'il y en a un, est souvent un indice de la catégorie sémantique dont relève le mot. Par exemple, le suffixe *-eur* signalera souvent que le mot est un nom d'agent, et le suffixe *-age* un nom d'action. Pourtant, cette relation entre affixes et catégories sémantiques n'est pas toujours, voire rarement, exclusive. Un affixe donné est souvent associé à plusieurs catégories (agent et instrument pour *-eur* ; action, résultat, instrument, lieu, entre autres pour *-age*), et une catégorie peut être instanciée par plusieurs suffixes (*-eur*, *-ien*, *-ier*, entre autres, pour les noms d'agent ; *-age*, *-ion*, *-ment*, entre autres, pour les noms d'action).

Cette multiplicité des relations (désignée sous le nom de *many-to-many relationship* dans la littérature anglophone) amène à des situations de concurrence, aussi appelée rivalité, où des formes sont construites par deux suffixes à partir d'une même base (*pavage* et *pavement*, *débiteuse* et *débitrice*). Dans certains cas, les formes en concurrence désignent des réalités différentes, à l'image de *débiteuse* 'celle qui tient des propos qui ne sont pas dignes d'intérêt' et *débitrice* 'celle qui doit quelque chose' ; dans les autres cas – plus marginaux –, les deux formes s'avèrent sémantiquement identiques, comme *pavage*[1] et *pavement* 'action de paver'.

La concurrence affixale a fait l'objet de nombreuses études dans la littérature, au travers de la comparaison de deux affixes ou plus : *-ic* et *-ical*, *-ize* et *-ify*, *-ity* et *-ness* en anglais (Lindsay 2012), *-ción*, *-miento* et *-do*/*-da* en espagnol (Fábregas 2007), *-sk-*, *-n-* et *-Ov-* en russe (Bobkova et Montermini 2020), ou *-iste* et *-ien* (Lignon 2007), *-iser* et *-ifier* (Bonami et Thuilier 2019), et *-ité* et *-itude* en français (Koehl et Lignon 2014), entre autres. Ces travaux cherchent à évaluer si les affixes considérés sont réellement rivaux en synchronie, avec l'hypothèse sous-jacente que l'un des affixes est en train de remplacer l'autre en diachronie, ou s'ils occupent des niches distinctes qui ont leurs spécificités propres ? Il s'agit alors de décrire et d'expliquer ces spécificités à l'aide de facteurs formels, syntaxiques ou encore sémantiques. Cette question a notamment été abordée dans des approches quantitatives qui visent la systématisation et la modélisation de la concurrence comme une fonction de multiples facteurs – phonologiques, historiques, syntaxiques, etc. (Bonami et Thuilier 2019, Naccarato 2019, entre autres). De tous les aspects pouvant être envisagés dans la concurrence affixale, le facteur sémantique est l'un des plus difficiles à évaluer de façon empirique et statistique. Les travaux offrant une comparaison sémantique des procédés concurrents se fondent notamment sur l'introspection, au travers de tests linguistiques visant à évaluer

1. Nous ne tenons pas compte ici de la possible polysémie des formes, *pavage* étant aussi un nom collectif 'ensemble de pavés', contrairement à *pavement*.

l'acceptabilité de propositions. La portée de cette approche est potentiellement limitée, puisqu'elle repose sur l'examen d'un nombre limité d'exemples, et qu'elle est intrinsèquement soumise à la variabilité inter- (voire intra-) locuteurs. Tous les phénomènes linguistiques ne peuvent par ailleurs pas être formalisés à l'aide de tests linguistiques. D'autres études se sont construites autour des usages en corpus, et notamment sur l'analyse des contextes d'occurrences (Dal *et al.* 2018). Là encore, l'examen manuel des contextes ne permet pas un accès au sens systématique, et limite le degré de généralisation des observations. Le renouveau récent de l'approche distributionnelle permet de revoir les modalités d'analyse sémantique à grande échelle.

La sémantique distributionnelle est une approche statistique sur corpus, offrant une représentation du sens dans des espaces vectoriels. Elle permet une quantification de la proximité entre des items, et une manipulation des représentations. Cette approche quantitative offre de nouvelles perspectives sur de nombreux aspects sémantiques, et notamment sur la conccurrence affixale (Varvara *et al.* 2016, Missud 2019, Guzmá n Naranjo et Bonami 2020, entre autres). Nous mettons à profit cette approche pour analyser des dérivés morphologiques au regard des relations *many-to-many* évoquées ci-dessus. Plus précisément, nous comparons à l'intérieur des espaces vectoriels des procédés dérivationnels – les suffixations déverbales en *-eur*, *-euse* et *-rice* d'une part, et en *-age*, *-ion* et *-ment* d'autre part – et des catégories sémantiques, à savoir la classe des noms d'agent et celle des noms d'action. Cette confrontation s'instancie selon trois axes, déclinés au travers de trois questions distinctes qui articulent l'ensemble de ce travail.

Le premier axe que nous développons repose sur la confrontation des catégories lexicales, à savoir la classe des noms d'agent et celle des noms d'action, à l'aune de leur proximité à leur verbe de base. Cette première expérience implique l'opérationnalisation de l'hypothèse, communément admise, que les noms d'action sont sémantiquement plus proches de leur base verbale que ne le sont les noms d'agent, du fait de la non activation de l'opération sémantique lors de la dérivation (Chomsky 1970, Roché 2009). Nous évaluons la tendance des verbes à être plus proches de leur nom d'action que de leur nom d'agent en évaluant au sein d'un grand nombre de familles dérivationnelle la proximité distributionnelle entre le verbe, le nom d'agent et le nom d'action.

Le deuxième axe que nous développons est la confrontation de procédés dérivationnels, d'une part entre les suffixes *-eur*, *-euse* et *-rice* pour la deuxième expérience, et d'autre part entre les suffixes *-age*, *-ion* et *-ment* pour la quatrième et dernière expérience. Ces deux expériences visent la caractérisation des différences sémantiques entre les classes définies par ces suffixes, et plus précisément la validation empirique, et à l'échelles des classes, des différences évoquées dans la littérature – de genre et de connotation pour

les noms d'agent (Dawes 2003), et ontologique pour les noms d'action. Cette validation passe par la comparaison des classes dans les espaces vectoriels à l'aide d'une représentation unifiée dont nous examinons les propriétés morphologiques et sémantiques.

Le dernier axe développé repose sur la confrontation de procédés dérivationnels à une catégorie lexicale. Nous nous attachons dans la troisième expérience à caractériser la catégorie lexicale des noms d'agent à partir des noms d'agent en *-eur*. Nous évaluons au travers de cette étude l'extension morphologique et sémantique de la classe d'agent. Plus précisément, nous proposons d'identifier les noms d'agent potentiels sur la base de leur proximité sémantique à des noms d'agent avérés. Nous explorons par ailleurs l'organisation interne de cette catégorie en faisant émerger plusieurs sous-types de noms d'agent, que nous décrivons d'un point de vue morphologique, ontologique et référentiel.

Sur le plan méthodologique, les trois expériences témoignent d'une progression quant à l'intégration et l'exploitation des espaces vectoriels distributionnels dans le dispositif global. Nous exploitons dans la première expérience une propriété fondamentale des espaces vectoriels, à savoir le rapprochement spatial des mots sur la base de leur proximité sémantique, que nous appliquons de façon directe. La deuxième expérience exploite une seconde propriété des espaces vectoriels, à savoir l'uniformité de la représentation de certains traits sémantiques, pour proposer une approche unifiée d'un ensemble de mots. Cette méthode implique un dispositif plus complexe que celui de la première expérience, car elle met en jeu un vecteur moyen. Cette expérience, qui se décline en plusieurs tâches, permet la mise en place d'un contrôle bien plus important, tant sur les données lexicales et distributionnelles, que sur la granularité de l'analyse produite. La troisième expérience, enfin, intègre la sémantique distributionnelle dans un dispositif plus large qui combine plusieurs approches, dont la modélisation statistique.

La contribution de cette thèse est double. D'une part, elle apporte des réponses sémantiques sur la délimitation des classes et sur la caractérisation sémantique des suffixes. D'autre part, elle éprouve l'apport du dispositif distributionnel, en illustrant sa capacité à rendre compte d'un large ensemble de propriétés sémantiques, tout en déterminant les conditions d'une utilisation raisonnée et efficiente.

La partie I de cette thèse est consacrée à des considérations théoriques sur les espaces vectoriels, et notamment sur leur utilisation pour l'analyse sémantique de dérivés morphologiques. Dans la mesure où chaque question étudiée dans ce travail est caractérisée par ses propres problématiques, nous apportons les clés théoriques nécessaires à la contextualisation de nos expériences dans les parties respectives et non dans cette première partie. Les

parties II, III, IV et V présentent respectivement les trois études présentées précédemment.

L'ensemble des données présentées dans ce travail sont accessibles à l'adresse `https://github.com/mwauquier/PhdData`. Les données relatives à chacune des partie font l'objet de répertoires spécifiques, dont nous rappelons le lien au début de chaque partie.

Première partie

Contextualisation théorique et méthodologique

Ce travail s'articule autour de quatre questions linguistiques étudiées par le biais de la sémantique distributionnelle. Nous nous penchons dans un premier temps sur la validation de l'hypothèse d'une plus grande proximité entre le verbe et son nom d'action dérivé qu'entre le verbe et son nom d'agent dérivé. Nous étudions dans un second temps dans une approche comparative les suffixes *-eur*, *-euse* et *-rice* permettant de construire des noms d'agent et d'instrument déverbaux, afin de faire émerger sur le plan distributionnel les spécificités notamment axiologiques dépréciatives des noms féminins, et plus particulièrement des noms en *-euse*. La troisième expérience approfondit l'étude des noms d'agent en *-eur*, *-euse* et *-rice* afin d'apporter de nouvelles pistes quant à la circonscription de la catégorie lexicale des noms d'agent. Enfin, la quatrième expérience est dédiée à l'examen de la concurrence des noms d'action déverbaux en *-age*, *-ion* et *-ment*, dont nous montrons qu'ils diffèrent relativement à leur degré de technicité que nous modélisons statistiquement.

Ces quatre études se distinguent par les questions linguistiques qu'elles posent. Nous faisons le choix de présenter dans chaque partie l'état de l'art spécifique à la problématique développée. Les quatre thématiques abordées dans les expériences partagent cependant toutes quelque chose en commun, à savoir l'approche distributionnelle que nous adoptons.Dans cette approche, le sens des mots fait l'objet d'une représentation mathématique dans des espaces vectoriels calculés à partir des usages des mots en corpus. Nous utilisons les espaces vectoriels comme un outil d'exploration, de description et de validation, en exploitant le principe que l'organisation de l'espace vectoriel traduit la proximité sémantique des mots du corpus. Ce principe a fait l'objet d'implémentations et d'applications diverses, présentant chacune leur avantages et inconvénients. L'utilisation d'espaces vectoriels implique donc un certain nombre de choix méthodologiques.

Cette première partie est consacrée à la description des aspects méthodologiques transversaux à l'ensemble de ce travail. Dans un premier temps, nous présentons les principes sous jacents à la représentation du sens dans les espaces vectoriels (chapitre 1). Puis nous contextualisation notre approche et nos objectifs au regard de ce qui se fait dans l'analyse sémantique de dérivés morphologiques dans les espaces distributionnels (chapitre 2).

Chapitre 1

Représentation du sens dans les espaces vectoriels

La linguistique a besoin d'accéder à de nouvelles possibilités de modéliser le sens à partir des contextes d'apparition des mots. Les espaces vectoriels développés en sémantique distributionnelle offrent cet accès empirique et quantitatif au sens basé sur les usages en corpus. Mais cette approche est encore difficile à articuler avec les représentations classiques, notamment du fait du flou des méthodes de représentation qui rend l'interprétation des dimensions du sens captées par la seule information distributionnelle.

En guise de préambule, nous présentons l'approche distributionnelle et les propriétés remarquables des espaces vectoriels (section 1.1) avant de passer en revue les choix méthodologiques que nous opérons (section 1.2).

1.1 Approche statistique du sens

Les approches statistiques du sens utilisent des techniques mathématiques pour formaliser le sens. À ce titre, la sémantique distributionnelle propose de dériver de l'ensemble des contextes d'apparition des mots une représentation qui concentre ces informations. Ce mode de représentation présente de nombreux intérêts du fait du potentiel opératoire qu'il offre, au travers notamment de la comparaison et de la manipulation des représentations du sens des mots (Lenci 2018, Boleda 2020).

Cette approche a connu un renouveau récent avec le développement de nouveaux outils, démocratisant son utilisation, et multipliant ses implémentations. Nous présentons les principes qui la sous-tendent (section 1.1.1) avant d'aborder les aspects pratiques (section 1.1.2).

1.1.1 Propriétés des espaces vectoriels

La sémantique distributionnelle est une approche quantitative et statistique du sens, dans laquelle l'accès à l'information, notamment sémantique, se fait par le biais de la modélisation statistique des usages langagiers (Turney et Pantel 2010). Basée sur l'hypothèse distributionnelle (Harris 1954, Firth 1957), cette approche exploite l'idée que « the amount of meaning correspond[s] roughly to the amount of difference in their environments »(Harris 1954 : 157). En d'autres termes, plus la distribution de deux mots diffère, plus leur sens tend à différer, et réciproquement, les mots sémantiquement similaires auront tendance à partager un plus grand nombre de contextes. Conséquemment, la similarité sémantique de deux mots peut être évaluée sur la base du partage de contextes. De fait, l'approche distributionnelle établit une corrélation entre la similarité distributionnelle et la similarité sémantique (Sahlgren 2008, Lenci 2018, Boleda 2020), et les contextes des mots permettent de les représenter sous la forme de vecteurs dans un espace où la proximité spatiale traduit la similarité sémantique.

Nous présentons certaines propriétés des espaces vectoriels à l'aide d'un corpus fictif dont nous donnons la matrice de cooccurrence dans le tableau 1.1.

	gamelle	panier	croquette	souris	jeu-vidéo	clavier
chien	8	9	10	0	0	0
chat	8	6	9	12	7	1
geek	0	1	0	7	14	9

TABLE 1.1 – Matrice de cooccurence pour un corpus fictif

La matrice de cooccurence donnée dans le tableau 1.1 présente en ligne les mots cibles (*chien*, *chat*, *geek*) et en colonne les contextes (*gamelle* dans *gamelle du chien* ou *gamelle pour chien...*). Chaque contexte constitue une dimension, et l'ensemble des valeurs obtenues pour un mot (c'est-à-dire son nombre d'occurrences avec les différents contextes) constitue son vecteur. En tant que tel, un vecteur n'est pas directement interprétable, et son exploitation passe par sa comparaison à d'autres vecteurs. Plus deux mots cibles partagent des vecteurs similaires, plus les mots ont de chance d'être sémantiquement similaires, à l'image de *chien* et *chat* dans la matrice présentée dans le tableau 1.1. On observe que *chat* et *geek* ont des valeurs relativement similaires pour les dimensions *souris* et dans une certaine mesure *jeu-vidéo*, qui s'expliquent par la polysémie de *chat* et *souris*, qui peuvent tous deux désigner un animal et un dispositif informatique.

La description de la distribution de mots sous la forme de vecteurs (ou

word embedding) permet une spatialisation des mots selon plusieurs dimensions. Les mots proches sémantiquement sont de fait rapprochés dans l'espace, du fait de leurs vecteurs similaires. Ce mode de représentation présente un caractère intuitif du fait de la tendance humaine à la spatialisation de la proximité sémantique (Lakoff *et al.* 1999, cité par Sahlgren 2006). De fait, on peut représenter les mots cibles *chien*, *chat* et *geek* de notre corpus fictif dans un espace à 6 dimensions (réduit à deux dimensions dans la figure 1.1 pour des raisons de lisibilité).

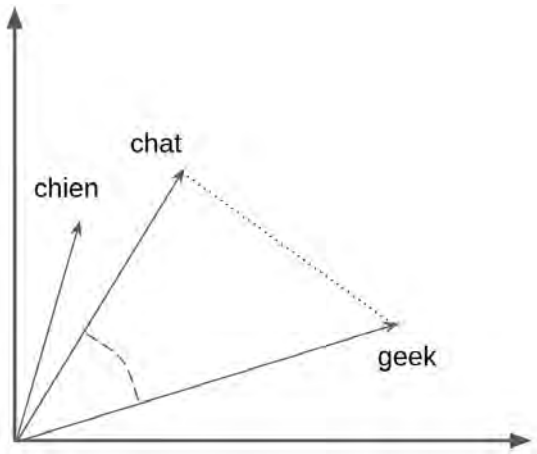

FIGURE 1.1 – Vecteurs des noms *chien*, *chat* et *geek* dans un espace vectoriel fictif

Le passage à des vecteurs, en tant qu'objets géométriques, permet la comparaison sémantique des mots représentés dans cet espace. La proximité des vecteurs – et donc des mots – est quantifiée de deux façons, à savoir la distance euclidienne (représentée par le trait discontu dans la figure 1.1) et le score cosinus (représenté par la ligne pointillée dans la figure 1.1). La distance euclidienne correspond à la distance entre les deux vecteurs, et est calculée de sorte à être minimisée à mesure que les items sont distributionnellement proches. Le score cosinus est quant à lui calculé à partir de l'angle entre les vecteurs, en faisant leur produit scalaire, de sorte à être maximisé à mesure que la distribution des mots se ressemble. La distance euclidienne et le score cosinus diffèrent principalement en ce qui concerne l'impact de la fréquence des mots sur le calcul de la proximité. Le score cosinus réduit l'impact de la fréquence, alors que la distance euclidienne en tient partiellement compte par le biais de la longueur des vecteurs [1]. Le score cosinus est la mesure la plus

1. Cela ne vaut que pour les modèles classiques, les modèles neuronaux comme Word2Vec(Mikolov *et al.* 2013a) impliquant une normalisation de la longueur des vec-

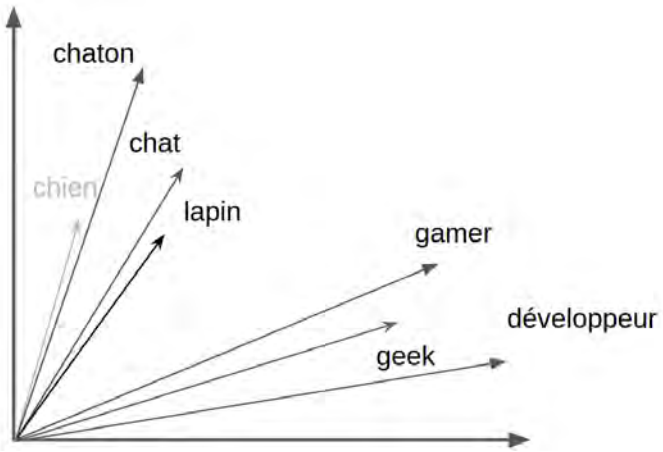

FIGURE 1.2 – 2 plus proches voisins de chat et geek dans un espace vectoriel fictif

largement utilisée, et est souvent la mesure optimisée par les algorithmes. Cette mesure est définie sur une échelle allant de 0, signifiant une différence complète, à 1, traduisant la stricte égalité des vecteurs comparés.

La figure 1.1 montre que les vecteurs de *chien* et *chat* sont plus proches entre eux qu'ils ne le sont de *geek*, et leur score cosinus est plus élevé que celui calculé avec *geek*. Notons que la valeur du score cosinus est difficile à interpréter dans l'absolu, et est utilisée de façon relative pour comparer différentes valeurs de proximité.Ces valeurs de proximité permettent d'identifier les vecteurs qui sont les plus proches des vecteurs ciblés, et de mettre au jour des mots proches, appelés voisins distributionnels. Dans le cadre de notre espace fictif, les voisins de *chat* et *geek* seraient respectivement *chaton* et *lapin*, et *gamer* et *développeur*. Nous illustrons cela à l'aide de la figure 1.2.

L'idée sous-jacente est que les vecteurs proches occupent des zones de l'espace vectoriel à partir desquelles on peut identifier des régions sémantiquement distinctes, cohérentes et définies (Mickus *et al.* 2020). Dans le cadre de notre exemple fictif, on identifie ainsi une zone occupée par les noms d'animaux d'une part, et une zone occupée par les noms relatifs à l'informatique d'autre part.

Outre la quantification de leur proximité qui permet d'identifier des voisins distributionnels, les vecteurs en tant qu'objets mathématiques peuvent être manipulés. De fait, la compositionnalité supposée des vecteurs est exploitée dans le cadre des modèles distributionnels compositionnels. Dans cette approche, les vecteurs sont combinés pour simuler la compositionnalité sé-

teurs.

mantique, et ainsi représenter des unités de sens complexes, allant des noms dérivés (voir chapitre 2) ou mots composés, aux syntagmes et phrases (Baroni et Zamparelli 2010, Mitchell et Lapata 2010, Baroni *et al.* 2014a). Diverses méthodes ont été proposées pour traduire mathématiquement la combinaison d'items lexicaux ou syntaxiques (Baroni et Zamparelli 2010, Mitchell et Lapata 2010, Lazaridou *et al.* 2013, entre autres). Ces méthodes sont généralement utilisées pour traduire mathématiquement la combinaison de mots ou de phrases, où le sens est construit d'un point de vue syntagmatique. Le sens de l'élément construit correspond à l'association des autres sens, de façon linéaire, avec ou sans pondération. Un premier type de méthodes, dites additives, consiste à calculer un vecteur $\overrightarrow{w}$ en additionnant les vecteurs $\overrightarrow{u}$ et $\overrightarrow{v}$, quels que soient les objets (mot, syntagme, phrase) représentés par ces vecteurs et la relation entretenue par les vecteurs $\overrightarrow{u}$ et $\overrightarrow{v}$. Les méthodes additives peuvent être simples (1.1) ou pondérées (1.2), auquel cas les vecteurs $\overrightarrow{u}$ et $\overrightarrow{v}$ se voient attribuer des poids α et β, afin de moduler l'importance donnée à chaque vecteur dans la composition du vecteur $\overrightarrow{w}$. Cette méthode vise à représenter des syntagmes comme *voiture rouge* à partir des vecteurs $\overrightarrow{voiture}$ et $\overrightarrow{rouge}$, et repose sur le postulat que le sens de *voiture_ rouge* peut être paraphrasé par *voiture qui est rouge*, qui peut être traduit par l'association linéaire du sens de *voiture* et du sens de l'adjectif *rouge*.

$$\overrightarrow{w} = \overrightarrow{u} + \overrightarrow{v} \tag{1.1}$$

$$\overrightarrow{w} = \alpha\overrightarrow{u} + \beta\overrightarrow{v} \tag{1.2}$$

D'autres méthodes permettent l'association linéaire de vecteurs, et notamment les méthodes dites dites multiplicatives, qui consistent à composer un vecteur $\overrightarrow{w}$ à partir du produit des vecteurs $\overrightarrow{u}$ et $\overrightarrow{v}$. Ces méthodes varient là encore du fait de la pondération des vecteurs, mais aussi du type de produit (scalaire, tensoriel, etc) utilisé.

Les approches compositionnelles offrent un panel plus large de méthodes, comprenant notamment des méthodes dites de dilatation (*dilation* en anglais), reposant sur la décomposition et la reconstruction de vecteurs, ou des méthodes à base de fonctions représentées sous la forme de matrices (Lazaridou *et al.* 2013). Les méthodes additives et multiplicatives, de par leur plus grande simplicité de mise en œuvre et leurs meilleurs résultats, sont cependant les méthodes les plus largement utilisées (Mitchell et Lapata 2010, Lazaridou *et al.* 2013) dans les approches lexicales, où l'ordre des composants n'est pas pertinent.

1.1.2 Évolution de la sémantique distributionnelle

Si les fondements de l'approche distributionnelle remontent notamment aux travaux de Harris (1954), ce mode de représentation vectorielle du sens a initialement été appliqué au traitement automatique des langues dans le cadre des tâches de recherche d'information (Salton *et al.* 1975). Dans cette approche, chaque document d'un ensemble donné est représenté dans un espace vectoriel. Pour déterminer le ou les documents jugés les plus similaires à une requête donnée, cette requête est à son tour représentée dans cet espace vectoriel, et est alors comparée aux autres représentations. On parle de modèles *document-based* (Fabre et Lenci 2015), ou de matrice *term-document* (Turney et Pantel 2010).

Ce mode de représentation a par la suite été étendu à d'autres types de contextes et d'objets, notamment les mots, menant à l'émergence de différents types de modèles. Certains sont basés sur la syntaxe, et prennent en compte les relations de dépendance syntaxique ou les motifs syntaxiques (*pair-pattern* chez Turney et Pantel 2010). Mais le type de modèles le plus largement utilisé, notamment dans les tâches sémantiques telles que le calcul de la similarité entre mots, le clustering ou la classification de mots, mais aussi des tâches de correction automatique, de désambiguïsation, d'étiquetage de rôles sémantiques, d'expansion de requêtes ou encore de génération automatique de thésaurus (Turney et Pantel 2010) est celui basé sur la cooccurence graphique, à savoir les modèles *word-context* ou *word-based*.

Ce type de modèles a fait l'objet de différentes implémentations. Ces implémentations diffèrent notamment par la méthode de construction des matrices et de calcul des vecteurs, comme la fréquence de cooccurrence, le *Pointwise Mutual Information* (PMI) ou sa variante le *Positive Pointwise Mutual Information* (PPMI), et le *Log-Likelihood Ratio* (LLR), entre autres (Turney et Pantel 2010). Les implémentations diffèrent aussi quant au traitement qui est fait de cette information distributionnelle, qui va généralement être condensée par le biais d'opérations de réduction des dimensions de la matrice, en se basant par exemple sur la redondance, la corrélation ou en amoindrissant l'impact de vecteurs considérés comme du bruit (Fabre et Lenci 2015). Cette réduction de dimensions peut être réalisée à l'aide de diverses méthodes, comme des algorithmes de décomposition en valeurs singulières (SVD) – la méthode la plus utilisée, notamment dans le cas des modèles d'analyse latente sémantique (LSA) (Landauer et Dumais 1997) – ou des algorithmes d'analyse en composantes principales (PCA) et de factorisation en matrices non-négatives (NMF), entre autres (Lenci 2018). Plus récemment, une nouvelle famille de modèles dits prédictifs, où l'information distributionnelle est déjà « condensée », est apparue (Lenci 2018). Ces mo-

dèles sont construits par apprentissage automatique non supervisé à l'aide d'algorithmes neuronaux. Ces algorithmes apprennent directement des représentations à partir de corpus, et cherchent à prédire les contextes d'un mot cible.

Ces modèles sont considérés comme plus obscurs que les précédents modèles, dits classiques, dans la mesure où l'apprentissage des représentations ne permet pas d'avoir accès aux contextes linguistiques discriminants [2]. C'est pour cela que l'on parle généralement de boîte noire (Han *et al.* 2011). L'utilisateur n'a accès qu'à une schématisation sémantique, en la nature d'un réseau sémantique ébauché par la proximité relative des vecteurs entre eux, ce qui nuit à la lisibilité des résultats. Certains auteurs, comme Levy et Goldberg (2014b), font l'hypothèse qu'une exploration des modèles distributionnels est possible par l'étude des contextes qui seraient activés par le mot cible, mais les travaux menés à cette fin n'ont pas encore abouti. Le positionnement relatif des vecteurs dans l'espace vectoriel et la quantification de leur proximité dans cet espace constituent une voie d'accès, indirecte, à l'information sémantique.

1.2 Les modèles distributionnels

Comme nous l'avons vu précédemment, plusieurs modèles et implémentations existent, chacun présentant leurs spécificités. Les modèles prédictifs se révèlent être plus intéressants que les modèles classiques dans la mesure où ils sont plus faciles à construire et à utiliser d'un point de vue computationnel, mais tout aussi, voire plus performants [3] (Baroni *et al.* 2014b, Bernier-Colborne et Drouin 2016), notamment dans la modélisation de la similarité des mots qui ont le même sens mais pas la même catégorie grammaticale (verbes, noms), à l'image des dérivés morphologiques et leurs bases. Dans ce travail, nous faisons le choix d'entraîner les modèles sur lesquels nous basons nos analyses avec Word2Vec (Mikolov *et al.* 2013a). Nous justifions ce choix dans la section 1.2.1 et nous présentons les paramètres que nous utilisons dans la section 1.2.2.

2. Notons que la réduction opérée sur les matrices dans les modèles classiques nuit elle aussi à la lisibilité et l'interprétabilité des modèles distributionnels, bien que les contextes puissent dans certains cas être retrouvés.

3. Lenci (2018) souligne néanmoins que cette plus grande efficacité des modèles prédictifs n'est pas systématique, et dépend notamment des paramètres utilisés pour leur entraînement.

1.2.1 Choix de l'outil

Les modèles distributionnels nous permettent d'analyser sur le plan sémantique les dérivés morphologiques sur la base de leur forme. Plusieurs outils sont disponibles pour ce faire, dont GloVe (Pennington *et al.* 2014), mais le plus populaire et le plus largement utilisé est sans doute Word2Vec (Mikolov *et al.* 2013a).

Word2Vec [4] procède à l'entraînement de modèles distributionnel par apprentissage automatique non supervisé à l'aide d'un réseau de neurones à une couche cachée. Les vecteurs sont initialisés aléatoirement par l'algorithme, et sont optimisés à mesure que le système rencontre des contextes de sorte à rapprocher les vecteurs qui partagent des contextes proches, et à éloigner les vecteurs qui n'en partagent pas. Le réseau de neurones transforme la tâche initiale, ici l'analyse distributionnelle, en une tâche de classification, puisque l'objectif du réseau de neurones est de prédire en couche de sortie une classe – ici un mot – sur la base des mots des contextes qu'il a rencontrés au niveau de la couche d'entrée (Han *et al.* 2011).

Word2Vec est l'algorithme le plus populaire, mais d'autres algorithmes ont depuis été développés. De nombreux travaux se sont attachés à adapter Word2Vec à d'autres types d'input. Levy et Goldberg (2014b) ont ainsi proposé une version modifiée de Word2Vec prenant en compte les relations de dépendances. D'autres modèles sont entraînés à partir de dictionnaires, ou de données limitées mais hautement informatives, comme les définitions (Tissier *et al.* 2017, Herbelot et Baroni 2017). Dans ce qui suit, nous utilisons Word2Vec dans sa version de base dans la mesure où nous souhaitons travailler sur les propriétés lexicales des dérivés morphologiques à partir des usages en corpus.

Des modèles inspirés de Word2Vec ont aussi cherché à enrichir les représentations distributionnelles par l'intégration d'informations dites morphologiques. L'exemple le plus notable est fastText (Bojanowski *et al.* 2016). Dans ces modèles, les mots sont représentés par des vecteurs qui sont la somme des vecteurs des n-grammes qui les composent. Le rapprochement des mots sur la base du partage de n-grammes est donc favorisé voire systématisé. Ce rapprochement formel se fait au détriment du sens, les performances de tâches sémantiques ne se voyant pas améliorées par l'intégration de ces composantes « morphologiques »(Avraham et Goldberg 2017). Puisque nous adoptons une approche sémantique de l'analyse de mots qui s'avèrent incidemment être des dérivés morphologiques, nous excluons l'utilisation d'outils mettant l'em-

4. Le code original en C est accessible à l'adresse `https://code.google.com/archive/p/word2vec/`. Une implémentation python est disponible dans `gensim` (Rehurek et Sojka 2010).

phase sur les caractéristiques morphologiques et plus précisément formelles.

Plus récemment encore, de nouveaux types de représentations vectorielles ont émergé, sous le nom de *contextualized word embeddings*, ou embeddings contextualisés. Ces représentations diffèrent des *word embeddings* classiques du fait de leur construction, puisque dans ces modèles, chaque occurrence fait l'objet d'une représentation vectorielle calculée à partir de la phrase dans laquelle se trouve l'occurrence (Peters *et al.* 2018). L'objectif d'une telle approche est de modéliser des phénomènes sémantiques plus complexes, à l'image de la polysémie, en permettant la représentation différenciée des différents sens d'un mot. Les représentants de ces approches sont ELMo (Peters *et al.* 2018) et BERT (Devlin *et al.* 2018).

De par leur nature, les *contextualized word embeddings* ne permettent pas d'analyser globalement le sens d'un mot. Cette incapacité à produire une généralisation pour une forme rend l'utilisation de ces modèles inadéquate dans ce travail. Par ailleurs, Mickus *et al.* (2020) montrent que la spatialisation des mots dans l'espace vectoriel en fonction de leur sens n'est pas aussi claire dans les modèles contextuels que dans les modèles prédictifs. Cela conforte notre choix d'utiliser Word2Vec.

1.2.2 Choix des paramètres

L'entraînement de modèles avec Word2Vec est conditionné par plusieurs paramètres. Certains sont relatifs au réseau de neurones et au déroulement de l'apprentissage, comme l'architecture et l'algorithme d'entraînement, et d'autres sont relatifs à l'espace vectoriel produit, comme le nombre de dimensions des vecteurs, la taille du corpus, le seuil de fréquence minimum, et la taille de la fenêtre.

L'architecture est ce qui définit la façon dont l'apprentissage se déroule. Dans le cas de Word2Vec, basé sur la cooccurrence, le choix se fait entre deux architectures : CBOW (Continuous Bag Of Word) ou Skip-gram.Dans les deux cas, les vecteurs sont initialisés de façon aléatoire. Puis les vecteurs sont mis à jour à partir des exemples rencontrés. Les deux architectures diffèrent relativement à la façon dont elles opèrent sur la mise à jour. L'architecture CBOW se sert de l'ensemble du contexte dans une fenêtre donnée autour du mot pour prédire le mot visé, sans tenir compte de l'ordre des mots au sein de cette fenêtre. L'architecture Skip-gram prédit quant à elle le contexte (c'est-à-dire les mots précédents et suivants) du mot visé. Ces deux architectures donnent des résultats variables en fonction des études et des tâches. Mikolov *et al.* (2013a) indiquent par exemple que l'architecture Skip-gram est plus performante que CBOW pour les mots peu fréquents, ou pour un nombre de dimensions et une fenêtre plus grands, mais cela se traduit par

un temps d'apprentissage beaucoup plus long. Bernier-Colborne et Drouin (2016) montrent quant à eux que les performances des architecturs dépendent de la tâche et notamment de la relation sémantique ciblée. les architectures CBOW et Skip-gram se montrent ainsi tour à tour plus performantes, mais elles obtiennent globalement des performances similaires. La comparaison avec des modèles *count-based* montre que l'architecture CBOW, bien que moins efficace que Skip-gram sur certaines tâches, se révèle néanmoins bien plus efficace que d'autres modèles de langage (Mikolov *et al.* 2013a). (Baroni *et al.* 2014b) confirment que l'architecture CBOW obtient systématiquement de meilleurs résultats que les modèles *count-based*, mais Ferret (2015) montre le contraire. Comparant des modèles classiques *count-based* à l'architecture Skip-gram, Levy *et al.* (2015) montrent qu'une optimisation fine des paramètres donne des résultats similaires entre les deux. En l'absence d'une nette différence, et du fait de ses avantages (rapidité, nombre faible de dimensions et fenêtre plus petite), nous choisissons l'architecture CBOW.

Quelle que soit l'architecture choisie, les mots traités en entrée passent par une couche neuronale cachée. La couche de classification *softmax* permet de transformer les informations de la couche cachée en probabilité et donc en prédiction, et plusieurs fonctions permettent de réduire le coût de calcul de cette couche. Les deux principales sont le *hierarchical softmax* et le *negative sampling*. Le *hierarchical softmax* est une extension de la fonction *softmax*, dont le but est de calculer la probabilité d'observer chaque mot présent dans le modèle afin de prédire le mot le plus probable (Faruqui 2016). Le *negative sampling* est une simplification du NCE (*Noise Constrastive Estimation*), qui vise à discriminer les données observées du bruit généré artificiellement lors de la création du modèle. L'optimisation et la mise à jour des vecteurs sont réalisés à l'aide d'exemples dits positifs, c'est-à-dire des contextes instanciés et extraits du corpus, et d'exemples dits négatifs, c'est-à-dire des contextes créés artificiellement en associant le mot cible à un mot avec lequel il ne coocurre jamais dans la fenêtre définie, et que le système va être entraîné à rejeter. Cette approche a pour but de mettre en avant les vecteurs similaires par contraste avec les vecteurs qui sont moins similaires (Goldberg et Levy 2014). Le *hierarchical softmax* est plus performant pour les mots fréquents et pour des vecteurs avec un nombre réduit de dimensions, et le *negative sampling* est plus performant pour les mots peu fréquents. Le *negative sampling* obtient par ailleurs globalement de meilleurs résultats (Bernier-Colborne et Drouin 2016). Nous choisissons donc d'utiliser la fonction *negative sampling*. Son utilisation requiert cependant de définir le nombre d'exemples négatifs, que nous fixons à 5, en l'absence de « contre-indications » dans la littérature à notre connaissance.

Parmi les paramètres liés à l'espace vectoriel, le nombre de dimensions des

vecteurs joue un impact sur la précision du modèle entraîné (Mikolov *et al.* 2013a). Plus leur nombre augmente, plus la précision augmente (Mikolov *et al.* 2013a, Bernier-Colborne et Drouin 2016). Mais cela augmente le temps de calcul, et nécessite de fournir davantage de données en entrée. Pour des raisons de coûts de calcul, le nombre de dimensions généralement utilisé est compris entre 100 et 300. Nous choisissons dans ce travail de fixer le nombre de dimensions à 100.

Le choix du corpus a un effet important sur l'analyse distributionnelle (Fabre et Lenci 2015). Schütze et Pedersen (1995) signalent ainsi que l'approximation de la similarité sur la base de la distribution ne vaut que si les contextes sont disponibles en quantité suffisante. Plus le corpus sera important, plus la similarité de certains mots sera perceptible dans le bruit que représentent de simples cooccurrences non significatives (Rychlỳ et Kilgarriff 2007). Le choix du corpus est aussi conditionné par sa nature : utiliser des corpus aussi larges et variés que possible permet de couvrir un plus grand champ lexical, de façon plus complète, ce qui améliore les performances des systèmes distributionnels. Nous revenons sur le choix du corpus dans la section 3.2.2.

La taille du corpus est liée à un autre critère important concernant l'entraînement de représentations vectorielles, à savoir la fréquence. Plus un mot sera fréquent, plus son vecteur sera robuste (Bullinaria et Levy 2007), puisque le système pourra construire sa représentation sur la base d'un plus grand nombre de contextes. Tanguy *et al.* (2015) montrent plus précisément que la fréquence du mot cible est le paramètre qui a l'impact le plus important sur la qualité des résultats de toute tâche exploitant un modèle distributionnel. Le seuil minimum à définir dépendra cependant de l'objet étudié et de la tâche à réaliser (Sahlgren 2006). Des outils comme Word2Vec proposent par défaut un seuil minimum de 5 occurrences, mais certains auteurs utilisent des seuils de 20 ou 50 occurrences (Sahlgren 2006). Certains auteurs choisissent quant à eux de ne pas définir arbitrairement un seuil de fréquence, mais de travailler sur le rang de fréquence, en analysant uniquement les 1 000, 2 000 ou 3 000 mots les plus fréquents par exemple. Augmenter le seuil de fréquence permet de garantir des représentations de bonne qualité, mais exclut aussi un plus grand nombre de mots, puisque seuls ceux qui apparaissent suffisamment souvent feront l'objet d'un vecteur. Comme nous souhaitons adopter une approche aussi extensive que possible, nous fixons le seuil de fréquence à 5 (sauf exception).

Le dernier critère largement discuté dans la littérature concerne la taille et la direction de la fenêtre. La fenêtre correspond à la zone autour du mot cible qui définit le contexte de chaque occurrence. Plus elle sera grande, plus un grand nombre de cooccurrences pourront être pris en compte. La direc-

tion de la fenêtre correspond à la région étudiée (gauche, droite, les deux). La définition de la taille et de la direction idéales de la fenêtre va elle aussi dépendre de la tâche visée. Si l'augmentation de la taille de la fenêtre contribue dans certaines tâches à une amélioration de la précision (Bernier-Colborne et Drouin 2016), ce n'est pas forcément pertinent, et peut potentiellement augmenter le bruit (Bullinaria et Levy 2007). Ainsi, des fenêtres réduites seraient plus efficaces pour l'acquisition d'informations sémantiques (Sahlgren 2006), favorisant une approche paradigmatique. Il n'y a pas de règles en la matière, les performances des modèles dépendant de la taille de la fenêtre, de la tâche, ou encore de la mesure d'association utilisée. Nous faisons ici le choix de fixer la taille à 5, qui est la valeur recommandée pour l'architecture CBOW.

D'autres paramètres peuvent être définis pour Word2Vec, comme le sous-échantillonnage des mots fréquents. Dans la mesure où nous n'avons pas d'hypothèses spécifiques au sujet de ce critère, et comme il n'a pas été à notre connaissance contesté dans la littérature, nous prenons la valeur par défaut, fixant à 10^{-3} la probabilité d'un mot d'être conservé.

Chapitre 2

Sémantique distributionnelle et dérivés morphologiques

L'objectif de ce travail est d'étudier et d'analyser sur le plan sémantique des dérivés morphologiques. Nous nous situons donc à l'interface entre morphologie et sémantique. Si la sémantique distributionnelle se veut avant tout une modélisation du sens, des liens avec la morphologie ont été faits à de nombreuses reprises, en témoignent notamment les modèles intégrant des informations dites 'sub-word' (voir section 1.2.1).

Ce chapitre présente le positionnement de notre travail au regard des travaux situés à l'interface entre morphologie et sémantique, tant sur le plan théorique que méthodologique. Les travaux proches se divisent en deux grandes familles, qui traduisent la relation mutuelle entretenue par la morphologie et la sémantique distributionnelle. On identifie d'une part des travaux qui utilisent les liens sémantiques et morphologiques entre lexèmes pour contribuer à l'amélioration des représentations distributionnelles (section 2.1), et d'autre part des travaux qui utilisent les représentations distributionnelles pour analyser sémantiquement des lexèmes caractérisés par certaines propriétés morphologiques (section 2.2). C'est dans le 2^{e} groupe que se place notre travail. Notons que ces deux approches ne s'excluent pas. Nous ciblons ici les études et les aspects les plus saillants relativement à cette interaction.

2.1 La morphologie au service de la sémantique distributionnelle

Un premier ensemble d'études exploite les liens morphologiques entre des lexèmes comme indice de lien sémantique. Ce lien est alors utilisé pour inter-

venir sur la représentation de ces lexèmes dans les espaces distributionnels, afin d'améliorer les performances des tâches dans lesquelles ces modèles sont utilisés, et pour proposer des représentations vectorielles de mots initialement absents des modèles distributionnels, parce qu'absents des corpus ou ayant une fréquence trop faible. Dans ces deux cas de figure, l'information sémantique extrapolée des propriétés morphologiques des lexèmes sert à enrichir ou à construire la représentation des autres mots. Ces travaux reposent pour la majeure partie sur l'hypothèse que la dérivation est une opération compositionnelle, conçue comme l'ajout du sens d'un affixe au sens de la base, et que cela peut se traduire en opérations de composition des vecteurs.

C'est notamment le cas de Luong *et al.* (2013), qui représentent par des vecteurs les morphèmes qu'ils ont identifiés dans une ressource extérieure. Le corpus d'entraînement est segmenté morphologiquement, et chaque unité minimale est ensuite modélisée sous le forme d'un vecteur dans l'espace distributionnel. Les vecteurs des mots complexes sont alors obtenus par composition à partir des vecteurs des morphèmes, de façon récursive. Ainsi, le vecteur du mot *unfortunately* est obtenu par la composition du vecteur du mot *unfortunate*, lui-même obtenu à partir des vecteurs du morphème *un-* et du mot *fortunate*, avec le vecteur du morphème *-ly*. Cette approche permet aux auteurs de mieux traiter les mots complexes rares, puisque leur représentation se base dès lors sur celle de morphèmes, bien plus fréquents. Botha et Blunsom (2014) utilisent la même approche appliquée à des modèles de langue. Le vecteur du mot *imperfection* est ainsi calculé à partir des vecteurs des morphèmes *im-*, *perfect* et *-ion*.

Soricut et Och (2015) développent une approche similaire, dans la mesure où il s'agit d'obtenir un vecteur correspondant à une affixation, mais elle diffère dans son opérationnalisation. Les auteurs ne cherchent pas à créer un vecteur pour un affixe donné à partir des coocurrences de cet affixe, qui pourrait être utilisé pour construire la représentation de tout nom porteur de ce n-gramme. Ici, il s'agit de représenter la transformation morphologique liant une base à son dérivé. Cette représentation est apprise à partir de paires base-dérivé acquises automatiquement dans le modèle, et pour lesquelles un lien sémantique peut être établi (au moyen d'un score de proximité). La présence d'un lien sémantique permet de garantir que l'apprentissage de la transformation repose sur des paires pertinentes comme *officially-official* et non sur des paires comme *only-on* dans le cadre de la transformation associée à *-ly*. La construction d'une représentation de la transformation associée à *-ly* permet alors de calculer une représentation pour les mots en *-ly* rares ou absents du corpus.

Cette approche compositionnelle est au cœur des modèles entraînés par fastText (Bojanowski *et al.* 2016), se détachant ici cependant de la notion

d'affixes ou de morphèmes pour s'appliquer à l'échelle des n-grammes. La représentation vectorielle des mots est ainsi vue comme la composition des vecteurs des n-grammes constitutifs du mot.

Baroni et Zamparelli (2010) proposent quant à eux de ne pas représenter l'affixe par un vecteur, mais par une matrice. Pour cela, ils étendent aux paires base-dérivé la méthodologie qu'ils ont utilisée pour représenter des syntagmes adjectif-noms. Les noms étaient alors représentés par des vecteurs, et les adjectifs par des matrices constituées des noms qu'ils peuvent modifier. Le vecteur du syntagme adjectif-nom était alors obtenu en multipliant le vecteur du nom avec la matrice de l'adjectif, elle-même calculée à partir des vecteurs des noms avec lesquels il cooccurre. L'application de cette méthode à l'affixation fait que l'affixe n'est pas représenté par un vecteur, mais par l'ensemble des vecteurs des paires base-dérivés impliquant cette affixation.

Dans leur comparaison des méthodes compositionnelles (voir section 1.1.1), Lazaridou *et al.* (2013) reprennent l'idée d'une représentation des affixes comme *re-* (pour la préfixation verbale à l'origine de *redo* ou *remake*) sous la forme d'une matrice calculée à partir des vecteurs des paires V-reV fréquentes. La méthode présentée par Lazaridou *et al.* (2013) diffère cependant de celle de Baroni et Zamparelli (2010) en cela que la composition du vecteur du dérivé par l'association du verbe de la base et de la matrice de l'affixe fait l'objet d'une pondération afin que la représentation compositionnelle du dérivé soit la plus proche possible de sa représentation apprise. Cette méthode leur permet d'améliorer la représentation des noms dérivés, mais elle n'a pas fait l'objet d'une exploitation en morphologie.

Une autre méthode qui repose sur l'hypothèse compositionnelle est l'analogie. Mikolov *et al.* (2013a) et Mikolov *et al.* (2013b) proposent de prédire par analogie des dérivés morphologiques en partant du postulat que certaines relations linguistiques (lexicales ou morphologiques) peuvent être représentées par des vecteurs. Les auteurs montrent que l'on peut prédire le terme *queen* à partir d'un couple initial, tel que *man* et *woman*, et d'un troisième terme, comme *king*. Les auteurs montrent que cette application de l'analogie est opérationnalisable pour représenter des relations morphologiques, notamment la formation d'adjectifs comparatifs ou superlatifs, visant notamment à prédire la forme *rougher* pour la base *rough* à partir de la paire formée par l'adjectif *good* et de l'adjectif comparatif *better*.

Le calcul ou l'amélioration de la représentation de mots rares ou absents proposé par Padó *et al.* (2013) ne repose pas sur la compositionnalité des vecteurs, mais sur la représentation d'autres membres de la famille dérivationnelle du mot cible. L'idée est ici d'améliorer les représentation des mots rares ou absents dans les modèles distributionnels en se servant de la forte similarité sémantique entre membres des familles dérivationnelles. La repré-

sentation d'un mot appartenant à la même famille peut servir d'approximation pour représenter un mot rare ou absent. La représentation de *oldish* est calculée en utilisant la représentation de *old*.

Ces différentes études montrent la multitude des approches proposées pour représenter un affixe dans les modèles distributionnels, ou pour calculer la représentation de dérivés morphologiques. Notons qu'aucune ne s'est imposée comme une méthode stantard, à l'exception de fastText, qui reste cependant assez marginalement utilisée. Si cela montre la capacité des modèles distributionnels à capter des régularités sémantiques associées à des propriétés morphologiques, notre étude ne vise pas les mêmes objectifs. Nous souhaitons étudier la valeur sémantique effective des mots, à travers leurs emplois (y compris ses idiosyncrasies), et pas la calculer sur la base d'un fonctionnement compositionnel idéal (voir Marelli et Baroni 2015 et l'opposition qu'ils dressent entre *full-form meaning* et *combinatorial meaning*). Nous n'envisageons l'hypothèse d'une compositionnalité des vecteurs que comme un outil de représentation uniforme de certaines caractéristiques sémantiques, comme nous allons le voir dans la section 2.2.

2.2 La sémantique distributionnelle pour l'analyse des dérivés morphologiques

Dans la deuxième approche, les dérivés sont des objets à analyser par le biais de la sémantique distributionnelle. Certaines études prennent comme point d'entrée les procédés dérivationnels, et d'autres les catégories sémantiques.

Certaines études comparent les dérivés en fonction de leur procédé. Le point d'entrée est alors la forme, en cherchant par exemple à comparer sémantiquement des mots construits par différents procédés.

Varvara *et al.* (2016) examinent par exemple la concurrence entre deux procédés de nominalisation en allemand. Leur hypothèse est que le glissement sémantique opéré lors de la dérivation des infinitifs nominaux et des noms déverbaux suffixés en *-ung* n'est pas représenté de la même façon dans les modèles distributionnels, les premiers se rapprochant de noms fléchis comme les participes présents, et les seconds se rapprochant des noms d'agent déverbaux en *-er*. Les auteurs comparent donc ces quatre groupes de mots à l'aide de deux critères, la transparence – c'est-à-dire le degré de similarité entre le dérivé et sa base – et la spécificité – c'est-à-dire la conservation ou l'ajout d'information sémantique – calculés à partir des informations distributionnelles. Les auteurs montrent à partir de ces indices distributionnels que

les quatre procédés se distinguent dans les modèles distributionnels relativement à leur transparence et à leur spécificité. Si la discrimination ne repose qu'indirectement sur les représentations vectorielles, elle montre néanmoins la capcité à comparer des ensembles de noms formés par des procédés définis dans les espaces distributionnels.

Le français a aussi fait l'objet de comparaisons de procédés dérivationnels dans les modèles distributionnels. Citons Missud (2019) qui compare les noms d'action suffixés en *-age* avec ceux converts de verbe, ou Guzmá n Naranjo et Bonami (2020) qui évaluent la proximité d'une trentaine de procédés dérivationnels.

Zeller *et al.* (2014) utilisent quant à eux une approche distributionnelle pour valider sémantiquement une ressource morphologique de l'allemand, DErivBase (Zeller *et al.* 2013). Ils vérifient si les mots rapprochés sur le plan dérivationnel dans la ressource le sont aussi sur le plan sémantique grâce au cosinus des vecteurs de ces mots dans l'espace vectoriel. Ce cosinus est ensuite traité par un classifieur qui détermine, dans le cadre d'une tâche binaire, s'il faut considérer cette paire comme sémantiquement valide ou non.

Plusieurs travaux se sont intéressés à l'identification de l'orientation de certains procédés dérivationnels. Kisselew *et al.* (2016) se sont notamment intéressés à la conversion, pour laquelle l'orientation n'est pas identifiable (Tribout 2010). Dans ce travail, l'identification de l'orientation repose sur des critères quantitatifs tirés de l'hypothèse que les noms dérivés tendent à être moins fréquents et sémantiquement plus spécifiques que leur base. La fréquence des items est calculée directement en corpus, et la spécificité est calculée à partir des représentations distributionnelles. La combinaison de ces deux critères permet de diagnostiquer de façon relativement satisfaisante les conversions Nom-Verbe et Verbe-Nom. Cette étude fait suite à un précédent travail sur l'orientation des procédés dérivationnels, mais qui ne s'attachait pas spécifiquement à la conversion (Padó *et al.* 2015)

D'autres travaux utilisent la sémantique distributionnelle pour étudier des dérivés morphologiques en termes purement sémantiques. Verhoeven *et al.* (2012) procèdent à la classification automatique de noms composés néerlandais et afrikaans en 6 classes sémantiques. Cette classification repose sur les représentations vectorielles des noms, fournies à un classifieur de type SVM. Les auteurs font ainsi l'hypothèse que la catégorie sémantique d'un nom composé peut être prédite en comparant ce nom à d'autres composés sémantiquement similaires.

Lapesa *et al.* (2018) utilisent quant à eux les représentations vectorielles pour désambiguiser en contexte des nominalisations néologiques en *-ment* pour l'anglais. La désambiguisation est à ce titre envisagée comme une tâche de classification visant à prédire le caractère événementiel ou non événe-

mentiel du nom sur la base de vecteurs d'occurrences de nominalisations néologiques.

Ces études confirment que les espaces vectoriels permettent de capter des phénomènes à l'interface entre morphologie et sémantique. L'examen de la proximité des dérivés et la comparaison des procédés à partir de propriétés sémantiques sur le plan distributionnel semblent donc tout à fait pertinents. Cet outil nous semble approprié pour l'étude extensive et systématique que nous envisageons.

Deuxième partie

Proximité sémantique de dérivés morphologiques

Nous nous intéressons à la proximité sémantique entre les lexèmes qui appartiennent à une même famille dérivationnelle, en considérant plus particulièrement les verbes et leurs noms d'agent et d'action dérivés. Le point de départ de ce travail est le souhait de tester l'hypothèse de Roché (2009) selon laquelle le sens d'un verbe (*protéger*) et de son nom d'action dérivé (*protection*) seraient identiques : la variation formelle et catégorielle, simple conséquence d'un choix de construction syntaxique, ne s'accompagnerait pas d'une variation sémantique significative. *A contrario*, la dérivation agentive implique une opération sémantique puisqu'elle vise à former le nom qui dénotera l'entité réalisant l'action dénotée par le verbe (*protecteur*). Certains voient dans cette opération un ajout de sens au lexème initial (Laca 2001). Cette position implique donc que le sens du nom d'agent construit diffère de celui de la base et, par conséquence, du nom d'action dérivé.

La proximité sémantique entre le verbe et ses dérivés a fait l'objet de nombreux travaux, principalement focalisés sur la prise en compte de 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 généralement sur l'application de tests d'acceptabilité, éventuellement complétés par des procédures d'annotation de corpus (Balvet *et al.* 2011). Nous avons fait le choix d'une autre approche, visant à considérer l'usage de ces lexèmes en corpus, en nous appuyant sur le critère de proximité distributionnelle comme indice de similarité sémantique. Nous cherchons à vérifier si l'hypothèse d'une plus grande proximité du verbe et du nom d'action se traduit en termes de profil distributionnel des lexèmes en corpus, et ce de façon homogène. Pour cela, nous considérons les liens de proximité sémantique au sein des triplets (verbe, nom d'agent, nom d'action) comme (*poncer*, *ponceur*, *ponçage*). L'hypothèse est que le verbe (*poncer*) est plus proche sur le plan distributionnel de son nom d'action (*ponçage*) qu'il ne l'est de son nom d'agent (*ponceur*). Il s'agit donc ici de comparer la proximité des membres entre eux, tâche pour laquelle les modèles distributionnels sont parfaitement adaptés.

L'objectif de cette première étude est de vérifier de façon empirique, sur le plan distributionnel, la validité de l'hypothèse d'une plus forte proximité sémantique entre le verbe et son nom d'action au sein des familles dérivationnelles. Nous exploitons pour cela le principe même des modèles distributionnels, que nous avons exposé dans le chapitre 1, à savoir la proximité des vecteurs dans l'espace vectoriel comme conséquence d'une proximité distributionnelle et donc sémantique. Cette approche nous permet de quantifier la proximité des lexèmes entre eux. Nous comparons à ce titre sur le plan distributionnel 2 585 triplets composés d'un verbe, d'un nom d'agent ou d'instrument et d'un nom d'action issus d'une même famille dérivationnelle,

en calculant leurs scores de proximité dans un espace vectoriel.

Nous consacrons le chapitre 3 à l'opérationnalisation de l'hypothèse. Nous contextualisons la question de la comparaison sémantique des nominalisations, et justifions notre recours aux espaces vectoriels. Nous présentons ensuite les ressources lexicales et distributionnelles sur lesquelles nous basons notre analyse. Nous décrivons dans un troisième temps les résultats permettant de corroborer l'hypothèse. Nous nous arrêtons plus longuement dans le chapitre 4 sur les résultats, en analysant de façon plus qualitative un nombre limité de triplets, que nous sélectionnons sur la base de leurs scores de proximité remarquables. Cela nous permet de mettre au jour certaines limites et précautions liées à l'utilisation de modèles distributionnels.

Les contributions de cette étude sont à la fois linguistiques et méthodologiques. D'une part, nous confirmons de façon empirique et à grande échelle que les verbes tendent effectivement à être sémantiquement plus proches des noms d'action que des noms d'agent ou d'instrument. Nous apportons une quantification, qui permet d'évaluer l'ampleur du phénomène sur plus de 2 500 cas. L'analyse plus spécifique de certains triplets nous permet de mettre en lumière certains phénomènes à l'œuvre dans les cas où l'hypothèse est invalidée, et incidemment les limites liées aux espaces vectoriels.

D'autre part, nous illustrons de façon concrète l'utilisation de modèles distributionnels pour l'analyse de dérivés morphologiques dans une approche quantitative. Nous proposons une première exploitation, caractérisée par une opérationnalisation simple d'une hypothèse, avec un minimum de contrôle et de traitement, tant computationnel que linguistique. Nous montrons l'influence du niveau d'analyse – global ou local – sur l'adéquation de l'outil, les granularités plus fines d'analyse invitant à un raffinement de la méthode ou des données.

Les données présentées dans cette partie sont accessibles à l'adresse `https://github.com/mwauquier/PhdData/tree/main/Part2`.

Chapitre 3

Validation de l'hypothèse

La nominalisation permet de construire des noms d'agent (*acheteur*, *danseur*), d'instrument (*essoreuse*, *étendoir*), d'action (*ponçage*, *affrontement*), de lieu (*lavoir*, *brasserie*), entre autres, à partir d'un verbe par le biais d'opérations formelle, catégorielle et possiblement sémantique. Les noms d'action diffèrent des autres dérivés notamment par la non activation de l'opération sémantique. Ils sont à ce titre considérés comme les lexèmes sémantiquement les plus similaires au verbe relativement aux autres dérivés comme les noms d'agent et d'instrument. Ce postulat est cependant difficile à évaluer du fait de la différence catégorielle entre le verbe et ses dérivés, les relations sémantiques traditionnelles telles que la synonymie ne s'appliquant donc pas, et en l'absence d'outils permettant la quantification du sens.

Notre objectif est d'évaluer empiriquement ce phénomène à l'échelle du lexique, dans un grand nombre de familles dérivationnelles. Nous utilisons pour cela les espaces vectoriels qui permettent d'expérimenter l'hypothèse en fournissant une quantification de la proximité sémantique. S'il se confirme que le verbe est plus proche du nom d'action que du nom d'agent, alors le score de proximité entre le verbe et le nom d'action dans l'espace vectoriel sera plus élevé que celui entre le verbe et le nom d'action. La comparaison systématique des scores de proximité entre les membres de 2 585 triplets formés de noms d'agent, verbes et noms d'action montre que le verbe tend effectivement à être plus proche du nom d'action que du nom d'agent, tant en termes de proportion de triplets validant l'hypothèse que de moyenne de proximité à l'échelle des 2 585 triplets.

Nous présentons dans un premier temps l'hypothèse et les difficultés quant à son évaluation empirique (section 3.1). Nous détaillons dans un second temps (section 3.2) le dispositif sur lequel nous basons l'analyse que nous commentons dans la section 3.3.

3.1 Les dérivés morphologiques déverbaux

La nominalisation implique des opérations qui ne sont pas toutes activées en fonction du type de nom ciblé. Cela implique des variations sémantiques plus ou moins systématisées entre dérivés, mais aussi vis-à-vis de la base. Nous passons en revue quelques travaux qui examinent l'impact sémantique de la nominalisation (section 3.1.1), puis nous présentons succinctement les propriétés des noms d'agent et d'instrument (section 3.1.2) ainsi que des noms d'action (section 3.1.3).

3.1.1 Nominalisation et opération sémantique

La nominalisation couvre les procédés dérivationnels permettant de créer des noms à partir de verbes. Ces procédés interviennent théoriquement sur trois niveaux : formel, catégoriel et sémantique (Roché 2009). Par exemple, la formation du nom d'agent *chanteur* à partir du verbe *chanter* se traduit par l'ajout du suffixe (œʁ), le passage d'un verbe à un nom, et la dénotation de l'agent qui réalise l'action dénotée par le verbe de base, ici l'action de chanter.

Le terme de nominalisation est communément utilisé pour les noms d'action, mais certains auteurs distinguent en réalité plusieurs types de nominalisations, dont la nominalisation active, la nominalisation résultative, la nominalisation agentive, et la nominalisation instrumentale (Dubois et Dubois-Charlier 1999). Du fait de l'hypothèse sur laquelle nous basons notre étude, nous nous concentrons dans ce travail plus spécialement sur la nominalisation dite active, c'est-à-dire celle des noms d'action, la nominalisation agentive, des noms d'agent, et la nominalisation instrumentale, des noms d'instrument (Dubois et Dubois-Charlier 1999).

Le terme de nominalisation recouvre plusieurs réalités puisqu'il désigne toute réalisation d'une proposition par une expression nominale (Chomsky 1970). Cette transformation est qualifiée d'enchâssement d'une des deux propositions dans le groupe nominal de la seconde proposition chez Dubois et Dubois-Charlier (1999). Si la nominalisation de la phrase 1 en anglais peut reposer sur trois types de procédés, à savoir le gérondif (1a), la dérivation (1b) et une approche mixte (1c) (exemples tirés de Chomsky (1970)), elle passe en français principalement par la dérivation ((2a)).

(1) John refused the offer

   a. John's refusing the offer

   b. John's refusal of the offer

   c. John's refusing of the offer

(2) Jean a refusé l'offre

a. le refus (par Jean) de l'offre

Selon Chomsky (1970), toutes les propositions n'ont pas forcément d'équivalence nominale, et la relation qui unit la proposition initiale et le syntagme nominal qui en découle varie énormément, et dépend de la proposition et du syntagme étudiés. Dubois et Dubois-Charlier (1999) suggèrent ainsi que le contenu enchâssé et son existence dans la proposition d'origine distinguent notamment la nominalisation active des nominalisations agentives et instrumentales. Les auteurs soutiennent ainsi que les nominalisations actives et résultatives conservent, par leur enchâssement, les arguments du verbe (bien que l'argument SN0 du verbe devienne facultatif). A l'inverse, la nominalisation agentive et instrumentale implique l'enchâssement d'un groupe nominal implicite, absent de la proposition initiale.

De fait, l'étude des dérivés déverbaux s'est souvent faite au regard de la préservation de la structure argumentale du verbe (Grimshaw 1990, Balvet *et al.* 2011, Condette *et al.* 2012) et de l'héritage de propriétés sémantiques, en particulier aspectuelles, du verbe par le nom (Haas *et al.* 2008). Ces travaux se fondent généralement sur l'application de tests d'acceptabilité, éventuellement complétés par des procédures d'annotation de corpus (Balvet *et al.* 2011).

Les opérations sémantique et formelle liées à la nominalisation – et plus largement à la dérivation – sont analysées différemment selon les auteurs. D'une part, certains auteurs soutiennent l'idée que le dérivé est toujours sémantiquement plus riche (Laca 2001), et que la dérivation peut ajouter du contenu sémantique ou le maintenir, mais jamais en retirer (Koontz-Garboden 2007). Roché (2009) nuance quant à lui cette idée, en affirmant qu'il n'y a pas d'ajout de contenu sémantique dans le cas de la nominalisation active, mais qu'il peut y avoir une perte de contenu. Le contenu sémantique est alors au mieux conservé, mais avec d'éventuelles restrictions, dans les cas où la base se révèle polysémique. La dérivation implique alors la sélection d'un sens spécifique.

De fait, lors de la dérivation de nominalisations actives et agentives ou instrumentales sur une même base, ces nominalisations diffèrent quant à leur relation vis-à-vis de la base. Leur appartenance à un même paradigme dérivationnel se traduit cependant par le partage de certaines composantes et propriétés sémantiques. Les relations morphologiques représentent ainsi un outil intéressant, puisqu'elles permettent de garantir un minimum de proximité sémantique. Cette proximité sémantique est doublée d'une proximité formelle, ce qui en facilite l'identification. Sauf cas spécifique, par exemple lié à la lexicalisation ou à la synonymie, le verbe d'une famille donnée sera

plus proche du nom d'agent issu de cette même famille que du nom d'agent d'une autre famille. Mais la comparaison au sein de la famille reste néanmoins limitée par la différence catégorielle.

Pour passer outre cette difficulté, nous nous proposons d'utiliser la sémantique distributionnelle, qui nous offre un point d'entrée différent sur le sens des dérivés morphologiques.

3.1.2 Les nominalisations agentives et instrumentales

Il est communément admis qu'un nom d'agent est un nom dérivé d'un verbe par l'ajout d'un suffixe, et moins couramment d'un autre nom, comme dans le cas de *bridgeur* qui dérive de *bridge*. Huyghe et Tribout définissent ainsi le nom d'agent comme étant un « nom déverbal qui dénote l'entité animée réalisant intentionnellement l'action décrite par le verbe de base »(2015 : 3).

On peut noter une certaine similarité des nominalisations agentives et instrumentales dans la mesure où les noms formés désignent l'entité (personne ou objet) réalisant l'action dénotée par la base. Huyghe et Tribout définissent le nom d'instrument comme un « nom déverbal qui dénote l'artefact prototypiquement utilisé pour réaliser l'action décrite par le verbe de base »(2015 : 3). Ainsi, il est communément admis que les noms d'agent et d'instrument déverbaux diffèrent par le caractère animé, voire humain, de l'entité à l'origine de l'action et par le caractère intentionnel de l'action réalisée (Huyghe et Tribout 2015). La distinction entre les deux types de noms n'est cependant pas toujours évidente, les tests utilisés pour diagnostiquer les uns et les autres ne permettant pas toujours de trancher clairement (Huyghe et Tribout 2015). À titre d'exemple, Huyghe et Tribout (2015) ont annoté le caractère agentif ou instrumental de 1 547 noms en *-eur*, *-euse* et *-rice* dérivés de verbes dynamiques extraits de Lexique3 [1]. Il ressort de l'annotation que 70% des noms sont des noms d'agent, 9% des noms d'instrument, 10% ont la double lecture agentive et instrumentale, et 12% sont indéterminés (le nom non instrumental *successeur* étant par exemple sous-déterminé concernant le trait animé). Ces derniers montrent parfois la difficulté de diagnostiquer l'une ou l'autre classe.

Bien que cela ne soit pas toujours clairement discuté (nous revenons plus en détail sur ce point dans le chapitre 8), les suffixes qui marquent la nominalisation agentive sont multiples et incluent, entre autres, les suffixes *-ant* (bien que son statut suffixal soit discuté, notamment par Anscombre 2003 – voir chapitre 8), *-eur*, *-ien*, *-ier*, et *-iste*, ainsi que les variantes féminines *-ante*,

1. Accessible à l'adresse `http://www.lexique.org/`.

-ienne, *-ière*, *-euse* ou *-rice*, dont une partie permet aussi la construction de noms d'instrument. Mais la construction prototypique des noms d'agent et d'instrument relève de la suffixation déverbale en *-eur* pour le masculin, et *-euse* et *-rice* pour le féminin (Huyghe et Tribout 2015).

3.1.3 La nominalisation active

La classe des noms d'action est définie comme "l'ensemble des noms qui dénotent des actions, i.e. des situations temporelles dynamiques, causant un changement" (Huyghe 2014 : 2). La nominalisation active, ou processive, en est l'un des principaux contributeurs, au travers de la suffixation et de la conversion.

Du fait de l'absence d'opération sémantique, Mel'čuk (1994) qualifie ce procédé de « dérivation syntaxique ». Le verbe et le nom d'action seraient donc maximalement proches sémantiquement, puisque ce dernier dénoterait, sous une autre forme syntaxique, la situation dynamique décrite par le verbe.

Selon Croft (1991), les verbes sont prototypiquement corrélés à la classe des actions avec un rôle de prédication, et les noms à la classe des objets avec un rôle de référence. Les noms d'action sont donc des items issus de la classe sémantique des actions mais dôtés de la fonction référentielle. Les noms d'agent sont quant à eux des items issus de la classe sémantique des objets mais réalisent une fonction de prédication. Il s'agit donc, selon Croft (1991), d'une configuration non prototypique, au regard des catégories syntaxiques, de la fonction pragmatique et de la classe sémantique.

Chomsky (1970) soutient que les nominalisations ont une structure interne semblable à celle des noms. Elles ne peuvent donc pas traduire l'aspectualité que l'on retrouve dans certains verbes et dans certaines propositions, ce qui explique qu'il n'y ait pas de nominalisations pour la proposition 3.

(3) John's having criticized the book

Si l'aspect dans son acception la plus générale, à l'image de l'aspect progressif traduit précédemment, est spécifique au verbe, un autre genre d'aspect, l'aspect lexical (ou Aktionsart), est commun aux verbes et aux noms d'actions déverbaux. Ainsi, les verbes sont généralement répartis en quatre classes, que sont les classes des achèvements, des accomplissements, des activités et des états, sur la base de la typologie dressée par Vendler (1957) et Huyghe et Marín (2007). Cette typologie est aussi appliquée aux noms d'action déverbaux, avec l'intégration d'une classe d'événement, généralement proposée pour subsumer les classes des achèvements et des accomplissements (voire celle des activités) (Balvet *et al.* 2011). L'héritage des propriétés aspectuelles du verbe n'est cependant pas systématique chez les noms déverbaux

(Huyghe et Marín 2007, Huyghe 2014), et cela peut ainsi contribuer à une différence sémantique entre le verbe et son nom d'action.

D'un point de vue morphologique, le nom d'action peut être un nom simple (Huyghe 2014) (*bal*, *conférence*, *stage*) ou construit (*ponçage*, *soulèvement*), auquel cas le nom d'action (ou processif) est généralement déverbal, à l'image du nom d'agent et du nom d'instrument. Le nom déverbal peut être construit par suffixation ou par conversion. Les suffixes disponibles pour dériver des noms d'action déverbaux sont nombreux, et incluent entre autres les suffixes *-age*, *-ion*, *-ment*, *-ance*, *-ure*, *-ade*, *-erie*, etc. Les suffixes les plus productifs sont cependant les suffixes *-age*, *-ion* et *-ment*. Notons que les noms formés par ces suffixes présentent une forte polysémie, notamment entre nom d'action, de résultat, et d'état. La sélection d'une de ces acceptions par le dérivé va ainsi aussi contribuer à la variabilité de la relation sémantique entre le nom déverbal et sa base.

Il émerge de ces caractérisations que la relation sémantique entre le verbe et ses noms déverbaux varie en fonction des familles dérivationnelles considérées, du fait de différents facteurs (morphologie, aspect, etc.). S'il ressort qu'il n'y a pas une équivalence stricte entre le verbe et le nom d'action du fait de l'opération catégorielle, les différences entre le verbe et son nom d'action semblent néanmoins moindres que celles entre le verbe et son nom d'agent. Ces tendances sont-elles perceptibles dans les modèles distributionnels ? Ce mode de représentation du sens permet-il de quantifier plus précisément ces différences de proximité ? Nous présentons dans ce qui suit les données puis l'expérience que nous mettons en place pour répondre à ces questions.

3.2 Dispositif expérimental

Pour savoir si le verbe est sémantiquent très proche de son nom d'action dérivé, et donc plus proche du nom d'action que du nom d'agent, nous nous basons sur la capacité des espaces vectoriels à rapprocher les noms sémantiquement proches en quantifiant cette proximité. Nous comparons donc la proximité sémantique entre le verbe, le nom d'agent et le nom d'instrument en comparant leur proximité distributionnelle. Nous présentons dans un premier temps la ressource à partir de laquelle nous récupérons les familles dérivationnelles (section 3.2.1), puis l'espace vectoriel sur lequel nous nous basons (section 3.2.2).

3.2.1 La ressource Lexeur

La principale ressource lexicale morphologique sur laquelle nous nous basons dans ce travail est la ressource Lexeur. Elle a initialement été constituée pour comparer en contexte la structure argumentale des noms d'agent masculins en *-eur* et celle des prédicats (verbaux ou nominaux) morphologiquement associés (Hathout et Namer 2016). Sa constitution a consisté en trois grandes étapes :

— la collecte automatique des formes déverbales (*porteur*), dénominales (*camionneur*) et complexes non construites (*auteur*, *vecteur*) en *-eur*,

— l'appariement automatique à une base,

— et la validation manuelle.

La collecte s'est basée sur trois ressources principales que sont *Frantext*, la nomenclature et l'index du *TLF*, et la nomenclature du *Robert Électronique*. Une première étape de validation manuelle a été effectuée sur les formes collectées à ce stade pour exclure les formes jugées erronées (*vadeboncoeur*, *binfaiteur*), les noms construits dénotant des qualités (*blancheur*), et pour identifier les noms non construits (*coeur*, *vecteur*). L'utilisation du programme Décor (Dal *et al.* 1999) a alors permis d'associer automatiquement à chaque forme conservée une liste ordonnée de verbes ayant la plus grande probabilité d'être associés morphologiquement (c'est-à-dire ici formellement) à la forme. Si aucun verbe attesté ne peut être associé à la forme, on cherche un nom. Si aucun nom n'est trouvé, la forme a été considérée comme non construite. Une étape de validation a ensuite permis d'ajouter les noms d'action associés aux verbes (si la base était verbale), ou les verbes rétro-construits ou construits par conversion ainsi que les noms associés (si la base était nominale). Cette étape a été réalisée à l'aide du lexique Verbaction (Hathout *et al.* 2002). La ressource a finalement bénéficié de l'ajout des formes féminines en *-euse* et *-rice* dans le cadre de la constitution de Démonette (Hathout et Namer 2014 2016).

Dans son état actuel [2], la ressource compte 5 974 sous-familles dérivationnelles. Les sous-familles sont constituées de :

— une entrée, à savoir un nom masculin en *-eur* (*abatteur*, *camionneur*, *prédateur*),

— un ou plusieurs équivalents féminins en *-euse* et/ou *-rice* (*abatteuse*, *camionneuse*, *prédatrice*),

— la base verbale ou nominale s'il y en a une (*abattre*, *camion* ; *prédateur* n'a pas de lexème de base attesté en français),

2. Lexeur est accessible à l'adresse http://redac.univ-tlse2.fr/lexiques/lexeur.html/.

— une liste de verbes morphologiquement associés (*sélectionner* pour *sélecteur*),

— une liste de noms morphologiquement associés (*rectorat* pour *recteur*) quand l'entrée est dénominale ou qu'il n'y a pas de lexème de base attesté,

— une liste de nominalisations de la base ou de nominaux associés à l'entrée (*abattage*, *prédation*).

Seul le premier champ est requis, tous les autres pouvant être vides. Tous les lexèmes sont associés à une description morphosyntaxique au format Multext-Grace. Les noms masculins sont étiquetés Ncms, les noms féminins Ncfs, et les verbes à l'infinitif Vmn----. Un extrait de Lexeur est donné dans le tableau 3.1. Pour des questions de place et de lisibilité, la description morphosyntaxique n'est pas indiquée dans le tableau.

MascAgt	FemAgt	Base	Verbes Associés	Noms Associés	Nominalisations
sculpteur	sculpteuse ; sculptrice	sculpter			sculpture ; sculptage
inflammateur	inflammatrice	enflammer			inflammation
inflammateur	inflammatrice	inflammer			inflammation
danseur	danseuse	danser			danse ; dansage danserie ; dansement
danseur	danseuse	danse	danser		
autostoppeur	autostoppeuse	autostop			
chiropracteur	chiropractrice				chiropraxie

TABLE 3.1 – Six sous-familles extraites de Lexeur

Le tableau 3.1 illustre la diversité des sous-familles incluses dans Lexeur. Toutes les familles contiennent *a minima* un nom masculin en *-eur* et au moins un nom féminin en *-euse* ou *-rice*. Certaines sous-familles présentent plusieurs équivalents féminins, à l'image de *sculpteuse* et *sculptrice*. C'est le cas de 82 familles (soit 1% de Lexeur), dont 43 ont pour verbe associé *cultiver*, à l'image du doublon formé par *apicultrice* et *apiculteuse*. Le suffixe *-euse* est globalement plus représenté que le suffixe *-rice*, à raison de 4 542 noms en *-euse*, contre 1 514 noms en *-rice* (ratio de 1 à 3).

Sur l'ensemble des 5 974 familles, 78% ont pour base un verbe (comme *sculpteur*), 14% un nom (comme *autostoppeur*) et 8% ne sont associées à aucun lexème de base attesté (*chiropracteur*). Les noms en *-eur* pour lesquels plusieurs bases peuvent être identifiées, comme *inflammateur* (de *inflammer* ou *enflammer*) et *danseur* (de *danser* ou *danse*) font l'objet de sous-familles distinctes, à raison d'une famille par base.

Notons que les noms en *-eur*, *-euse* et *-rice* de la ressource Lexeur regroupent à la fois des noms d'agent (*chanteur*), des noms d'instrument (*détonateur*) et des noms à la double lecture agentive et instrumentale (*navigateur*). La représentation de chaque type reste à quantifier, en l'absence

d'étiquettes préalables et de tests automatiques. De la même façon, les noms présents dans le champs 'Nominalisations' de Lexeur couvrent des réalités multiples et regroupent entre autres des noms d'action (*identification*), de résultat (*construction*), d'objet (*goûter*), d'état (*abattement*), et de lieu (*fumerie*). Là encore, une quantification précise des différents types reste à faire.

Les champs facultatifs 'Verbes associés' et 'Noms associés' (absents du tableau 3.1) s'avèrent globalement peu peuplés. Le champ 'Verbes associés' est ainsi renseigné pour 660 familles, qui se voient alors attribuer un ou plusieurs verbes. C'est notamment le cas de la famille associée à l'entrée *accenseur*, pour laquelle la base identifiée est le nom féminin *acense* et qui se voit attribuer les verbes *accenser* et *acenser*. De même, l'entrée *coadjurateur*, qui n'a pas de base identifiée dans Lexeur, est associée au verbe *aider*. Le champ 'Noms associés' est quant à lui renseigné pour 162 familles. C'est notamment le cas de *cardioaccélérateur*, qui n'a pas de base identifiée dans Lexeur mais qui est associé au verbe *accélérer* et aux noms *accélération*, *accélèrement* et *accélérage*, ou de *aventurieur*, qui n'a pas de base ni de verbe associé, mais qui a le nom associé *aventurier*.

Lexeur offre une grande couverture tant pour les noms d'agent et d'instrument que pour les noms d'action, et est le fruit d'une constitution manuelle par des linguistes. Cela constitue donc une bonne base pour l'étude des noms d'agent et d'action déverbaux.

3.2.2 Espace vectoriel

Nous estimons la proximité des nominalisations déverbales avec leur base à l'aide de la sémantique distributionnelle car c'est un des rares cadres qui fournit une caractérisation sémantique automatique et systématique. Dans cette approche, la quantification repose sur la représentation des mots d'un corpus dans un espace vectoriel (voir chapitre 1).

Comme indiqué dans la section 1.2, nous faisons le choix de construire notre modèle [3] à l'aide de Word2Vec (Mikolov *et al.* 2013a). Nous choisissons pour cela d'utiliser l'architecture CBOW et l'algorithme *Negative Sampling*. Nous fixons le nombre de dimensions des vecteurs à 100, la taille de la fenêtre à 5, et le seuil de fréquence minimum à 5. Les autres paramètres sont laissés par défaut, et notamment le sous-échantillonnage (10^{-3}).

L'entraînement d'un modèle distributionnel nécessite l'utilisation de corpus conséquents. Nous faisons le choix de travailler à partir du corpus *Wiki-*

3. L'ensemble des modèles distributionnels utilisés dans ce travail ont été construits sur la plateforme Osirim, administrée par l'IRIT et soutenue par le fonds européen de développement régional (FEDER), le gouvernement français, la région Midi-Pyrénées et le Centre National de la Recherche Scientifique (CNRS).

pedia2018, constitué à partir du *dump* (daté de 2018) de la version française de l'encyclopédie en ligne du même nom, pour un total d'un milliard de mots [4]. La taille de ce corpus est le premier critère de choix. Le corpus *Wikipedia2018* est l'un des plus grands corpus français actuellement disponibles, avec le corpus *frWaC* (1.3 milliard de mots), et le corpus *frCoW* (9 milliards de mots). Le corpus *frWaC* n'a pas été choisi du fait de la qualité insuffisante des données liée à sa constitution. Le corpus *frCoW* n'était quant à lui pas encore disponible au commencement de ce travail, et il n'a pas été utilisé par la suite pour garder des résultats cohérents et comparables. Le choix du corpus *Wikipedia2018* se justifie aussi par sa nature. Issu de l'encyclopédie collaborative en ligne, il couvre un grand nombre de domaines et de sujets, et présente donc un vocabulaire très varié. Fruit du travail collaboratif d'internautes, il constitue un échantillon du français contemporain. Les contributions diverses, entre ajouts et corrections, garantissent la présence d'usages généralisés, et limite la présence trop marquée d'idiolectes.

Notons en passant que si nous qualifions *Wikipedia2018* de corpus, cette description est parfois discutée, notamment au regard des standards en linguistique de corpus. En effet, il ne s'agit pas d'un corpus dans le sens prototypique (Gilquin et Gries 2009), dans la mesure où il n'a pas été constitué en suivant les critères préconisés en linguistique de corpus relatifs à sa représentativité (Biber 1993). Nous y référerons néanmoins dans la suite de ce travail par le terme « corpus »dans la mesure où il s'agit d'un ensemble de textes que l'on fournit à l'algorithme pour produire un espace vectoriel.

Wikipedia2018 est préalablement lemmatisé pour permettre à Word2Vec de capturer davantage de régularités sémantiques, limitant la dispersion des contextes. La lemmatisation est effectuée à l'aide du parseur syntaxique Talismane (Urieli 2013). Lorsque le lemme d'une forme n'est pas identifié, la forme est conservée. Word2Vec construit donc un vecteur par lemme dont la fréquence est supérieure ou égale à 5.

3.3 Scores de proximité

Nous évaluons la proximité entre trois membres des familles dérivationnelles partielles contenues dans Lexeur, à savoir le verbe (désormais *Vb*), ses noms d'agent et d'instrument (*Nag*), et ses noms d'action dérivés (*Nac*). Nous constituons dans un premier temps des triplets {*Vb, Nag, Nac*} à partir de Lexeur, puis nous calculons la proximité distributionnelle entre chaque paire de mots (section 3.3.1). Nous évaluons dans un second temps notre hypothèse au regard de ces proximités (section 3.3.2).

4. Le corpus a été construit par Franck Sajous.

3.3.1 Extraction et annotation des triplets

Nous extrayons de Lexeur tous les triplets {*Vb, Nag, Nac*}, où pour une famille donnée, *Vb* est extrait du champ 'Base', *Nag* du champ 'MascAgt' ou 'FemAgt', et *Nac* du champ 'Nominalisation', soit un total de 13 739 triplets distincts. Du fait des caractéristiques de l'espace vectoriel présenté dans la section 3.2.2, seuls les triplets dont les trois membres apparaissent au moins cinq fois dans le corpus sont conservés, soit 2 585 triplets. Les noms d'agent ou d'instrument de ces triplets sont majoritairement des noms en *-eur* (2 068), 437 noms en *-euse* et 80 noms en *-rice*. Les noms d'action se répartissent plus ou moins équitablement entre noms en *-age* (640), en *-ion* (340), en *-ment* (762) et 839 noms construits par d'autres procédés (catégorie 'autre').

Parmi les 839 noms d'action qui ne sont pas suffixés en *-age*, *-ion* ou *-ment* se trouvent des noms suffixés en *-ure* (*argenture*, *lecture*, *rupture*), en *-ance/-ence* (*naissance*, *maintenance*, *régénérescence*), en *-erie* (*batellerie*, *beuverie*, *broderie*), des conversions (*montée*, *ralenti*, *vente*), des emprunts (*jogging*), etc.

Un aperçu du type de triplets ainsi obtenus est donné ci-dessous en (4), (5), (6) et (7).

(4) a. *broyer - broyeur - broiement*

b. *broyer - broyeur - broyage*

(5) a. *sculpter - sculpteur - sculpture*

b. *sculpter - sculpteuse - sculpture*

(6) a. *violer - violeur - viol*

b. *violer - violeur - violence*

c. *violer - violeuse - viol*

d. *violer - violeuse - violence*

(7) a. *vulgariser - vulgarisateur - vulgarisation*

Ces triplets exhibent des configurations diverses. Ainsi, certains triplets contiennent le même verbe et le même nom d'agent ou d'instrument, mais un nom d'action différent (à l'image de (4a) et (4b)). D'autres triplets proposent un même nom d'action et un même verbe, mais un nom d'agent ou d'instrument différent, comme illustré par (5). Ces cas de figure correspondent généralement à l'alternance entre le nom d'agent masculin et féminin (à l'image de (5a) et (5b)), mais intègrent aussi parfois l'alternance entre agent et instrument, à l'image de *sertisseur* et *sertisseuse*. Certains triplets combinent ces deux types de variation, à l'image de (6). Enfin, certaines familles ne

fournissent qu'un seul triplet, à l'image de (7), soit parce que la famille dérivationnelle ne contient pas d'autres dérivés, soit parce que les autres dérivés ont moins de cinq occurrences en corpus.

Nous calculons la proximité P de chaque couple de lexèmes composant un triplet dans le modèle distributionnel. Nous obtenons ainsi trois scores : celui entre le verbe et le nom d'agent ou d'instrument P(VbAg), celui entre le verbe et le nom d'action P(VbAc) et celui entre le nom d'agent ou d'instrument et le nom d'action P(AgAc). Un exemple de cette annotation est donné dans le tableau 3.2.

Verbe	Nom d'agent	Nom d'action	P(VbAg)	P(VbAc)	P(AgAc)
exécuter	*exécuteur*	*exécution*	0.373	**0.566**	0.375
itérer	*itérateur*	*itération*	0.503	**0.704**	0.343
détenir	*détenteur*	*détention*	**0.704**	0.160	0.106
réguler	*régulateur*	*régulation*	0.673	0.687	**0.754**

TABLE 3.2 – Scores de proximité pour quatre triplets

Le tableau 3.2 montre les différentes configurations possibles. La première configuration (illustrée par *exécuter* et *itérer*) correspond au cas où le verbe est effectivement plus proche du nom d'action que du nom d'agent, c'est-à-dire où le score P(VbAc) est supérieur aux deux autres scores. Cette configuration valide notre hypothèse. Le tableau présente aussi des cas où c'est le nom d'agent qui est le dérivé le plus proche du verbe (*détenir*), c'est-à-dire où P(VbAg) est supérieur aux deux autres scores. Le triplet de la 4^{e} ligne (*réguler*) montre le cas où les noms d'agent ou d'instrument et d'action sont plus proches entre eux qu'ils ne le sont du verbe, c'est-à-dire où P(AgAc) est supérieur aux deux autres scores. Notons que pour tous ces triplets, le second score de proximité le plus élevé peut varier, comme le montrent les triplets construits autour de *exécuter* et *itérer*. On observe pour le premier que le second score le plus élevé est P(AgAc), alors que pour le second, il s'agit du score P(VbAg).

3.3.2 Analyse des scores de proximité

Si l'on s'éloigne des triplets pris individuellement et que l'on regarde à l'échelle de l'ensemble des 2 585 triplets, on constate que le verbe a tendance à être plus proche du nom d'action que du nom d'agent ou d'instrument. En effet, 59% des triplets (1 520 sur 2 585) ont comme score de proximité le plus élevé celui entre le verbe et le nom d'action. Le score de proximité P(AgAc) est le plus élevé dans 24% des cas (619 triplets). Enfin, pour 17% des triplets

(soit 446 sur 2 585), le verbe et le nom d'agent ou d'instrument sont les items les plus proches. Dans près de 60% des cas, donc, le verbe est plus proche du nom d'action que du nom d'agent, et ces deux mots sont plus proches entre eux que ne le sont les autres items du triplet. Cela va donc dans le sens de l'hypothèse présentée plus haut.

Ce constat est confirmé par le fait que le score moyen P(VbAc) est de 0.400, alors que le score moyen P(VbAg) est de 0.253. Le score moyen de proximité entre les deux dérivés P(AgAc) est quant à lui de 0.289. Notons que les différences observées entre ces trois scores de proximité sont significatives ($p<2.2e\text{-}16$ au test de Student apparié). La comparaison des trois scores moyens de proximité montre donc qu'en moyenne, au sein d'une famille dérivationnelle, les membres les plus proches sont le verbe et le nom d'action. Il est intéressant à ce titre de constater que la proximité entre le verbe et le nom d'agent ou d'instrument d'une part, et celle entre le nom d'agent ou d'instrument et le nom d'action d'autre part sont du même ordre (0.253 vs 0.289). Cette intuition est corroborée par les scores de corrélation entre les différents scores de proximité : la corrélation de Pearson indique une corrélation positive élevé (0.6) entre P(VbAg) et P(AgAc), alors que cette corrélation, bien que positive, est moins élevée entre P(VbAg) et P(VbAc) et entre P(VbAc) et P(AgAc) (0.3 dans les deux cas). Cela semble conforter l'idée d'une forte similarité entre le verbe et le nom d'action, dans la mesure où le nom d'agent ou d'instrument tend à se trouver à la même distance des deux autres lexèmes.

Nous présentons plus en détail la distribution de chaque score de proximité pour les 2 585 triplets analysés dans la figure 3.1.

La figure 3.1 montre que la dispersion est plus grande pour le score P(VbAc) que pour les autres scores, bien que l'écart-type soit similaire pour les trois scores (0,2 pour P(VbAc) contre 0.17 et 0.19 respectivement pour P(VbAg) et P(AgAc)). Le score de proximité entre le verbe et le nom d'action est donc plus variable que le score de proximité entre les autres membres des triplets. On constate cependant que près de 75% des triplets ont un score P(VbAc) plus élevé que le score moyen de proximité entre les autres membres du triplet. La figure 3.1 fait par ailleurs apparaître des cas où le score de proximité entre le verbe et le nom d'action est égal à 1. Nous analysons plus en détail ces triplets dans le chapitre 4, mais cela semble indiquer que dans certains triplets, le verbe et le nom d'action sont effectivement très similaires puisque leurs vecteurs sont identiques. Cela contribue potentiellement au score moyen de proximité plus élevé affiché par la paire Verbe-Nom d'action.

La proximité affichée par les différents membres des triplets varie cependant en fonction des procédés impliqués. Nous détaillons dans les tableaux 3.3 et 3.4 les scores moyens de proximité obtenus entre les différents lexèmes

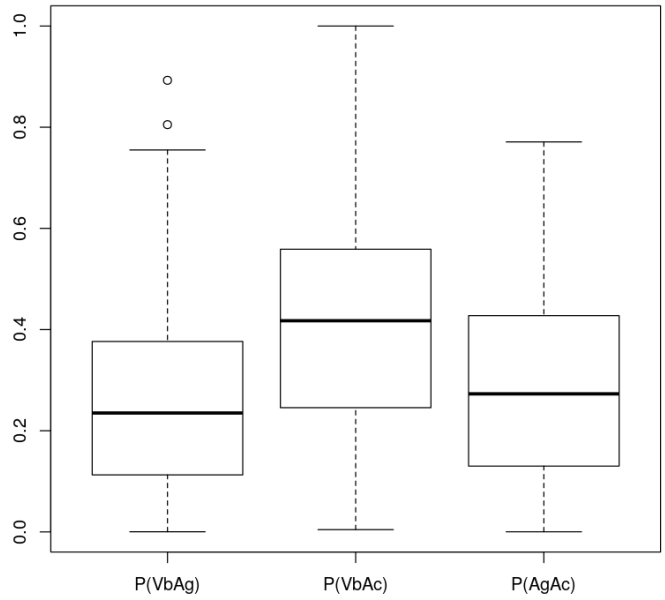

FIGURE 3.1 – Dispersion des scores de proximité des 2 585 triplets

en fonction de leur suffixe. Sont uniquement donnés dans le tableau 3.4 les résultats des noms d'action qui nous intéressent principalement dans cette thèse, à savoir les déverbaux suffixés en *-age*, *-ion* et *-ment*.

	P(VbAg)	P(VbAc)	P(AgAc)
-eur	0.267	**0.402**	0.307
-euse	0.190	**0.377**	0.207
-rice	0.228	**0.473**	0.244

TABLE 3.3 – Scores de proximité moyens en fonction du suffixe agentif

Le tableau 3.3 confirme que la tendance générale d'une plus forte proximité entre le verbe et le nom d'action se vérifie pour les trois suffixes *-eur*, *-euse* et *-rice*. Nous en tirons cependant trois observations supplémentaires. Tout d'abord, nous remarquons que les scores de proximité moyens sont globalement plus faibles pour les triplets impliquant un nom d'agent en *-euse*, et dans une moindre mesure en *-rice*. Leurs membres sont globalement plus éloignés distributionnellement les uns des autres. Les triplets semblent moins homogènes sémantiquement. Deuxièmement, on observe que les triplets im-

pliquant *-rice* montrent le score moyen P(VbAc) le plus élevé, ce qui signifie que c'est dans ces triplets que le verbe est en moyenne le plus proche du nom d'action. Enfin, on remarque que la proximité entre les noms d'agent et les verbes est plus importante pour les noms suffixés en *-eur* que pour les autres suffixes. Les noms d'agent ou d'instrument en *-eur* sont donc en moyenne plus proches du verbe que ne le sont les noms en *-euse* ou *-rice*. Une explication quant à ces différences pourrait résider dans la plus grande neutralité du suffixe masculin, tout particulièrement vis-à-vis du suffixe *-euse*, porteur d'une forte connotation dépréciative (que nous explorons dans la partie III).

	P(VbAg)	P(VbAc)	P(AgAc)
-age	0.245	**0.367**	0.299
-ion	0.276	**0.458**	0.319
-ment	0.225	**0.359**	0.228
autre	0.248	**0.389**	0.277

TABLE 3.4 – Scores de proximité moyens en fonction du suffixe de nominalisation

Le tableau 3.4 montre la variation des scores de proximité en fonction du suffixe des noms d'action. On constate que les noms en *-ion* semblent en moyenne plus proches de leur verbe que ne le sont les noms en *-age* ou en *-ment*. Cette plus forte proximité s'observe aussi avec le nom d'agent ou d'instrument, que ce soit vis-à-vis du verbe ou du nom d'action. Les familles dont les dérivés se construisent sur une base savante semblent donc plus cohésives. Cela est cohérent avec la plus forte proximité observée dans le cadre des triplets impliquant un nom d'agent ou d'instrument en *-rice* dans le tableau 3.3. Les triplets impliquant un nom d'action en *-ment* affichent quant à eux la plus grande hétérogénéité, les scores étant globalement les plus faibles.

Notons que le score moyen de proximité P(VbAg) entre le nom d'agent et le verbe n'est jamais plus élevé que le score moyen de proximité P(VbAc), quelles que soient les configurations suffixales des noms d'agent ou d'instrument et des noms d'action. On remarque que la proximité entre le nom d'action et le verbe est la plus élevée dans le cas des triplets impliquant un nom d'agent en *-eur* et un nom d'action en *-ion*, avec un score P(VbAc) moyen de 0.462 (sur un total de 657 triplets), et les triplets impliquant des noms en *-rice* et *-ion*, avec un score moyen de 0,485 (pour 68 cas). À l'inverse, cette proximité est minimale dans le cas des triplets impliquant un nom d'agent en *-rice* et un nom d'action en *-age* (soit trois triplets), avec un score moyen de 0.225. Le score P(VbAg) est quant à lui toujours le score le plus faible, sauf dans le cas des triplets impliquant un nom d'agent en *-euse*

et un nom d'action en *-ment* (soit 59 triplets), où le score P(VbAg) moyen est de 0.152 et le score P(AgAc) moyen de 0.150. Cela corrobore la plus grande hétérogénéité observée pour ces deux suffixations.

Un examen plus précis des triplets suggère que l'on retrouve une proportion importante de noms d'instrument parmi les triplets exhibant une plus forte proximité P(VbAc), à l'image de *inhalateur*, *mitrailleuse*, *cuiseur*, *ioniseur*, etc. Pour tenter de quantifier cette observation, nous sous-échantillonnons les triplets pour ne conserver que des triplets impliquant un nom d'agent et un nom d'instrument. Pour chaque catégorie, nous sélectionnons 30 noms monosémiques (ou du moins monotypé, c'est-à-dire dont les acceptions relèvent toutes de la même catégorie – soit agentive, soit instrumentale). La monosémie est vérifiée à l'aide du *TLFi* et de *Wiktionnaire*. Nous sélectionnons alors les triplets où apparaissent ces noms. Nous obtenons ainsi 41 triplets incluant un nom d'agent, et 36 triplets incluant un nom d'instrument. Nous comparons les scores de proximité affichés par ces deux groupes de noms, présentés dans le tableau 3.5.

	P(VbAg)	P(VbAc)	P(AgAc)
All	0.362	0.493	0.398
Triplets agentifs	0.355	0.442	0.358
Triplets instrumentaux	0.369	0.550	0.443

TABLE 3.5 – Scores de proximité moyens en fonction de la nature agentive ou instrumentale des triplets

Le tableau 3.5 montre que les scores moyens de proximité sont globalement plus élevés pour les triplets de notre échantillon que pour l'ensemble des 2 585 triplets, et plus spécifiquement pour les triplets impliquant un nom d'instrument. On constate par ailleurs que cette augmentation est particulièrement marquée pour les scores P(VbAc) et P(AgAc). Le verbe tend donc à être plus proche du nom d'action dans le cas de triplets impliquant un nom d'instrument que de triplets avec un nom d'agent. Par ailleurs, on peut voir que le verbe tend à être à la même distance du nom d'agent que du nom d'instrument. On pourrait ainsi faire l'hypothèse que le trait humain différenciant les noms d'agent et d'instrument ne fait pas varier sensiblement la proximité au verbe. Le corollaire de cette hypothèse serait que l'explication de la différence observée entre les deux groupes de triplets résiderait dans le contenu sémantique de la base et dans l'action dénotée elle-même. On peut faire l'hypothèse que les verbes et noms dénotant des actions réalisées par des instruments se sont moins lexicalisés que ceux dénotant des actions réalisées par des agents. Le tableau 3.5 présente une tendance évaluée sur un nombre

limité de triplets, qui ne sont pas forcément représentatifs du comportement de l'ensemble des 2 585 triplets (comme le montrent les valeurs plus élevées des trois scores, bien que suivant les mêmes tendances). Il faudrait analyser de manière plus systématique ces différences.

Notons que le nombre de triplets ayant comme score de proximité le plus élevé P(VbAc) ne varie pas significativement entre les groupes agentifs et instrumentaux ($p = 0.582784$ au test du χ^2), à savoir que 25 des triplets ont P(VbAc) comme score le plus élevé dans les deux groupes, et respectivement 8 et 4 pour P(VbAG), et 8 et 7 pour P(AgAc). L'hypothèse d'une plus grande proximité entre le verbe et le nom d'action est donc validée dans les mêmes proportions entre les deux groupes, bien que le degré de proximité varie.

Chapitre 4

Explorations méthodologiques

Nous avons montré dans le chapitre précédent que l'hypothèse d'une plus grande proximité sémantique entre les verbes et leurs noms d'action au sein des familles dérivationnelles était validée dans les espaces vectoriels, par la comparaison à grande échelle des scores de proximité distributionnelle de verbes avec leur nom d'agent et leur nom d'action dérivé. Nous explorons dans ce qui suit plus finement les résultats, afin d'en tirer notamment des enseignements méthodologiques. Pour ce faire, nous explorons les différentes tendances que l'on observe en termes de proximité, afin de faire émerger les phénomènes à l'œuvre. Nous prenons comme point d'entrée la comparaison des scores P(VbAg) et P(VbAc), puisque ce sont les proximités qui nous intéressent le plus au sein des triplets, sans pour autant écarter l'examen de leur proximité au nom d'action.

Nous étudions plusieurs cas de figure, à savoir les cas où l'on observe une grande différence entre les deux scores de proximité (section 4.1), et au contraire, les cas où la différence entre les deux scores est minime (section 4.2). Bien que les scores de proximité soient difficiles à analyser dans l'absolu, nous fixons pour les besoins de cette analyse, et de façon arbitraire, des seuils pour juger de l'importance de différences entre les scores. Nous considérons que l'écart est minime dès lors qu'il est inférieur à 0,1, et que l'écart est important dès lorsqu'il dépasse 0,3. Notons que l'écart entre les scores de proximité est un point d'entrée pour procéder à une analyse plus qualitative des résultats. Il ne s'agit pas tant d'expliquer les raisons des rapprochements dans l'espace vectoriel que de comprendre un peu mieux les possibilités et les limites d'une telle approche. Nous montrons à ce titre que l'invalidation de l'hypothèse s'explique pour certaines familles par des phénomènes linguistiques comme la lexicalisation et la polysémie. L'examen de différents types de triplets montre ainsi les avantages mais aussi les limites de l'utilisation des espaces vectoriels en fonction du niveau d'analyse choisi.

4.1 Écart important entre les scores P(VbAc) et P(VbAg)

Nous examinons les triplets pour lesquels l'écart entre les scores P(VbAc) et P(VbAg) est important, dans un sens comme dans l'autre. Nous nous concentrons sur ces noms car cela signifie soit que ces triplets valident de totalement l'hypothèse, soit au contraire que l'hypothèse est fortement invalidée par ces triplets. Nous détaillons ces deux cas de figures dans les sections 4.1.1 et 4.1.2 respectivement.

4.1.1 Le verbe est plus proche du nom d'action

Nous examinons ici les triplets où le verbe est plus proche du nom d'action que du nom d'agent ou d'instrument, c'est-à-dire pour lesquels P(VbAc) est supérieur à P(VbAg). Il s'avère que la majorité de ces triplets valident l'hypothèse d'une plus grande proximité entre le verbe et le nom d'action qu'entre les autres membres du triplet, mais ce n'est pas systématique, à l'image de triplets comme {*centrer*, *centreur*, *centrage*} pour lequel P(VbAc) (0.405), bien que plus élevé que P(VbAg) (0.091), est néanmoins plus faible que P(AgAc) (0.455). Nous essayons de comprendre ce qui justifie pour ces triplets un tel écart, plus que la validation de l'hypothèse elle-même, et ce au travers de trois grands types de phénomène : (i) l'homonymie et la polysémie, (ii) la lexicalisation et le glissement sémantique, et (iii) les usages.

L'impact de l'homonymie et de la polysémie est instancié par un cas des plus flagrants, mais aussi très particulier, qu'est celui des triplets tels qu'illustrés dans le tableau 4.1. Ces triplets se caractérisent par l'identité formelle de deux de leurs trois membres, à savoir le verbe et le nom d'action, qui se révèlent homographes.

Verbe	Nom d'agent	Nom d'action	P(VbAg)	P(VbAc)	P(AgAc)
goûter	*goûteur*	*goûter*	0.465	1	0.465
porter	*porteur*	*porter*	0.523	1	0.523
toucher	*toucheur*	*toucher*	0.166	1	0.166

TABLE 4.1 – Scores de proximité des triplets construits autour de *goûter*, *porter* et *toucher*

Dans ces triplets, les items identifiés comme verbe et nom d'action ont la même forme. De fait, les deux lexèmes partagent la même représentation vectorielle, qui explique donc le score de proximité P(VbAc) de 1. Dans cette

situation, il est impossible d'évaluer la proximité réelle du verbe et du nom. Cette ambiguïté concerne ainsi un total 16 formes dans nos données : *manger*, *dîner*, *souper*, *goûter*, *dire*, *parler*, *lever*, *grimper*, *vivre*, *lâcher*, *lancer*, *porter*, *baiser*, *rire*, *tomber*, et *coucher*. Une façon de résoudre ce problème inhérent aux modèles distributionnels, s'il s'agissait de conserver ces données, serait d'entraîner ce dernier sur un corpus lemmatisé et étiqueté morphosyntaxiquement, ce qui permettrait de calculer des représentations vectorielles distinctes pour les formes verbales et les formes nominales. Nous pourrions alors cibler plus précisément le vecteur du verbe et le vecteur du nom d'action pour faire nos comparaisons. Cela donnerait une importance très grande à l'annotation automatique des catégories grammaticales, et aux erreurs que cela peut induire, et rendrait les calculs – liés au prétraitement du corpus et à l'entraînement du modèle – plus lourds. Cela entraînerait aussi une perte liée à la plus grande dispersion des contextes.

Le second cas de figure que nous identifions est celui de la lexicalisation ou du glissement sémantique d'au moins un des membres du triplets. Nous l'illustrons à l'aide du triplet construit autour du verbe *chauffer* et impliquant les noms *chauffeuse* et *chauffage* (P(VbAg)=0.064 ; P(VbAc)=0.710 ; P(AgAc)=0.146). Si *chauffeuse* est initialement associé à l'idée de chauffage, puisqu'elle désigne la 'chaise pour se chauffer près du feu' (définition issue du *TLFi*), le lien avec la notion de chaleur s'est perdu, l'éloignant de fait du verbe et du nom d'action. Le nom *chauffeuse* est par ailleurs polysémique et peut désigner l'agent féminin qui conduit, équivalent de *chauffeur*, qui l'éloigne d'autant plus du paradigme sémantique formé par *chauffer* et *chauffage*. De fait, cet éloignement du nom *chauffeuse* au sein d'un triplet se retrouve aussi dans les deux autres triplets impliquant le nom *chauffeuse* et le verbe *chauffer*, complétés respectivement des noms d'action *chauffe* et *chaufferie*.

Un troisième type de cas où l'écart entre P(VbAc) et P(VbAg) est très marqué concerne les triplets pour lesquels les usages en corpus du nom d'agent ne permettent pas de le rapprocher du verbe et du nom d'action, malgré l'existence d'un lien sémantique avéré. C'est notamment le cas du nom *franchisseur*, dans le triplet contenant le verbe *franchir* et le nom d'action *franchissement* (P(VbAg)=0.019 ; P(VbAc)=0.707 ; P(AgAc)=0.062). Ainsi les contextes de *franchisseur*, qui désigne un véhicule, ne sont pas suffisamment informatifs relativement au sémantisme du triplet, comme l'illustre l'exemple donné en (8), extrait de l'article *Toyota Land Cruiser* du corpus *Wikipedia2018*.

(8) *Toujours très bon* ***franchisseur****, ce VDJ200 (appellation Toyota) peut désormais côtoyer les SUV « Grand Luxe » avec des performances sur*

route, comme un 0 à 100 km/h abattu en 8,2 secondes et une vitesse maximale de plus de 210 km/h.

Le sens de *franchisseur* est clairement construit sur celui de *franchir* et de *franchissement*, comme le prouve la présence du nom d'action *franchissement* plus en amont dans l'article, mais ne semble pas traduit par la distribution du nom.

Notons que la lexicalisation et la pauvreté des contextes illustrées par les triplets présentés précédemment ne concernent pas directement le verbe et le nom d'action, et n'ont donc pas d'influence directe sur leur proximité. L'éloignement du nom d'agent ou d'instrument que ces phénomènes induisent accentue cependant l'écart entre le verbe et le nom d'agent ou d'instrument, et donc possiblement permet à un triplet de valider l'hypothèse, quand bien même la proximité entre le verbe et le nom d'action est faible.

4.1.2 Le verbe est plus proche du nom d'agent ou d'instrument

Nous analysons ici les triplets qui invalident de façon significative l'hypothèse d'une plus grande proximité entre le verbe et le nom d'action. Là encore, il est possible que le score le plus élevé au sein des triplets considérés soit le score P(AgAc) entre le nom d'action et le nom d'agent, mais dans tous les cas, ces triplets ne valident pas l'hypothèse initiale. Plusieurs types de phénomènes peuvent expliquer ce résultat, qui sont de même nature que ceux présentés dans la section 4.1.1 : la polysémie et l'homonymie, la lexicalisation et le glissement sémantique et les usages en corpus.

Contrairement à la situation précédente, la polysémie et l'homonymie ont un impact important sur les proximités au sein des triplets, dans la mesure où elles vont réduire la proximité entre le verbe et le nom d'action, au profit du nom d'agent. Cela s'observe à des degrés variables, que nous illustrons par les quatre triplets du tableau 4.2.

Verbe	Nom d'agent	Nom d'action	P(VbAg)	P(VbAc)	P(AgAc)
détenir	*détenteur*	*détention*	0.704	0.160	0.106
porter	*porteur*	*port*	0.523	0.040	0.159
sprinter	*sprinteur*	*sprint*	0.893	0.697	0.616
essorer	*essoreuse*	*essor*	0.544	0.058	0.069

TABLE 4.2 – Scores de proximité des triplets construits autour des verbes *détenir*, *porter*, *sprinter* et *essorer*

Le degré minimal dans cette configuration correspond au triplet construit autour de *détenir*, dans lequel le nom *détention* s'avère polysémique entre l'action de détenir et l'état d'être détenu. Cette seconde acception contribue à la distanciation du verbe et du nom d'action dans la mesure où ce dernier ne réfère pas directement à l'action dénotée par le verbe, ce qui dilue le lien entre *détenir* et *détention* au sein de la représentation vectorielle. À l'inverse, le sens de *détenteur* est directement lié à l'action de *détenir*. Dans le cas du triplet construit autour de *porter*, l'homonymie porte sur le nom *port*, qui désigne l'action de porter et l'installation maritime, sans rapport aucun avec l'action de porter. Là encore, l'information sémantique du prédicat est diluée dans la représentation vectorielle, ce qui se traduit par un éloignement du verbe et du nom d'action au profit du nom d'agent. Dans ces deux cas, seule une annotation manuelle en corpus préalable à l'entraînement du modèle permettrait de distinguer les deux sens et de proposer des vecteurs distincts. Une telle annotation est cependant inenvisageable, du fait de son coût.

Le cas de *sprinter* est le cas extrême de cette configuration, comme pendant des cas illustrés dans le tableau 4.1 avec les verbes *goûter*, *toucher* ou *porter*. Ici aussi le problème réside dans l'identité formelle de deux lexèmes, impliquant ici le verbe et une variante du nom d'agent, absente du triplet, mais présente en corpus. La forme *sprinter* est en effet un anglicisme utilisé pour désigner l'action mais aussi (et surtout) le sportif qui pratique cette spécialité. La représentation vectorielle de la forme *sprinter* est donc fortement influencée par son usage agentif, ce qui se traduit par une très forte proximité du nom d'agent et du verbe, au détriment du nom d'action (même si la proximité avec celui-ci reste relativement élevée). Là encore, un éventuel étiquetage du corpus pourrait pallier cette ambiguïté.

Le tableau 4.2 illustre un dernier cas de figure relatif à l'homonymie et à la polysémie. Dans le triplet associé au nom *essoreuse*, on constate qu'aucune des acceptions du nom d'action polysémique *essor*, qui désigne un développement et l'envol d'un oiseau, n'est lié sémantiquement à *essoreuse*. On pourrait penser dans un premier temps à une erreur liée à la constitution de la ressource, induite par la similarité formelle des noms, mais leur association au sein d'un même triplet s'explique par la polysémie du verbe de base, *essorer*, qui désigne l'action de faire sécher quelque chose et l'action de prendre son envol pour un oiseau. De fait, le verbe *essorer* se trouve à l'interface de deux paradigmes sémantiques distincts auxquels appartiennent respectivement les noms *essoreuse* et *essor*. L'emploi de *essorer* dans son acception ornithologique étant plus marginal dans notre corpus, le verbe est de fait rapproché du nom d'instrument, qui relève du paradigme sémantique mobilisé par *essorer* dans le corpus.

Le second type de phénomènes à l'œuvre dans l'éloignement important

du verbe et du nom d'action est la lexicalisation. Nous l'illustrons par les deux triplets présentés dans le tableau

Verbe	Nom d'agent	Nom d'action	P(VbAg)	P(VbAc)	P(AgAc)
chanter	*chanteur*	*chantage*	0.533	0.033	0.081
violer	*violeur*	*violement*	0.496	0.074	0.381

TABLE 4.3 – Scores de proximité des triplets construits autour des verbes *chanter* et *violer*

Dans le premier exemple du tableau 4.3, *chantage* est lexicalisé. S'il y a un lien morphologique entre *chantage* et *chanter*, traduit dans le cadre de la paraphrase 'faire chanter', ce lien subsiste que dans le cadre de cette expression figée paraphrastique, qui de fait voit ses composants désémantisés (Rastier 2008). Il n'y a donc plus de lien sémantique entre *chantage* et *chanter*, ni entre *chantage* et *chanteur* par ailleurs, ce qui explique leur grande distance dans l'espace vectoriel.

Le second cas de lexicalisation illustré par le tableau 4.3, est celui du triplet construit autour du verbe *violer*. Le glissement sémantique concerne ici deux des trois membres du triplet, à savoir le nom d'agent *violeur* et le verbe *violer*, dont les sens se sont spécialisés sur la violence sexuelle imposée à autrui. À l'inverse, *violement* – bien que désuet – a gardé son sens plus général d'atteinte à quelque chose ou quelqu'un. La spécialisation du nom d'agent et du verbe se traduit de fait par une plus grande proximité de ces deux items, au détriment du nom d'action [1]. Le rapport entre les différents scores de proximité change ainsi fortement dès lors que le nom d'action *violement* est remplacé par le nom d'action *viol*, qui du fait de sa spécialisation est particulièrement proche du nom d'agent, au profit du score P(AgAc) (P(VbAg)=0.496 ; P(VbAc)=0.513 ; P(AgAc)=0.714).

Enfin, le dernier phénomène que nous retrouvons via l'examen de triplets affichant de forts écarts en termes de score de proximité est relatif aux usages en corpus. Ainsi, des noms d'action comme *ramassement*, dans le triplet qu'il forme avec le nom d'agent *ramasseur* et le verbe *ramasser*, *menterie*, dans le triplet formé avec *menteur* et *mentir*, ou *mangement* dans le triplet formé avec *mangeur* et *manger*, se caractérisent pas des usages marginaux, soit du fait de leur fréquence – *mangement* n'apparaissant que 8 fois dans le corpus *Wikipedia2018* –, soit du fait de leur caractère vieilli (d'après le *TLFi*) ou spécifique. Le nom *ramassement* est ainsi particulièrement associé au champ

1. Un examen en corpus de *violement* suggère qu'une partie des occurrences correspondent à une forme erronée de l'adverbe *violemment*, ce qui contribue d'autant plus à l'éloignement sémantique de *violement* vis-à-vis de *violeur* et *violer*.

lexical de la lutte, en tant que nom de prise, et *menterie* correspond à un régionalisme canadien.

4.2 Écart minime entre les scores P(VbAc) et P(VbAg)

Un écart minime entre les scores de proximité P(VbAc) et P(VbAg) signifie que le verbe est dans une certaine mesure aussi proche du nom d'agent que du nom d'action dans l'espace vectoriel. Plusieurs cas de figures permettent d'expliquer une telle similarité que nous explorons à l'aune de la proximité du nom d'agent et du nom d'action. En effet, on distingue les cas où l'ensemble des membres du triplet sont relativement équidistants (section 4.2.1) des cas où le nom d'action se distingue par sa proximité particulière au nom d'agent et au verbe, qu'il soit distant ou au contraire proche (section 4.2.2). Signalons qu'un écart minime entre P(VbAc) et P(VbAg) ne dit rien de la proximité respective entre le verbe et le nom d'action, et le verbe et le nom d'agent. Certains triplets se caractériseront ainsi par des scores P(VbAc) et P(AgVb) proches mais faibles – à l'image du triplet {*défiler, défileur, défilement*}, avec des scores P(VbAg), P(VbAc) et P(AgAc) respectifs de 0.192, 0.192 et 0.076 – ou au contraire élevés – à l'image du triplet {*prospecter*, *prospecteur*, *prospection*}, avec des scores P(VbAg), P(VbAc) et P(AgAc) respectifs de 0.598, 0.605 et 0.430.

4.2.1 Homogénéité sémantique des triplets

Nous examinons dans un premier temps les triplets pour lesquels les trois scores de proximité P(VbAg), P(VbAc) et P(AgAc) sont du même ordre. Nous identifions à ce titre deux types de configuration distinctes.

À l'image de ce que nous avons rencontré dans la section 4.1.2, la polysémie est une nouvelle fois à l'origine d'une proximité plus limitée entre le verbe et le nom d'action, qui contribue ici à niveler les trois scores de proximité. Nous l'illustrons avec les deux triplets présentés dans le tableau 4.4.

Verbe	Nom d'agent	Nom d'action	P(VbAg)	P(VbAc)	P(AgAc)
modérer	*modérateur*	*modération*	0.367	0.464	0.442
climatiser	*climatiseur*	*climatisation*	0.693	0.715	0.724

TABLE 4.4 – Scores de proximité des triplets construits autour des verbes *modérer* et *climatiser*

Les deux triplets présentés dans le tableau 4.4 se caractérisent par une polysémie de leur nom d'action, *modération* désignant à la fois une action et une entité collective (au même titre que *rédaction* ou *organisation*), et *climatisation* désignant à la fois une action et un instrument ('système de climatisation'). Les représentations vectorielles des noms identifiés comme noms d'action dans ces deux triplets intègrent de fait une composante agentive (ou instrumentale) qui vient diluer la composante actionnelle du prédicat, les distanciant par conséquent de leur verbe correspondant. On observe le même type de polysémie entre action et entité collective avec les triplets {*codiriger*, *co-directrice*, *codirection*} et {*codiriger*, *co-directeur*, *codirection*} par exemple.

On retrouve aussi parmi les triplets aux membres équidistants des triplets pour lesquels il est difficile de comprendre pourquoi ils ne se conforment que partiellement à l'hypothèse. C'est le cas notamment des triplets présentés dans le tableau 4.5.

Verbe	Nom d'agent	Nom d'action	P(VbAg)	P(VbAc)	P(AgAc)
aduler	*adulateur*	*adulation*	0.399	0.397	0.399
concasser	*concasseur*	*concassage*	0.587	0.647	0.642
pister	*pisteur*	*pistage*	0.440	0.427	0.422
torpiller	*torpilleur*	*torpillage*	0.544	0.616	0.586

TABLE 4.5 – Scores de proximité des triplets construits autour des verbes *aduler*, *concasser*, *pister* et *torpiller*

Les scores de proximité des triplets présentés dans le tableau 4.5 suggèrent, comme précédemment, une homogénéité sémantique relative au sein des triplets, puisque les trois items sont très proches dans l'espace vectoriel. Les noms d'agent, verbes et noms d'action impliqués dans ces triplets ne semblent pas présenter de polysémie ou de degré de lexicalisation induisant une prise de distance ou au contraire un rapprochement avec un des autres items du triplet. Ces triplets ne sont cependant pas des cas marginaux dans nos données, et si certains valident l'hypothèse d'une plus grande proximité du verbe et du nom d'action, à l'image des triplets construits autour de *torpiller* et *concasser* dans le tableau 4.5, il n'en reste pas moins que les différences semblent négligeables, et que la proximité attendue du verbe et du nom d'action n'est pas si significative. Une étude plus approfondie des contextes de ces mots, de leur fréquence, et de leurs spécificités serait donc nécessaire pour creuser davantage leur similarité.

4.2.2 Fort éloignement du nom d'action

Tous les triplets présentant des scores P(VbAc) et P(VbAg) similaires ne suivent cependant pas le schéma présenté dans la section 4.2.1 d'une homogénéité sémantique au sein des triplets. Ainsi, certains de ces triplets présentent un score P(AgAc) nettement différent des deux autres scores, qu'il soit bien plus élevé ou au contraire bien plus faible. Cette variation semble être une nouvelle fois liée à la polysémie de certains des membres des triplets, à l'image de ce qui a pu être observé précédemment.

Ainsi, il y a des cas où la polysémie induit une distanciation de deux des membres, mais sans autre conséquence majeure. C'est notamment le cas du triplet formé par *délégateur*, *déléguer* et *délégation* (P(VbAg)=0.336 ; P(VbAc)=0.334 ; P(AgAc)=0.115). Le nom *délégation* est polysémique, désignant à la fois l'action de déléguer et la personne ou l'entité collective qui a reçu une délégation. Le nom *délégation*, lorsqu'il dénote une personne ou entité, relève alors davantage du patient ou du bénéfiaire que de l'agent, ce qui l'éloigne à la fois du verbe et du nom d'agent.

Un second type de cas de polysémie est illustré, à l'image de ce qui a pu être présenté pour *essor* dans la section 4.1.2, par les triplets présentés dans le tableau 4.6. Les trois membres du triplet {*tirer*, *tireur*, *tirade*} sont polysémiques, mais *tirer* et *tireur* présentent les mêmes types d'acception, alors que *tirade* intègre, en plus de la dénotation de l'action de tirer – dans un sens vieilli et dans le domaine des sports (selon le TLFi) – une acception liée au discours. Le verbe *tirer* se trouve donc à l'interface entre deux paradigmes sémantiques, à l'image du verbe *essorer* dans le tableau 4.2. La représentation vectorielle de *tirade* diffère de fait fondamentalement de celle du verbe ou du nom d'agent. Il est cependant intéressant de noter que ce triplet valide néanmoins l'hypothèse d'une plus grande proximité du verbe et du nom d'action, puisque le score le plus élevé est le score P(VbAc). Les scores sont néanmoins globalement assez faibles, ce qui semble conforter l'hétérogénéité sémantique de ce triplet. Le même constat peut être fait pour le triplet formé par *doubler*, *doubleur* et *doublement*, où le nom d'agent et le verbe intègrent une dimension – celle du doublage de cinéma – dont est exempt le nom d'action. Cela devient particulièrement flagrant lorsque l'on examine le triplet {*doubler*, *doubleur*, *doublage*} (P(VbAg)=0.312 ; P(VbAc)=0.326 ; P(AgAc)=0.748), où les scores augmentent, notamment entre le nom d'agent et le nom d'action.

L'examen des triplets ayant des scores P(VbAg) et P(VbAc) similaires, mais un score P(AgAc) différent, fait aussi émerger les triplets présentés dans le tableau 4.7. Ces triplets nous font nous interroger sur l'impact de la synonymie, vu comme une concurrence, sur la proximité d'items lexicaux. Ainsi,

Verbe	Nom d'agent	Nom d'action	P(VbAg)	P(VbAc)	P(AgAc)
tirer	*tireur*	*tirade*	0.155	0.233	0.020
doubler	*doubleur*	*doublement*	0.312	0.284	0.009

TABLE 4.6 – Scores de proximité des triplets construits autour des verbes *tirer* et *doubler*

si *envahisseur* est bien l'agent de *envahir*, et *envahissement* l'action d'envahir, il nous semble que *envahisseur* n'est pas tant l'agent d'*envahissement*, mais plutôt d'*invasion*, qui occupe la même position sémantique que *envahissement* – du moins partiellement. On peut donc se demander dans quelle mesure cela influence la proximité entre le verbe et le nom d'action. Le même constat peut être fait pour *sauvetage* vis-à-vis de *sauveur* et *sauver*. Si on retrouve effectivement le prédicat associé à *sauver* dans *sauvetage*, *sauvetage* nous semble davantage associé au nom d'agent *sauveteur* (celui qui fait des sauvetages) qu'à *sauveur* (celui qui sauve). Cela semble confirmé par les scores de proximité associés au triplet formé par *sauver*, *sauveteur* et *sauvetage* (P(VbAg)=0.301 ; P(VbAc)=0.340 ; P(AgAc)=0.638).

Verbe	Nom d'agent	Nom d'action	P(VbAg)	P(VbAc)	P(AgAc)
envahir	*envahisseur*	*envahissement*	0.527	0.514	0.301
sauver	*sauveur*	*sauvetage*	0.396	0.340	0.121

TABLE 4.7 – Scores de proximité des triplets construits autour des verbes *envahir* et *sauver*

L'examen qualitatif des triplets sur la base des scores de proximité de leurs membres fait ainsi ressortir différents phénomènes pouvant expliquer l'invalidation de l'hypothèse par certaines familles dérivationnelles. La lexicalisation et le glissement de sens en jeu dans certains triplets relèvent du lexique, et questionnent donc l'association de certains mots sur des bases morphologiques. Il n'est donc pas étonnant que cela ressorte dans les espaces vectoriels, et est même souhaitable. Cela conforte l'utilisation des espaces vectoriels comme outil d'exploration des corpus, puisque les représentations qu'ils offrent sont cohérentes avec notre intuition et avec les phénomènes instanciés dans les corpus.

D'un autre côté, l'homonymie et la polysémie de certains mots montrent une des limites des modèles vectoriels, puisque ils ne permettent pas de discriminer les différents sens. Au contraire, ils les aggrègent. Cela nous empêche de fait de comparer de façon pertinente certains items. Par ailleurs, les usages

en corpus, en termes de contexte et de fréquence, ne permettent parfois pas de générer une représentation satisfaisante des mots, nuisant à leur comparaison.

Notons cependant que ces observations sont faites sur la base de l'examen d'un nombre relativement limité de triplets. Il faudrait systématiser leur analyse qualitative pour avoir une vision plus complète. Ces premiers constats suffisent néanmoins à montrer les avantages et les limites de notre approche. Le dispositif que nous avons présenté se montre intéressant pour le genre d'hypothèses que nous venons de traiter, impliquant la quantification d'une proximité sémantique à grande échelle, mais montre ses limites dès lors qu'on produit une analyse à l'échelle des familles dérivationnelles.

Cet outil n'est donc pas le plus adapté lorsqu'on veut produire une analyse fine, sans contrôle spécifique sur les données. Les modèles prédictifs ne permettent pas de gérer l'ambiguïté formelle, sémantique ou syntaxique, ni les situations d'homonymie et de polysémie qui viennent brouiller les résultats. Si des solutions peuvent être apportées lorsque l'ambiguïté subsiste entre deux lexèmes de catégories grammaticales distinctes, seuls les modèles contextuels pourraient permettre de pallier l'ambiguïté sémantique. Or, comme nous l'avons signalé dans la section 1.2.1, ces modèles ne sont pas adaptés à nos besoins.

Le dispositif présenté se montre néanmoins suffisant dans le cadre d'une étude quantitative à grande échelle et à gros grain, et offre des résultats satisfaisants, pour un coût computationnel minime. Nous avons ainsi pu valider globalement l'hypothèse d'une plus grande proximité entre le verbe et son nom d'action au sein de familles dérivationnelles. Un plus grand contrôle sur les données, par une sélection visant à limiter la lexicalisation et la polysémie, permettrait sans doute de conforter cette hypothèse, notamment à l'échelle d'un plus grand nombre de familles dérivationnelles, mais cela implique un traitement préalable qui ne semble pas nécessaire dans une approche quantitative à gros grain comme celle que nous présentons.

Troisième partie

Analyse comparative des noms déverbaux en *-eur*, *-euse* et *-rice*

L'objectif de l'expérience présentée en partie II était de tester l'hypothèse d'une différence de proximité entre les membres de familles dérivationnelles. La représentation vectorielle du sens dans un espace distributionnel fournit un moyen d'opérationnaliser cette hypothèse en quantifiant la proximité sémantique. Cela nous a permis de mettre en évidence, à l'échelle de l'ensemble des familles dérivationnelles considérées, une tendance à une plus grande proximité entre un verbe et ses nominalisations actives qu'entre le verbe et ses noms d'agent ou d'instrument en *-eur*, *-euse* et *-rice*. Nous avons aussi pu observer des comportements variables en termes de proximité au verbe en fonction du suffixe. Des différences s'observaient notamment pour les suffixes *-eur*, *-euse* et *-rice*, sur lesquels nous allons nous focaliser dans le cadre de cette partie.

Ces différences ont fait l'objet d'un certain nombre d'études (Houdebine-Gravaud 1998, Schafroth 2001, Dawes 2003, entre autres), visant à souligner la présence d'une valeur sémantique supplémentaire du féminin par rapport au masculin d'une part, et l'existence d'une différence de connotation entre les deux suffixations féminines d'autre part. Mais ces études reposent sur des bases introspectives, ou limitées en termes d'items observés.

Dans cette partie III, nous souhaitons creuser l'étude des noms déverbaux en *-eur*, *-euse* et *-rice*. Plus précisément, nous cherchons à comparer sur le plan sémantique les noms construits à partir des suffixations en *-eur*, *-euse* et *-rice*, afin de savoir si la féminisation introduit un profil sémantique différent, et s'il est observable sur le plan distributionnel. Nous proposons dans ce travail d'exploiter les représentations vectorielles pour approcher cette comparaison dans une approche extensive et quantitative. Nous nous demandons (i) comment la sémantique distributionnelle permet de capter le sens d'un ensemble de mots à l'échelle d'une catégorie sémantique, à savoir celle des noms d'agent, et (ii) dans quelle mesure elle permet d'apporter un éclairage nouveau sur les différences et similarités observées localement entre les noms construits par les suffixes *-euse* et *-rice*.

Contrairement à l'étude présentée précédemment, qui se situait à l'échelle des familles dérivationnelles, nous cherchons ici à représenter les classes formées par les noms en *-eur*, *-euse* et *-rice*, afin de pouvoir les caractériser dans leur globalité sur le plan sémantique. Il s'agit plus précisément de caractériser les zones de l'espace vectoriel distributionnel occupées par les noms en *-eur*, *-euse* et *-rice*. Pour ce faire, nous utilisons une représentation globale de ces classes, que nous analysons ensuite à l'aune de leurs voisins distributionnels. Ceux-ci fournissent des indices permettant de caractériser la zone dans laquelle un point de l'espace vectoriel se situe, autrement dit des repères pour l'interprétation des représentations globales dans l'espace distributionnel.

Outre cet apport méthodologique, nous montrons dans cette partie que la

représentation unifiée des classes formées par les noms en *-eur*, *-euse* et *-rice* permet d'accéder à des différences sémantiques associées à ces suffixes. Nous vérifions ainsi pour le français l'ancrage de ces différences dans les usages, tant entre le masculin et le féminin, plus orientées vers le genre féminin du référent, qu'entre les suffixes *-euse* et *-rice*, basées sur des connotations péjoratives ou sexuelles.

Nous développons et testons notre méthode de représentation de classes de mots dans le chapitre 5 sur la base de l'étude des noms en *-eur*, puis nous appliquons dans le chapitre 6 cette méthode aux noms en *-euse* et *-rice* afin de faire émerger les différences sémantiques de ces deux classes de noms dérivés.

Le travail présenté dans cette partie a fait l'objet de plusieurs publications (Wauquier 2018, Wauquier *et al.* 2018 2020a). Les données présentées sont accessibles à l'adresse `https://github.com/mwauquier/PhdData/tree/main/Part3`

Chapitre 5

Représentation de la classe des noms en *-eur*

La variation dans la proximité distributionnelle observée entre le verbe et ses dérivés agentifs et instrumentaux en fonction du suffixe de ces derniers suggère l'existence d'une différence sémantique entre les noms que les suffixes *-eur*, *-euse* et *-rice* construisent. Ces différences ont été étudiées dans la littérature, particulièrement celles entre les noms d'agent masculins (*entraîneur*) et féminins (*entraîneuse*), mais pas de façon exhaustive pour le français. Notre objectif est d'en proposer une analyse fondée sur la sémantique distributionnelle.

Ce chapitre expose une méthode de représentation d'une classe de mots qui permet de capter certaines composantes sémantiques partagées par les membres de cette classe. Cette méthode unifie l'analyse d'une classe par l'examen de ses plus proches voisins distributionnels. La méthode est appliquée aux noms déverbaux en *-eur*, et met en exergue comme attendu les interprétations agentive et instrumentale associées à cette suffixation. L'examen de cette classe nous permet par ailleurs d'approfondir plusieurs aspects méthodologiques liés aux représentations vectorielles.

Nous présentons dans ce chapitre les fondements linguistiques et méthodologiques de l'étude comparative systématique des suffixes déverbaux *-eur*, *-euse* et *-rice*. Nous examinons dans la section 5.1 les différences et spécificités des trois suffixes, en mettant l'accent sur celles que nous souhaitons explorer. Nous présentons ensuite dans la section 5.2 les choix méthodologiques que nous adoptons, et en proposons une première évaluation qualitative. Enfin nous explorons dans la section 5.3 l'impact de nos choix méthodologiques en évaluant l'impact de différentes configurations sur nos analyses du suffixe *-eur*.

5.1 Les suffixes *-eur*, *-euse* et *-rice*

Les suffixes *-euse* et *-rice*, du fait de la similarité de leurs règles de contruction, sont considérés comme les équivalents féminins du suffixe déverbal *-eur*. S'ils permettent au même titre que le suffixe *-eur* de dénoter des noms d'instrument (*moissonneuse*, *calculatrice*) et des noms d'agent (*nageuse*, *directrice*), les dérivés qu'ils construisent diffèrent, tant des dérivés masculins qu'entre eux. Nous présentons dans la section 5.1.1 les différences entre masculin et féminins évoquées dans la littérature, puis nous creusons plus spécifiquement dans la section 5.1.2 les différences entre les deux suffixes féminins.

5.1.1 Oppositions de genre

Comme nous l'avons vu dans la section 3.1.2, *-eur* est le suffixe prototypique pour la formation des noms d'agent. Il prend traditionnellement en *input* une base verbale (*danser*, *organiser*) et forme un nom désignant l'agent qui réalise l'action dénotée par la base verbale (*danseur*, *organisateur*). Les suffixes *-euse* et *-rice* en sont les équivalents féminins les plus productifs (Anscombre 2001, Dawes 2003). Ces suffixes permettent en effet de construire des noms d'agent déverbaux dénotant les référents féminins qui réalisent l'action dénotée par la base verbale (*danseuse*, *organisatrice*).

De fait, les noms d'agent déverbaux en *-eur*, *-euse* et *-rice* diffèrent sur la base du genre. Cette distinction opère sur plusieurs plans, les plus évidents étant le plan grammatical et le plan référentiel. Un troisième plan a été évoqué : le plan axiologique. En effet, outre le genre du référent, la stricte équivalence des noms d'agent féminins et masculins a longuement été remise en question. Des travaux soulignent ainsi la présence d'une valeur sémantique supplémentaire du féminin liée aux biais culturels, à l'image de *mister* 'monsieur' et *mistress* 'maîtresse' en anglais (Marcato et Thüne 2002, Hellinger et Bußmann 2001), ou *entraîneur* et *entraîneuse* en français.

Historiquement, la féminisation progressive des noms de métiers et fonctions depuis le courant du 20^{e} siècle, a permis une translation de la dénotation de l'épouse de celui qui occupe une fonction donnée – à l'image des noms *ambassadrice*, *sénatrice* et *maréchale* qui désignaient la femme de l'ambassadeur, du sénateur et du maréchal, (Le Draoulec et Péry-Woodley 2016b, Cerquiglini 2018) – à la dénotation de l'agent féminin (le nom *directrice* désigne la femme qui dirige quelque chose) (Dawes 2003). Cette féminisation s'est traduite par deux courants parallèles. Le premier courant consiste en l'utilisation du genre unique pour les métiers valorisés ou jugés valorisants, notamment à l'aide de formes épicènes (*maire*, *médecin*) ou masculines (Dawes 2003). Le

second courant a vu le développement de la féminisation mais uniquement à destination des métiers peu valorisés ou jugés peu valorisants (*travailleuse*, *institutrice*) (Dawes 2003). De fait, les noms d'agent féminins ont longtemps été plus dépréciatifs que les noms d'agent non marqués grammaticalement.

Plus récemment, un certain nombre d'études ont montré que les noms d'agents masculins et féminins ne sont pas sémantiquement identiques. Zeller *et al.* (2014) remarquent que les agentifs féminins contenus dans la ressource morphologique de l'allemand DerivBase (Zeller *et al.* 2013) sont globalement plus distants du lexème de base, sur le plan sémantique, que les agentifs masculins. Ils expliquent cela par le fait que l'agent féminin ne serait utilisé que dans des contextes où le genre a de l'importance. Le suffixe agentif féminin en allemand serait donc particulièrement marqué vis-à-vis du genre. Dans le même ordre d'idée, Dawes (2003) souligne la connotation sexuelle des noms de métiers et fonctions au féminin, où le nom féminin fait davantage référence au sexe, là où le nom équivalent masculin fait simplement référence à l'humain agent. Bolukbasi *et al.* (2016) et Caliskan *et al.* (2017) montrent quant à eux que les représentations sémantiques obtenues de façon automatique à partir de corpus traduisent des biais culturels et sociologiques notamment liés au genre. Les auteurs soulignent par exemple que des mots féminins comme *woman* ou *girl* ne sont pas rapprochés des mêmes champs lexicaux que des mots masculins, et que ces champs lexicaux traduisent les stéréotypes culturels en cours. Des mots comme *woman* et *girl* sont davantage associés aux arts, à l'inverse des noms masculins, associés aux mathématiques et aux sciences. De même, les prénoms féminins sont davantage rapprochés de termes liés à la famille qu'à la carrière, contrairement aux prénoms masculins.

Notons que l'opposition référentielle et axiologique entre les noms déverbaux en *-eur*, *-euse* et *-rice* ne vaut que pour les noms d'agent. Car à l'image du suffixe *-eur*, les suffixes *-euse* et *-rice* permettent aussi de construire des noms d'instrument (*essoreuse*, *calculatrice*). L'opposition entre masculin et féminin n'a alors qu'une valeur grammaticale, et n'a pas d'effet sur le plan sémantique. Des hypothèses ont cependant été avancées concernant l'évolution de la dénotation d'instruments par ces suffixes en diachronie.

Dubois (1962) suggère en effet que le suffixe *-eur* servait historiquement à désigner de façon exclusive un agent ou un instrument, à l'image des noms *téléspectateur* et *téléviseur*, un même nom ne pouvant désigner à la fois l'agent et l'instrument. L'utilisation du suffixe féminin permettait de remédier à cela, à l'image du couple *moissonneur* (agent) et *moissonneuse* (instrument), distinguant de façon explicite les deux lectures construites sur la base d'un même prédicat. Dubois (1962) émet l'hypothèse qu'avec l'utilisation croissante des machines, et leur substitution à l'homme, cette distinction s'est effacée, les suffixes masculins et féminins s'utilisant alors de façon indifférenciée pour

former des noms d'agents ou des noms d'instruments. On se retrouve ainsi avec un ensemble non cohérent sur le plan sémantique de noms en *-eur* ne désignant pas uniquement des noms d'agents et de noms en *-euse* ne désignant pas uniquement des noms d'instruments.

De fait, les noms grammaticalement masculin et féminin issus d'un même paradigme dérivationnel ne désignent pas nécessairement la même réalité, et la variation du genre grammatical ne traduit pas exclusivement une variation du genre référentiel de l'entité dénotée. Dans certains cas, les deux noms désignent de façon complémentaire – ou presque – l'agent (*moissonneur*) et l'instrument (*moissonneuse*[1]). On observe aussi dans d'autres cas que les deux noms désignent tous les deux des instruments (*calculateur* / *calculatrice*). Les deux suffixes sont alors utilisés pour désigner deux entités distinctes, dénotant deux réalités distinctes bien que fondées sur un même prédicat (l'action de calculer).

L'hypothèse d'une spécialisation des suffixes féminins pour la dénotation de noms d'instrument n'a cependant à notre connaissance pas été démontrée, et le phénomène reste à évaluer en synchronie. Des différences attestées existent aussi entre les suffixes *-eur*, *-euse* et *-rice* à l'échelle des noms d'agent – dont nous avons vu dans la section 3.1.2 qu'ils tendent à être plus représentés parmi les noms déverbaux en *-eur* que les noms d'instrument. L'étude des suffixes *-eur*, *-euse* et *-rice* met cependant en lumière d'autres différences, cette fois entre les suffixes féminins eux-mêmes.

5.1.2 Hiérarchisation des suffixes

L'existence de plusieurs équivalents féminins pour le suffixe *-eur* pousse à s'interroger sur la spécificité de chaque suffixe. Si la similarité de leurs règles de construction suggère que *-euse* et *-rice* sont en réalité deux exposants d'une unique règle, plusieurs différences ont été avancées.

Une première différence relève de considérations formelles. Le suffixe *-euse* ne peut produire un nom d'agent que s'il est ajouté à une base de type verbal, à partir du thème de présent. Le suffixe *-rice* sélectionne préférentiellement pour sa part le thème de supin des bases verbales, mais admet marginalement d'autres types de formation, notamment par analogie (Dal 2003). L'analogie explique notamment la formation de *sénatrice*, à l'image de *sénateur*, sans passer par un verbe **séner* ou un nom d'action **sénation* non attestés.

Ces contraintes formelles n'empêchent cependant pas l'existence de cas de concurrence entre deux formes construites sur une même base, à l'image

1. Le nom *moissonneuse* désigne aussi l'agent féminin réalisant l'action de moissonner, mais l'usage de *moissonneuse* en corpus correspond très largement à la lecture instrumentale.

de *sculpteuse* et *sculptrice* à partir de *sculpter*. Des différences d'ordre sémantique ont ainsi été mises en avant, permettant d'expliquer l'existence de tels doublets. De nombreuses études ont montré l'existence d'une opposition axiologique entre ces noms.

Si nous avons vu dans la section 5.1.1 que la féminisation marquée était intrinsèquement dépréciative en diachronie, cette féminisation s'est aussi caractérisée par une hiérarchisation axiologique des suffixes féminins (Dawes 2003). Le suffixe *-euse* est moins bien considéré que *-rice*, jugé plus « noble »(mais aussi moins que *-eure*) (Houdebine-Gravaud 1998, Dawes 2003, Lenoble-Pinson 2008). Le suffixe *-euse* renvoie ainsi à des métiers peu qualifiés, comme *repasseuse*, *nettoyeuse*, *vendeuse* ou *coiffeuse*, quand le suffixe *-rice* serait favorisé pour des positions valorisantes ou supérieures, comme *directrice* ou *sénatrice*. Houdebine-Gravaud (1998) évoque quant à elle l'existence de normes fictives, en plus de considérations esthétiques (Le Draoulec et Péry-Woodley 2016a) qu'appliqueraient les locuteurs, liées au poids du biais sociolinguistique de ces suffixes. Notons que cette tendance n'est pas exclusive au français, et qu'elle se retrouve dans d'autres langues romanes et germaniques. Les suffixes italien *-essa*, roumain *-esa*, et allemand *-ess* sont tous fortement connotés, dans un sens dépréciatif ou vulgaire (Maurice 2001, Marcato et Thüne 2002, Bußmann et Hellinger 2003, Dawes 2003). Des suffixes dénués de connotation existent en parallèle de ces suffixes, à l'image du suffixe italien *-trice* et du suffixe allemand *-in*.

Une des limites de ces études relève de leur méthodologie. Les analyses sont largement menées sur la base d'observations localisées, non systématiques, notamment dans le cas des différences de connotation entre les suffixes *-euse* et *-rice*, ou sont proposées pour d'autres langues que le français, à l'image de la comparaison du masculin et du féminin. Nous cherchons dans cette partie III à comparer à large échelle les noms formés par les suffixes *-eur*, *-euse* et *-rice* au regard de ces différents constats. Il s'agit d'une part de se demander si on vérifie de façon plus systématique pour le français les différences sémantiques observées entre les suffixes masculin et féminins, et si, d'autre part, la différence de connotation entre les deux suffixes féminins *-euse* et *-rice* est une tendance propre à l'ensemble des noms formés par ces suffixes. Nous déployons pour cela une approche extensive et automatique, que nous décrivons dans la section 5.2. Précisons que notre approche est centrée autour des suffixe, et non des classes lexicales. Nous considérons à ce titre l'ensemble des noms d'agent et d'instrument construits par les suffixes ciblés, contrairement à la démarche que nous déployons dans la partie IV, où nous concentrons spécifiquement sur la catégorie lexicale des noms d'agent.

5.2 Une représentation pour les aggréger tous

Notre étude porte sur les différences entre les trois suffixes non pas à l'échelle du mot mais à l'échelle de la suffixation. Pour pouvoir tirer des généralisations, il est nécessaire d'envisager la classe formée par les mots construits par la suffixation comme un tout, qu'il faut analyser comme une entité unique. Nous présentons pour cela dans la section 5.2.1 une méthode qui repose sur l'hypothèse compositionnelle des vecteurs et d'une représentation uniforme des traits sémantiques dans les espaces vectoriels. Nous appliquons cette méthode dans la section 5.2.2 aux noms déverbaux pour permettre première évaluation subjective.

5.2.1 Du mot à la classe

Comme nous venons de le signaler, une grande partie des différences sémantiques et pragmatiques entre les noms déverbaux en *-euse* et en *-rice* évoquées dans la section 5.1 reposent sur une analyse locale du contexte ou sur l'introspection, à partir d'un nombre d'exemples limités. Or, ces méthodes ont des limites car elles ne donnent qu'un aperçu partiel de la situation. D'une part, elles permettent uniquement une analyse de mots individuels, et n'offrent pas de vision globale, à l'échelle de l'ensemble des noms d'un groupe donné. D'autre part, ces analyses ne prennent pas nécessairement en compte les usages d'un mot, notamment lorsqu'elles reposent sur l'introspection. De ce fait, elles ne fournissent pas une véritable vision d'ensemble des propriétés sémantiques de ce groupe.

Les modèles distributionnels permettent justement d'adopter une approche plus extensive en prenant en compte la distribution de l'ensemble des occurrences d'une forme donnée pour le calcul de sa représentation vectorielle. Ces modèles peuvent ensuite être « interrogés »afin d'accéder à la description sémantique (par le biais de leurs voisins) d'un très grand nombre de formes.

Comme nous l'avons vu dans la section 1.1.1, les voisins distributionnels d'un nom en *-eur* donné fournissent un mini-réseau de mots qui situe ce nom dans un espace sémantique plus ou moins bien identifié. À titre d'exemple, le nom *danseur* a comme trois plus proches voisins dans le modèle utilisé dans la partie II les noms *chorégraphe*, *ballerine* et *musicien*. L'examen des plus proches voisins offre ainsi une première approximation sémantique de l'usage des mots. Au-delà de cette utilisation classique du voisinage distributionnel, nous nous intéressons ici à la possibilité d'accéder à la représentation d'une classe de mots, sans devoir passer par l'examen individuel de l'ensemble des mots de la classe. L'objectif est donc de représenter la classe des mots formés

par un suffixe donné au moyen d'un vecteur unique que l'on peut caractériser à l'aune de ses voisins distributionnels.

Nous admettons dans la suite de ce travail que les noms formés par une même suffixation partagent une composante sémantique qui leur permet de former un groupe suffisamment homogène sur le plan distributionnel, malgré des variations intrinsèques. Nous formulons alors l'hypothèse que cette composante sémantiquement homogène, construite autour du sens partagé par les membres du groupe, peut être décrite comme une entité. Se pose alors la question de la réprésentation de cette entité dans les modèles distributionnels, puisque, en tant qu'abstraction sémantique, elle n'est pas directement instanciée dans le corpus et n'est donc pas directement accessible sous la forme d'un vecteur distributionnel.

Comme nous l'avons vu dans la section 1.1.1, certains auteurs eploitent la compositionnalité des vecteurs pour construire la représentation vectorielle d'affixes ou de procédés morphologiques à partir des vecteurs de lexèmes dérivés. L'objectif de ces études est principalement d'améliorer la représentation des dérivés morphologiques, à l'aide des représentations de leurs suffixes. Notre objectif est tout autre puisque nous ne souhaitons pas représenter un suffixe mais un groupe sémantique formé par les noms construits par ce suffixe. Nous ne cherchons pas à intervenir sur la représentation de ces mots, dont la construction morphologique n'est qu'un fait et non un but, mais souhaitons seulement analyser ces représentations telles qu'elles apparaissent dans les modèles distributionnels.

C'est donc dans les travaux en syntaxe et sémantique, plutôt qu'en morphologie, que nous puisons notre inspiration. Bien qu'ils répondent à des objectifs différents, des travaux comme (Kintsch 2001) et (Erk et Padó 2008) entre autres proposent de représenter un groupe de mots partageant des caractérisques sémantiques communes sous la forme d'un unique vecteur.

Kintsch (2001) présente un algorithme permettant de désambiguiser le sens d'un prédicat en fonction de ses arguments dans des modèles de LSA. Cet algorithme construit ainsi des représentations distinctes pour les formes *ran* présentes dans les phrases *the horse ran* 'le cheval a couru' et *the color ran* 'la couleur a coulé'. Ces représentations sont obtenues en calculant le barycentre (*centroid* dans le texte) du prédicat, de l'argument et de cinq mots sémantiquement proches à la fois du prédicat et de l'argument. Notons qu'ici les mots proches sont sélectionnés automatiquement parmi les voisins du prédicat sur la base de leur proximité avec l'argument. Ainsi, le barycentre de la phrase *the horse ran* est calculé à partir des vecteurs des mots *horse* (l'argument), *ran* (le prédicat), *stopped*, *yell*, *came*, *saw* et *ran* (les cinq mots les plus fortement liés au prédicat et à l'argument, le prédicat étant son propre voisin). Ce nouveau vecteur pour la forme *ran* s'avère ainsi plus proche du

vecteur du verbe *gallop* (0.75) que ne l'était le vecteur non désambiguisé (0.33). Le barycentre construit à partir de mots partageant des spécificités sémantiques permet ainsi de capter celles-ci.

Un principe similaire est utilisé par Erk et Padó (2008) dans leur travail sur les restrictions de sélection des verbes. Afin d'évaluer de façon automatique la plausibilité pour un mot d'être l'argument d'un verbe donné, les auteurs construisent un barycentre pondéré à partir des vecteurs des arguments attestés du verbe. La proximité distributionnelle d'un argument candidat à ce vecteur moyen dans un modèle compositionnel permet alors d'évaluer la plausabilité que ce candidat soit un argument du verbe cible. Les auteurs montrent ainsi que cette méthode est très performante, et que les scores de proximité obtenus sont significativement corrélés aux jugements des locuteurs sur cette même tâche.

Ces deux études suggèrent que le vecteur moyen des vecteurs de mots partageant des propriétés sémantiques fournit une bonne approximation de la façon dont les sens de ces mots se réalisent dans différentes configurations. Plus précisément, l'approche adoptée par Erk et Padó (2008) fait appel à l'intuition que le vecteur construit à partir des arguments attestés du prédicat aggrège les sens de ces arguments, et constitue de fait une représentation du sens exprimé par la classe formée par les arguments (Padó *et al.* 2013). Le barycentre peut être pensé comme la représentation d'une sorte de sens moyen des arguments. C'est cette approche que nous choisissons de transposer dans cette étude, car nos objectifs sont similaires. Notons cependant que nous raisonnons au niveau paradigmatique, alors que les travaux précédents opèrent sur le plan syntaxique. Nous cherchons en effet à construire un vecteur à partir d'un ensemble de noms dont nous estimons qu'ils partagent les propriétés sémantiques induites par leur construction, afin de capter ces propriétés partagées. Nous postulons donc que les noms en *-eur* partagent les propriétés sémantiques associées à cette suffixation – l'agentivité pour les noms d'agent et l'instrumentalité pour les noms d'instrument –, indépendamment des spécificités liées à leur base. Nous faisons l'hypothèse subséquente que ces propriétés sémantiques partagées sont représentées de manière uniforme et qu'elles seront captées par ce vecteur moyen. En observant les barycentres calculés à partir de noms d'agent et d'instrument construits par différents procédés dérivationnels, nous pouvons comparer l'expression de cette propriété sémantique en fonction du procédé.

Nous choisissons donc de créer une représentation vectorielle unique pour l'ensemble des noms porteurs d'un même suffixe en faisant la moyenne de leurs représentations vectorielles, selon la formule (5.1) où $\overrightarrow{C}$ correspond au

vecteur moyen, que nous appellerons désormais *barycentre* [2], construit à partir des n vecteurs $\vec{v_i}$ (que nous appellerons *amorces*) des noms w appartenant à la classe S définie par un suffixe donné.

$$\begin{aligned} S &= \{w_1, ..., w_n\} \\ \vec{v_i} &= V(w_i), 1 \leq i \leq n \\ \overrightarrow{C} &= \frac{1}{n} \sum_{i=1}^{n} \vec{v_i} \end{aligned} \tag{5.1}$$

Le vecteur ainsi construit est un objet abstrait puisqu'il ne correspond pas à la représentation vectorielle d'une forme présente dans le corpus construite sur la base de ses contextes. Pour évaluer le contenu sémantique encapsulé par le barycentre, nous observons les p plus proches voisins du vecteur moyen. Nous considérons les plus proches voisins distributionnels comme étant représentatifs du groupe de noms agrégés par le barycentre. À ce titre, on peut considérer que le barycentre est un vecteur construit qui nous permet d'aller sonder une zone spécifique de l'espace vectoriel dont on fait l'hypothèse qu'elle s'organise autour d'un trait sémantique spécifique, ici l'agentivité et l'instrumentalité.

Le choix du nombre des plus proches voisins à analyser est arbitraire. Il n'existe à notre connaissance aucune étude définissant un seuil à partir duquel le voisinage n'est plus pertinent [3]. Dans la mesure où nous souhaitons faire une description sémantique riche et extensive du barycentre, nous fixons ce nombre à 100, afin que nos observations soient suffisamment générales tout en permettant une analyse manuelle fine du voisinage.

2. Nous avons dans un état antérieur du travail utilisé le terme de vecteur ou représentation prototypique pour désigner ce vecteur moyen construit. En s'inscrivant dans la perspective d'une catégorisation graduelle telle que décrite par Kleiber (1990), le barycentre était considéré comme le dérivé prototypique d'une suffixation donnée, et donc le représentant prototypique de la catégorie sémantique associée à cette suffixation, en cela qu'il instancierait le plus de propriétés caractéristiques de cette catégorie. Rien ne dit cependant que les propriétés prototypiques des noms d'agent soient toutes encapsulées par le barycentre. Nous considérons plutôt ce vecteur moyen comme le dérivé moyen construit par le procédé, car il moyenne les propriétés de l'ensemble des noms qui le construisent.

3. Des seuils ont été calculés et proposés, mais ces seuils sont envisagés à l'aune de la variation des voisinages, dans une approche quantitative (Pierrejean 2020). La définition de ces seuils ne préjuge en rien de la significativité et de la pertinence des voisins conservés pour la description du vecteur ciblé.

5.2.2 Étude appliquée aux noms déverbaux en *-eur*

Si la méthode des barycentres s'est révélée efficace dans les travaux précédemment cités, nous devons vérifier qu'elle permet bien de capter le trait sémantique partagé à l'échelle des noms en *-eur*, *-euse* et *-rice* (*a priori* l'agentivité et l'instrumentalité). On peut cependant faire l'hypothèse que les spécificités sémantiques associées aux noms féminins, à savoir le genre féminin et la possible connotation des suffixes, peuvent compliquer l'observation des traits agentifs et instrumentaux. Nous analysons donc dans un premier temps la représentation de l'agentivité et de l'instrumentalité à l'aide des noms déverbaux en *-eur*, ce qui nous permettra de valider la méthode et qui nous fournira un point de référence pour la comparaison des suffixes *-euse* et *-rice* étudiés dans le chapitre 6. Nous considérons l'approche par barycentre comme valide si le barycentre parvient à capter les traits sémantiques supposément partagés par les noms déverbaux en *-eur*. Plus précisément, la validation de la méthode repose sur l'examen qualitatif des voisins, dont nous faisons l'hypothèse qu'ils seront très majoritairement des noms d'agent et d'instrument.

Pour ce faire, nous utilisons le modèle distributionnel présenté dans la partie II, construit par Word2Vec avec les paramètres de base, sur le corpus *Wikipedia2018*. Nous utilisons comme amorces les noms masculins déverbaux en *-eur* de la ressource Lexeur [4]. Nous extrayons dans un premier temps l'ensemble des formes en *-eur* issues de familles déverbales, soit un total de 4 588 formes distinctes, et ne conservons que les formes présentes dans le modèle distributionnel utilisé (c'est-à-dire ayant une fréquence supérieure ou égale à 5), soit 1 675 formes. Le barycentre est construit en faisant la moyenne des vecteurs de ces 1 675 noms déverbaux en *-eur* selon la formule 5.1. Nous en étudions les 100 plus proches voisins. Un aperçu des dix premiers voisins et de leur score de proximité au barycentre est donné dans le tableau 5.1.

On remarque premièrement que les 100 plus proches voisins du barycentre correspondent tous à des noms (*technicien*, *client*) ou à des formes ayant au moins une acception nominale (*tranquillisant*, *accessoire*). Parmi ces 100 voisins, 45 correspondent à des amorces utilisées pour la construction du barycentre (à l'image de *aspirateur* ou de *programmeur*). Notons que ces 45 noms se répartissent uniformément dans le voisinage considéré, et ne

4. Comme indiqué dans la section 3.2.1, Lexeur intègre indistinctement des noms d'agent et des noms d'instrument. Dans la mesure où nous adoptons une approche à l'échelle du suffixe et pas de la classe lexicale, nous ne cherchons pas à distinguer les deux lectures. Nous faisons l'hypothèse que cela ne nuira pas à la comparaison ultérieure, dans le chapitre 6, des suffixations en *-eur*, *-euse* et *-rice*, dans la mesure où la double lecture agentive et instrumentale concerne les trois suffixations.

Voisin	Score de proximité
aspirateur	0.683
plombier	0.677
client	0.660
mécano	0.657
machiniste	0.652
pousseur	0.650
garagiste	0.647
conducteur	0.641
nettoyeur	.639
soudeur	0.635

TABLE 5.1 – Dix plus proches voisins du barycentre des noms déverbaux en *-eur*

correspondent pas aux 45 plus proches voisins du barycentre. Nous en tirons deux conclusions. Tout d'abord, cela suggère qu'une partie des noms déverbaux en *-eur* occupent bien une région spécifique de l'espace vectoriel, au cœur de laquelle se trouve le barycentre. Cela montre par ailleurs le caractère heuristique de la méthode, qui permet d'identifier d'autres noms au cœur de la classe considérée (les noms d'agent et d'instrument), et pas seulement les noms en *-eur* ayant servi d'amorces. Parmi ces autres voisins, on trouve un seul autre nom en *-eur*, *prestidigitateur*. Bien que présent dans Lexeur, *prestidigitateur* n'a pas été pris comme amorce car il ne s'agit pas d'un nom déverbal mais d'un nom complexe non construit. Les 54 autres voisins sont morphologiquement très variés. On retrouve notamment des noms composés (*marteau-piqueur*, *pied-de-biche*, *microphone*), des emprunts à l'anglais (*pacemaker*, *taser*), des noms simples (*client*), ainsi que des noms construits par d'autres affixes (*technicien*, *armurier*, *garagiste*, *antivol*). La région de l'espace dans laquelle on se situe ne se résume donc pas aux seuls noms en *-eur*. À présent, la capacité à interpréter ou non les caractéristiques communes de ces voisins définira l'homogénéité ou l'hétérogénéité sémantique de la région sondée par le barycentre.

Sur le plan sémantique, les 100 plus proches voisins du barycentre des noms en *-eur* semblent relativement homogènes. On retrouve 32 noms d'agent (*plombier*, *mécano*, *machiniste*), 45 noms d'instrument (*grille-pain*, *lance-pierre*, *grappin*), ainsi que 16 noms à double lecture agent et instrument (*mouchard*, *polisseur*). Cette homogénéité semble confirmée par la présence minoritaire, au nombre de 7, de voisins dont le sens ne relève pas de la lecture agentive et/ou instrumentale (*double-saut*, *accessoire*, *chien*). Nous qualifions de nom d'agent tout nom dénotant l'humain réalisant une action.

Est considéré comme tel tout nom ayant au moins une acception agentive, quand bien même la forme est associé à un autre lexème qu'un nom d'agent. Nous simplifions ici la définition donnée dans la section 3.1.2 en limitant les entités animées dénotées aux humains, afin de la rendre opérationnelle et de permettre une annotation sémantique à gros grain des voisins. Nous verrons dans la partie IV que cette définition n'est pas pleinement satisfaisante, et qu'elle nécessite d'être raffinée pour un travail à grain fin. Cette définition suffit néanmoins à ce stade du travail, pour l'analyse que nous proposons. Les noms d'agent correspondent principalement à des métiers, pour la plupart manuels (*déménageur*, *soudeur*, *plombier*). Une grande partie des noms d'instrument correspondent quant à eux à des instruments du quotidien [5] (*sèche-cheveux*, *tournevis*, *interrupteur*). Enfin, parmi les sept voisins restants, on retrouve cinq noms d'artefact (*robot*, *arceau-cage*, *gadget*).

Les deux voisins restants semblent se démarquer, à savoir *chien* et *double-saut*. Le premier désigne en effet un animal, dont on peut questionner le caractère agentif notamment dans des syntagmes comme *chien de garde* ou *chien de berger* (selon qu'on considère le trait humain comme nécessaire ou non à la qualification en tant que nom d'agent – voir chapitre 7). Sa présence dans le voisinage peut cependant s'expliquer par son utilisation sur le plan syntaxique en tant qu'agent de verbes couramment utilisés pour des agents humains, le rapprochant de fait de certains noms d'agent. Le nom *double-saut* désigne quant à lui une action dans le domaine des jeux vidéo. On peut donc à ce titre penser que sa présence dans le voisinage du barycentre des noms d'agent en *-eur* est idiosyncratique, lié à la nature du corpus *Wikipedia2018*. Son caractère instrumental dans le corpus est en tout cas confirmé par la présence parmi ses plus dix plus proches voisins de noms d'instruments comme *taser*, *grappin*, ou *téléporteurs*.

À l'issue de cette expérience, nous pouvons conclure que la zone sondée à l'aide du barycentre des noms déverbaux en *-eur* est sémantiquement homogène, et semble caractérisée par l'agentivité et l'instrumentalité. La représentation d'une catégorie sémantique au moyen d'un vecteur moyen semble de fait être une méthode efficace pour un travail exploratoire de ce type.

5.3 Variations de paramètres

L'implémentation de l'expérience que nous venons de présenter implique à différents plans des choix pouvant avoir un impact sur les résultats. Ces choix peuvent être relatifs aux données utilisées en *input* (section 5.3.1) et

5. On peut faire l'hypothèse que les instruments plus techniques sont des termes complexes, ne faisant donc pas l'objet d'un vecteur unique.

aux espaces vectoriels sur lesquels l'expérience est bâtie (section 5.3.2). Nous explorons ces deux aspects dans les sections suivantes.

5.3.1 Choix relatifs aux amorces

Il est légitime de se poser la question de l'impact des amorces sur le contenu sémantique encapsulé par le barycentre. Nous avons fait le choix dans la section 5.2.2 d'intégrer tous les noms déverbaux en *-eur* issus de Lexeur et présents dans le modèle distributionnel. Mais on peut se demander dans quelle mesure la représentation sémantique moyenne d'un groupe varie en fonction des amorces sur lesquelles on construit cette représentation, l'enjeu étant de vérifier la stabilité de la représentation, et sa relative indépendance par rapport à des variations sur les individus utilisés

Pour tenter d'évaluer ce paramètre, nous reprenons l'expérience mais en modifiant certains facteurs. Nous faisons ainsi varier le nombre et les propriétés des amorces, et nous évaluons l'impact de ces facteurs en comparant les barycentres obtenus, toujours sur la base de leurs 100 plus proches voisins. Cette évaluation vise à répondre à deux questions, à savoir (i) dans quelle mesure les voisins varient, et (ii) comment se traduit la variation sur le sens encapsulé par le vecteur moyen, notamment en termes de représentation de l'agentivité et de l'instrumentalité. Quatre facteurs sont ici testés : l'homogénéité sémantique des amorces, le nombre d'amorces utilisées pour construire le barycentre, le possible biais sémantique lié à la sélection des amorces, et la fréquence des amorces.

Notons que les conclusions que nous en tirons n'ont pas valeur de recommandation. Nous ne prétendons pas ici définir des règles quant à la sélection des amorces. Nous souhaitons simplement donner une idée de l'impact de ces paramètres.

Homogénéité sémantique des amorces

Comme nous l'avons vu précédemment, l'analyse du barycentre repose sur l'interprétation de ses voisins. Or, on peut se demander si tout barycentre ne présente pas des voisins interprétables, et par conséquent que l'idée d'une cohérence sémantique du voisinage est illusoire. On peut à ce titre se demander dans quelle mesure le voisinage du barycentre des noms déverbaux en *-eur* traduit réellement une réalité sémantique captée par le barycentre, et dans quelle mesure la représentation est un artefact lié à notre quête d'interprétation.

Pour évaluer cela, nous calculons un barycentre à partir de 100 mots sélectionnés aléatoirement parmi l'ensemble des mots représentés dans l'espace

vectoriel. Ces amorces se révèlent très hétérogènes sur le plan sémantique, et contiennent des noms – d'animaux (*meiobenthos*, *Halitherium*), de plats culinaires (*méchoui*) ou encore d'instruments (*plugiciel*, *steamer*), des adjectifs (*trichromatiques*, *génératif*, *athéologique*), des toponymes (*Munstergeleen*, *Aigny*), des patronymes (*Günther-Frédéric-Charles*, *Kožená*), des acronymes (*KSAN*, *NXIVM*, *ESADMM*) mais aussi des mots d'autres langues que le français (*encuentro*, *heißem*, *difficili*) et des nombres (*12,81*, *635*).

L'analyse des 100 plus proches voisins du barycentre ainsi construit montre que l'on retrouve les mêmes types d'unités, mais dans des proportions différentes. On retrouve ainsi des nombres (*9591*, *16631*) mais à raison de 50% des voisins, alors qu'ils ne représentaient que 7% des amorces. Parmi les autres types de voisins, on retrouve des toponymes – sous leur forme phonologique (*bartɔˈd͡ʒejɛ*, *aˈɲelin*) ou graphique (*Goodsoil*, *Pullinque*), des patronymes (*Heske*), mais aussi des noms de plantes (*×emasculata*, *gossypiifolius*) ou d'animaux (*Acrocinus*, *gambiensis*, *Godartiana*) – principalement sous leur appelation scientifique –, des noms de plats (*Ragda*) et enfin des mots étrangers (*misfatti*, *muistoja*, *angenehmer*). Il ressort de ce voisinage qu'un type sémantique unique n'émerge pas, mais qu'on retrouve au contraire plusieurs types sémantiques, qui étaient représentés dans les amorces. Cette absence de cohérence sémantique semble confirmée par le fait que l'on ne retrouve aucune amorce dans le voisinage. On peut faire l'hypothèse que le barycentre se trouve dans une zone grise de l'espace, regroupant des formes difficiles à classer et sans lien entre elles, et que les voisins rendent compte de cet amalgame de différents vecteurs. On observe la présence notable de nombres parmi les voisins, auxquels il est difficile d'associer un quelconque contenu sémantique. On identifie bien des groupes de sens (animaux, plantes, noms propres), mais ces différents groupes ne forment pas un ensemble cohérent, et ne sont pas liés par une notion ou un concept sémantique partagé. Le barycentre tendrait cependant à mettre en avant certains groupes (ici les nombres), sur des bases qui restent cependant à définir.

Sélection des amorces selon leur terminaison

Nous précisons ce constat en répétant l'expérience précédente avec un barycentre calculé à partir de 100 amorces sélectionnées aléatoirement parmi l'ensemble des formes finissant par la chaîne de caractères *-eur* du modèle. Nous souhaitons montrer que le sens n'émerge pas à partir d'un ensemble de mots qui ne sont pas sémantiquement homogènes au départ, quand bien même ils partagent des caractéristiques formelles. On retrouve parmi ces amorces des noms d'agent (*pâtissier-confiseur*, *soigneur*) et d'instrument (*déboucheur*, *applicateur*) mais aussi d'autres types des toponymes (*Prilly-*

Chasseur, *Yssel-Supérieur*, *Baleur*), des patronymes (*Lechasseur*, *Demeur*), ainsi que des artefacts liés au traitement du corpus (*jpg/fleur*, *[1]Meilleur*), des adjectifs (*ultérieur*), des noms de propriété (*mi-longueur*, *micropensanteur*, *douleur*) et de substances (*sueur*).

L'analyse des 100 plus proches voisins de ce nouveau barycentre fait émerger le même constat que précédemment, à savoir que l'on retrouve globalement les mêmes types sémantiques que ceux des amorces. On observe parmi les voisins des toponymes (*Piez*), des patronymes (*Barroyer*, *Duchezeau*, *Banderet*), des noms d'agents (*garde-caverne*, *marteleur*, *farceur*), d'animés (*glouton*), d'instruments (*ouvre-boîte*, *déchiqueteur*, *michaudine*) de substances (*mexyl*, *antipelliculaire*), et de procédés (*sanforisage*, *microbillage*). Par ailleurs, on ne retrouve qu'une seule amorce parmi les 100 plus proches voisins. Globalement donc, le voisinage forme un ensemble peu cohérent sémantiquent, où plusieurs types sémantiques coexistent. Cela corrobore l'idée d'une absence d'homogénéité sémantique du barycentre lorsque les amorces ne sont elles-mêmes pas sémantiquement homogènes. L'apparition des noms de procédés – absents des amorces – suggère une orientation légèrement plus actionnelle, qui est cohérente avec la représentation des noms d'agent et d'instrument, mais qui reste néanmoins très marginale. Cette composante actionnelle est accompagnée par une prédominance des patronymes (bien plus importante que parmi les amorces), qui va dans le sens d'une spécialisation du barycentre vers des entités humaines, potentiellement définies par des actions. La présence d'un nombre important d'autres sous-groupes montre néanmoins l'hétérogénéité du barycentre.

On note que l'augmentation du nombre d'amorces en *-eur* sélectionnées aléatoirement se traduit par une homogénéité sémantique croissante du voisinage. Ainsi, le passage de 100 à 200 amorces entraîne l'apparition notable d'un très grand nombre de noms d'agent (*plombier*, *machiniste*, *coiffeur*, *magasinier*, *perchman*, *gabier*, *maître-chirurgien*) et de quelques noms d'instrument (*serre-frein*, *arceau-cage*, *jumar*) ou de noms à la double lecture agentive et instrumentale (*aiguiseur*, *mouchard*). Les toponymes (*Frainville*) et patronymes (*Guibot*, *Demas*) deviennent marginaux. Cela s'explique par la présence plus importante de noms d'agent et d'instrument parmi les amorces. On peut ainsi faire l'hypothèse que l'homogénéité du barycentre est liée à celle des amorces d'une classe sémantique donnée. Plus une classe est représentée parmi les amorces, plus le barycentre parvient à capter leur régularité, et plus cette classe s'impose dans le voisinage, réduisant l'importance des formes qui ne font pas sens vis-à-vis de la classe. Il faudrait cependant quantifier précisément la représentation des différentes classes au sein des amorces et des différents voisinages pour conforter cette impression.

Taille de l'effectif

Une première interrogation relative aux amorces concerne leur nombre. Dans l'expérience présentée en section 5.2.2, nous avons choisi d'inclure tous les noms d'agent déverbaux en *-eur* que nous avions à notre disposition dans Lexeur, mais on peut se demander dans quelle mesure les résultats varient si nous construisons ce barycentre à partir d'un nombre plus limité de noms. Nous évaluons cette variation sous un angle quantitatif – combien de voisins changent ? – et qualitatif – les voisins sont-ils toujours majoritairement des noms d'agent et d'instrument ? – en comparant les différents voisinages. Une question sous-jacente est de savoir s'il est nécessaire d'avoir un grand nombre d'amorces pour produire un vecteur moyen représentatif de la catégorie sémantique que l'on vise. Dans notre cas, nous nous demandons si les traits agentifs et instrumentaux restent visibles quel que soit l'effectif. Pour évaluer ce facteur, nous nous proposons de faire varier le nombre n de noms d'agent utilisés comme amorces et de comparer les voisinages ainsi obtenus.

Nous nous servons de la configuration présentée dans la section 5.2.2, où $n = 1675$, comme référence par rapport à laquelle nous allons comparer deux autres configurations. À partir de l'ensemble initial d'amorces S, nous constituons deux sous-échantillons de noms. Le premier, S_{500}, est constitué de 500 amorces sélectionnées aléatoirement parmi l'échantillon S, soit 30% de l'effectif de S. Le second échantillon S_{50} est un sous-échantillon de S_{500} constitué de 50 amorces sélectionnées aléatoirement parmi les amorces de S_{500}, ce qui représente donc 3% de l'effectif de S, et 10% de l'effectif de S_{500}. Nous comparons dans un premier temps les trois voisinages d'un point de vue quantitatif, tant en termes de recouvrement entre amorces et voisins (pour chaque barycentre) qu'en termes de recouvrement entre les différents voisinages.

Notons tout d'abord que le taux de recouvrement entre amorces et voisins diminue à mesure que l'effectif diminue. Ce nombre d'amorces présentes parmi les 100 plus proches voisins du barycentre est de 45% pour l'échantillon S, de 4% lorsqu'on se limite à 500 amorces (échantillon S_{500}), et de 3% lorsqu'on ne prend plus en compte que 50 amorces (échantillon S_{50}). Le nombre d'amorces présentes dans les voisinages est donc nettement plus faible pour les deux sous-échantillons, ce qui indique que peu d'amorces ne sont pas dans le voisinage direct du barycentre. On peut faire l'hypothèse d'une homogénéité moindre de ces amorces dans l'espace. Ces taux sont cependant à nuancer. Ainsi, les 45 amorces que l'on retrouve dans le voisinage représentent 2.7% de l'ensemble des amorces S, alors que ce chiffre est inférieur à 1% dans le cas de S_{500}, mais à 6% de S_{50}.

La comparaison des trois listes de voisins montre par ailleurs d'autres dis-

parités. Ainsi, seuls 16 voisins sont partagés par les trois barycentres (donnés en (9)). Ces 16 voisins communs capturent cependant bien les propriétés sémantiques mises en évidence dans la section 5.2.2, puisque l'on y retrouve des noms d'agent et des noms d'instrument.

(9) *ouvre-boîtes*, *magasinier*, *coiffeur*, *plombier*, *armurier*, *manipulateur*, *artificier*, *confectionneur*, *fabricant*, *garagiste*, *gabier*, *rabatteur*, *polisseur*, *arceau-cage*, *contremaître*, *bottier*

Lorsqu'on compare les voisinages deux à deux, on observe que ceux de S_{50} et S partagent les 16 voisins en (9). Ceux de S_{50} et S_{500} partagent les 16 voisins présentés en (9), plus deux autres noms d'agent, *cuisinier* et *charlatan*, soit 18 voisins au total. Le recouvrement est par contre bien plus conséquent lorsqu'on compare les voisinages des barycentres construits avec S_{500} et S, puisque 71 des 100 plus proches voisins sont identiques. Il semblerait donc que le passage d'un effectif de 1 675 à 500 amorces ait un impact moindre sur la construction du barycentre que le passage à 50 amorces.

Ce constat est cependant à nuancer au regard de l'analyse sémantique des voisinages. Si les voisins distributionnels du barycentre varient en fonction du nombre d'amorces, le contenu sémantique encapsulé par celui-ci ne semble en effet pas vraiment évoluer. Une analyse des listes de voisins pour S, S_{500} et S_{50} montre qu'on obtient pour les trois valeurs de n des noms d'agent, des noms d'instrument, des noms à la double lecture agentive et instrumentale, ou plus globalement des noms d'humain ou d'artefact.

Quelques variations apparaissent cependant localement. Ainsi, on notera la plus grande présence de noms d'agent (*cuistot*, *pickpocket*, *camionneur*) parmi les voisins spécifiques au barycentre construit à partir de S_{500} par rapport au voisinage de S présenté en section 5.2.2, et une présence moindre de noms d'instrument (*appeau*, *tensiomètre*). On note notamment à ce titre l'apparition de *lycan*, qui désigne un groupe ethnique fictif. *A contrario*, une majeure partie des voisins spécifiques au barycentre de S (par rapport à celui construit avec S_{500}) sont des noms d'instrument (*interrupteur*, *extincteur*, *joystick*), et très rarement des noms d'agent (*optométriste*). Un constat similaire peut être fait pour le barycentre construit à partir de S_{50}. Les voisins qui lui sont spécifiques sont principalement des noms d'agent (*mégissier*, *avoué*), et rarement des noms d'instrument (*typomètre*, *fendoir*). On notera que dans ce cas précis, le nom *chien* disparaît, mais que le nom *roncin*, qui partage des caractéristiques similaires notamment sur le plan syntaxique, apparaît.

On observe donc globalement une diminution du nombre de noms d'instrument parmi les 100 plus proches voisins à mesure que le nombre d'amorces diminue. On peut faire l'hypothèse d'une corrélation avec la proportion restante de noms d'agent et d'instrument parmi les amorces, bien que l'on note

la présence non négligeable de noms d'instrument parmi les 50 amorces de S_{50}. Une quantification des noms d'agent et d'instrument parmi les amorces serait nécessaire pour valider cette tendance.

Pour conclure, si les voisins varient lorsque l'on fait varier le nombre d'amorces utilisées pour construire le barycentre, cela ne semble pas avoir un impact trop important sur la représentation sémantique construite, du moins pas dans le cas d'une étude à gros grain comme la nôtre. On observe des disparités quant à la représentation de certains types sémantiques au sein de cette représentation, mais l'ensemble reste néanmoins sémantiquement cohérent, et les résultats pertinents au vu de notre objectif, ici la représentation de la classe des noms en *-eur*. On peut cependant se demander dans quelle mesure les variations observées sont le fait des échantillons sur lesquels elles sont basées. La sélection aléatoire d'amorces ne nous permet pas en l'état de contrôler les proportions de noms d'agent et d'instrument, par exemple. Il faudrait à ce titre reproduire l'expérience un plus grand nombre de fois pour confirmer ces résultats et montrer leur significativité.

Dans la suite des expériences (section 5.3.1), nous fixerons arbitrairement l'effectif des amorces à 50 afin de neutraliser l'impact de ce paramètre, et pour rendre les analyses comparables. Nous nous concentrons désormais sur les propriétés des amorces, et non sur leur nombre.

Biais de sélection

Un second aspect relatif à la selection des amorces concerne l'influence individuelle de chaque amorce dans le contenu sémantique encapsulé par le vecteur moyen. La question se pose tout particulièrement lorsqu'on ne se situe pas dans une approche extensive, mais que l'on part d'un échantillon de taille réduite, ce qui peut induire un biais du fait de cette sélection d'amorces. À ce titre, on peut se demander si la sélection de certains noms d'agent ne donnerait pas trop d'importance aux composants sémantiques présents dans ces amorces. Ainsi, on pourrait imaginer que le contenu sémantique du barycentre soit influencé par la sur-représentation d'un domaine ontologique ou de certaines propriétés axiologiques. Dans quelle mesure la représentation d'une catégorie sémantique donnée varie-t-elle en fonction de ses représentants ? Une hypothèse pourrait être que la classe serait d'autant plus homogène que ce delta est faible. Quelle est par exemple la stabilité de la représentation de l'agentivité et de l'instrumentalité selon les diverses représentations construites à partir des noms d'agent et d'instrument ?

Pour éclaircir ces points, nous définissons trois échantillons de 50 noms d'agent déverbaux en *-eur* dans le but d'évaluer la variabilité de l'instruction sémantique du barycentre. Deux de ces trois échantillons, So_1 et So_2,

sont construits en sélectionnant aléatoirement 50 noms parmi les 1 675 noms d'agent déverbaux en *-eur* de Lexeur présents dans le modèle distributionnel [6]. Le troisième sous-échantillon So_3 est construit de sorte à favoriser un type spécifique de noms d'agent en *-eur*. Nous choisissons arbitrairement le domaine de la religion (*évangélisateur*, *blasphémateur*) et de la spiritualité (*profanateur*, *invocateur*), et sélectionnons manuellement parmi les 1 675 noms d'agent en *-eur* issus de Lexeur et présents dans le modèle distributionnel des noms liés au domaine ontologique ciblé, soit 34 noms. Nous complétons cette liste par 16 noms tirés aléatoirement [7] afin d'atteindre un effectif total de 50 noms. Nous construisons un barycentre pour chaque ensemble d'amorces, et nous en comparons les 100 plus proches voisins.

D'un point de vue quantitatif, seules 3 des amorces de So_1 se retrouvent dans le voisinage du barycentre, contre 5 pour So_2. Le recouvrement entre amorces et voisins concerne 11 noms pour l'échantillon ontologiquement orienté So_3. Cela tend à confirmer ce que nous avons vu avec les barycentres construits à partir d'amorces aléatoires, à savoir que la probabilité de trouver des amorces dans le voisinage d'un barycentre donné augmente avec la cohérence sémantique de celui-ci. La présence d'amorces dans le voisinage peut donc être envisagée comme un indicateur de cohérence sémantique.

Sémantiquement, la spécificité des plus proches voisins du barycentre calculé à partir de l'échantillon ontologiquement orienté So_3 ressort nettement. Tous les voisins sont en effet de près ou de loin liés au domaine du spirituel. Une majeure partie désigne des humains en fonction de leur croyance (*Juif*, *incroyant*), de leur pratique religieuse (*infidèle*, *excommunié*), de leur appartenance à une communauté (*Templier*, *pharisien*), ou encore de leur métier (*hagiographe*, *exorciste*). Quelques rares voisins désignent des entités divines ou mystiques (*Dieu*, *Satan*, *Samaël*, *djinn*). Seuls deux voisins ne désignent pas des humains ou des entités, tout en restant relatif au domaine de la spiritualité, à savoir le nom d'idéalité *hérésie* et le verbe *courroucer*.

Les deux autres voisinages sont bien moins orientés sémantiquement : on retrouve des noms d'agent et des noms d'instrument. Les voisins diffèrent d'un échantillon à l'autre, mais les types sémantiques représentés restent néanmoins stables. Ces deux barycentres aléatoires partagent ainsi 12 voisins, donnés en (10), qui illustrent l'orientation agentive et instrumentale des deux barycentres, ce qui confirme notre description en section 5.2.2. La spécialisation du barycentre orienté So_3 est corroborée par le faible nombre de

6. Nous réutilisons S_{50} en guise de premier échantillon aléatoire So_1.

7. Aucune contrainte n'a été définie pour empêcher la présence d'une même amorce dans les différents échantillons. À ce titre, nous signalons que les deux échantillons aléatoires So_1 et So_2 partagent 5 amorces, et que l'échantillon So_3 partage une amorce avec So_1 et So_2.

voisins partagés avec les barycentres So_1 (neuf voisins, donnés en (11)) et So_2 (aucun voisin partagé).

(10) *armurier*, *contremaître*, *plombier*, *arceau-cage*, *commerçant*, *rabatteur*, *gabier*, *garagiste*, *fabricant*, *confectionneur*, *artificier*, *magasinnier*

(11) *clerc*, *maître*, *serviteur*, *chantre*, *sacristain*, *débauché*, *charlatan*, *mendiant*, *guérisseur*

L'analyse comparative de ces trois listes de voisins suggère donc que la sur-représentation d'une propriété sémantique parmi les amorces (ici le domaine ontologique religieux) amène à une orientation du contenu sémantique encapsulé par le vecteur moyen, à l'image de ce que l'on observait pour l'homogénéité sémantique des amorces. Il faut donc tenir compte de l'influence de ce facteur lors de la sélection des amorces.

Signalons qu'il ne s'agit pas d'un problème. Au contraire, la capacité des modèles distributionnels à capter ces propriétés est un aspect fondamental des modèles distributionnels utilisés comme outil d'exploration linguistique, qui permet de mettre au jour les propriétés sémantiques des classes de mots que l'on étudie. Notre position diffère donc de celle de Bolukbasi *et al.* (2016), où ce type de propriétés est vu comme un biais qu'il s'agit de neutraliser. En tout état de cause, il s'agit d'une caractéristique fondamentale des modèles distributionnels dont il faut avoir conscience.

Fréquence des amorces

Le dernier paramètre relatif aux amorces que nous explorons ici est celui de la fréquence. En effet, les 1 675 noms déverbaux en *-eur* utilisés comme amorces dans la section 5.2.2 présentent des fréquences très variables, comme le montre le tableau 5.2. Nous donnons aussi à titre indicatif la fréquence des amorces féminines.

	minimum	q1	médiane	q3	max	moyenne
-eur	5	15	79	487	325 785	2 548
-euse	5	8	20	52	8 524	115
-rice	5	6	14	46	12 076	533

TABLE 5.2 – Fréquences des noms déverbaux en *-eur*, *-euse* et *-rice* de Lexeur présents dans le modèle distributionnel construit à partir du corpus *Wikipedia2018*

Comme nous l'avons vu dans la section 1.2.2, la fréquence influence directement les représentations vectorielles construites dans les modèles distributionnels. Nous ne cherchons pas ici à définir des seuils de fréquence, mais

à évaluer l'impact de ce facteur sur l'information sémantique encapsulée par le barycentre.

Nous constituons trois listes de 50 amorces tirées aléatoirement. Les deux premiers échantillons Sf_1 et Sf_2 sont constitués de 50 noms tirés aléatoirement, sans contrôle sur la fréquence des amorces. Nous reprenons là aussi les deux listes de 50 amorces So_1 et So_2 utilisées pour l'expérience précédente. La comparaison se fait donc uniquement avec le troisième échantillon, Sf_3, constitué de 50 noms tirés aléatoirement parmi un sous-échantillon de noms déverbaux en *-eur* issus de Lexeur ayant une fréquence comprise entre 100 et 500 (soit un total de 366 noms) afin de garantir un même ordre de grandeur. Nous comparons ensuite les 100 premiers voisins des barycentres construits à partir des trois échantillons.

Notons que l'échantillon contrôlé Sf_3 partage 2 amorces (*batailleur*, *déflecteur*) avec le premier échantillon aléatoire Sf_1, et une seule amorce avec le second échantillon aléatoire Sf_2 (*fileur*). Pour rappel, les deux échantillons aléatoires Sf_1 et Sf_2 partagent 5 amorces. La fréquence des amorces présentes dans les trois échantillons est donnée dans le tableau 5.3.

	min	q1	mediane	q3	max	moyenne	std
Sf_1	5	15	47	239	95 839	2 206	13 529
Sf_2	5	28	139	971	16 436	1 839	4 229
Sf_3	100	152	225	331	482	248	113

TABLE 5.3 – Fréquences des amorces des trois échantillons

Le tableau 5.3 montre de fortes variations entre les différents échantillons. Cette variation s'observe à la fois entre les deux échantillons non contrôlés, Sf_1 offrant une fréquence plus variable que Sf_2, mais aussi vis-à-vis de la liste contrôlée Sf_3.

Sur le plan quantitatif, l'analyse des voisinages montre qu'on retrouve quatre amorces parmi les 100 premiers voisins du barycentre calculé à partir de Sf_3, ce qui est du même ordre que ce que l'on avait observé pour les deux échantillons aléatoires Sf_1 et Sf_2, avec un taux de recouvrement de 3% et 5%. Ce barycentre de Sf_3 partage par ailleurs trois voisins avec le barycentre de Sf_1 (*manipulateur*, *ouvre-boîtes* et *soufflet*), et neuf voisins avec le barycentre de Sf_2. C'est un recouvrement globalement plus faible que celui observé entre les deux échantillons aléatoires, qui était de 12%.

L'analyse de ces voisins partagés, et plus globalement des voisins du barycentre Sf_3, fait émerger des propriétés sémantiques remarquables. On constate en effet que la quasi totalité des voisins – partagés ou non – du barycentre de Sf_3 sont des noms d'instrument (*aspirateur*, *sèche-cheveux*,

poussoir, *ouvre-boîte*) ou d'artefact (*accessoire*, *joystick*, *gadget*). Contrairement à ce que l'on observait pour les deux barycentres calculés à partir des échantillons aléatoires, on ne retrouve aucun nom d'agent monosémique, et très peu de noms à la double lecture agentive et instrumentale (*conducteur*, *nettoyeur*, *manipulateur*). Cet effacement des noms d'agent est propre au barycentre, et pas aux amorces, puisque Sf_3 contient plusieurs noms d'agent (*kidnappeur*, *persécuteur*, *batailleur*, *hypnotiseur*, *noceur*). On peut faire l'hypothèse que les noms d'instrument tendent à être plus fréquents que les noms d'agent, du moins au sein de Sf_3. Une analyse plus fine des amorces et la reprise de l'expérience à partir d'autres échantillons seraient cependant nécessaires pour conforter les résultats et valider cette hypothèse.

Discussion

Il semble évident que les amorces jouent un rôle important dans la construction des barycentres. Ces quelques expériences montrent que la sélection de ces amorces a une influence notable sur le contenu sémantique encapsulé par le barycentre, et sur les voisins que l'on analyse. Globalement, on retrouve des constantes – et notamment dans notre cas les propriétés agentive et instrumentale – mais on constate aussi que la représentation de ces composantes sémantiques varie fortement. La taille de l'effectif et la fréquence des amorces amènent ainsi à de fortes variations des voisins, poussant le barycentre à être tantôt plus agentif, tantôt plus instrumental. La sélection induite par ces deux paramètres a pour conséquence une possible orientation sémantique – qu'elle soit perceptible, comme dans le cas de notre analyse du biais sémantique, ou non – qui induit donc cette variation sémantique du barycentre.

À ce stade, l'impact de ces paramètres a été étudié un à un, mais nous n'avons pas analysé leurs interactions. Par ailleurs, l'influence de certains paramètres n'a pour le moment pas été évaluée, comme celui de la polysémie (agentive et instrumentale, mais pas que), ou de la proportion de noms d'agent et de noms d'instrument au sein des échantillons. Il semble en tout cas à ce stade que la sélection des amorces a un effet sur la représentation. Si cet effet peut être bienvenu, ces analyses plaident néanmoins en faveur d'une approche aussi extensive que possible dans la mesure où nous souhaitons dans ce chapitre analyser globalement les noms d'agent et d'instrument, pour les comparer à l'échelle des catégories qu'ils forment, et que l'augmentation du nombre d'amorces semble compenser – dans une certaine mesure – leur potentielle hétérogénéité.

5.3.2 Choix relatifs au modèles distributionnels

Comme nous venons de le voir, les résultats d'une analyse basée sur les espaces vectoriels dépendent des données lexicales que l'on fournit en entrée, mais aussi évidemment des modèles distributionnels utilisés. Or, ces modèles distributionnels présentent une certaine instabilité du fait des méthodes stochastiques impliquées dans l'entraînement non supervisé (Antoniak et Mimno 2018, Pierrejean 2020). Certaines étapes de l'entraînement font intervenir des traitements aléatoires qui induisent une variation de la disposition des espaces distributionnels. De fait, la localisation des vecteurs les uns par rapport aux autres fluctue, ce qui se traduit par une variation de leur proximité. Cela fait aussi varier les plus proches voisins d'un mot en fonction des modèles. On peut donc se demander comment cette instabilité des voisinages se traduit sur l'analyse des barycentres et de leurs voisinages.

Une solution préconisée par Antoniak et Mimno (2018) est de reproduire l'expérience que l'on mène sur m modèles distincts entraînés selon les mêmes paramètres, et de tirer des conclusions sur la base des $m \times 100$ voisins obtenus (Antoniak et Mimno 2018). Cela est aisément opérationnalisable pour des questions de recherche impliquant des résultats quantitatifs (et notamment pour l'expérience présentée dans la partie II), où l'on peut aisément faire la moyenne de scores de proximité. Cette option n'est cependant pas envisageable dans le cas d'une analyse qualitative sémantique, qui ne peut pas être soumise au moyennage, ou qui impliquerait de reproduire m fois l'annotation des voisins.

Puisqu'il n'est pas possible d'identifier un modèle comme étant plus représentatif qu'un autre, il s'agit de trouver une méthode permettant de moyenner les résultats mais qui évite la redondance d'un post-traitement linguistique manuel. Nous proposons dans la suite de cette section trois pistes pour répondre à cette problématique, basées respectivement sur (i) le moyennage des scores cosinus (Cos) sur plusieurs modèles, (ii) la construction d'un modèle unique à partir de la moyenne des vecteurs de plusieurs modèles (Ave), et (iii) la construction d'un modèle unique à partir de la concaténation des vecteurs de plusieurs modèles (Concat). Pour chaque méthode proposée, nous en comparons les résultats avec ceux obtenus pour la configuration initiale présentée dans la section 5.2.2 sur la base d'un unique modèle (désormais Dsm_1). Puisqu'il s'agit d'étudier l'impact de ces configurations sur la stabilité de résultats, nous les confrontons à un second modèle distributionnel (Dsm_2) calculé selon les mêmes paramètres que Dsm_1. Nous évaluons les méthodes en comparant les listes des 100 plus proches voisins du barycentre des noms déverbaux en *-eur* en observant plus précisément leur variation, tant en nombre qu'en types sémantiques. Nous fixons de façon arbitraire à 5 le

nombre de modèles distributionnels agrégés.

Moyenne des scores cosinus (Cos)

Le premier niveau de méthodes que nous testons implique simplement un moyennage. L'idée est que du fait de la variabilité des modèles distributionnels, certains voisins idiosyncratiques apparaissent parmi les 100 premiers voisins d'un barycentre dans un modèle donné de façon exceptionnelle. Cela signifie qu'à l'inverse les autres voisins forment le noyau du barycentre des mots de la classe. Notre objectif est d'identifier ce noyau, dont nous postulons désormais qu'il est constitué des 100 voisins « moyens »– sur les cinq modèles distributionnels – du barycentre.

Nous nous basons sur les scores de proximité moyens pour identifier les 100 plus proches voisins communs aux cinq modèles. Nous faisons l'hypothèse qu'un voisin idiosyncratique aura un score de proximité moyen plus faible que les autres voisins puisqu'il sera en général plus distant du barycentre, et qu'il n'apparaîtra donc pas parmi les 100 plus proches voisins communs. À l'inverse, les voisins qui se trouvent dans le voisinage proche du barycentre dans les cinq modèles auront un score de proximité moyen plus élevé.

Nous opérationnalisons cela en construisant le barycentre des noms déverbaux en *-eur* (voir section 5.2.2) dans cinq modèles différents, et nous calculons pour l'ensemble des mots de chaque modèle leur proximité au barycentre. Les cinq modèles partageant le même vocabulaire, nous pouvons calculer le score de proximité moyen de chaque mot aux barycentres. Nous extrayons alors les 100 mots dont le score de proximité moyen aux barycentres est le plus élevé (que nous appellerons désormais *voisins communs*). Nous formalisons cela selon la formule 5.2, où u_n $(1 \leq n \leq 5)$ est un mot du vocabulaire V commun aux modèles D_j $(1 \leq j \leq 5)$, $P_{Av}(u_n)$ le score de proximité moyen de u_n sur les cinq modèles et $P_j(u_n)$ le score de proximité de u_n au barycentre $\overrightarrow{C_j}$ dans le modèle D_j. Nous analysons ces voisins communs au regard des résultats obtenus dans la section 5.2.2 dans le DSM_1.

$$\begin{aligned} V &= \{u_1, ..., u_n\} \\ P_{Av}(u_n) &= \frac{1}{5} \sum_{j=1}^{5} P_j(u_n) \end{aligned} \tag{5.2}$$

Les 100 plus proches voisins obtenus avec cette méthode partagent 87 noms en commun avec les voisins calculés dans (DSM_1). Le cœur de la représentation vectorielle des noms déverbaux en *-eur* semble donc relativement stable. Cela semble corroboré par le fait que les voisins spécifiques à la version moyennée n'apparaissent pas avant le 58$^{\text{e}}$ rang (*charriot*).

Réciproquement, parmi les 13 voisins qui ne sont pas retrouvés dans le voisinage moyenné, on notera le nom *double-saut*, ce qui semble confirmer son caractère idiosyncratique. Le nom *chien* par contre se maintient. D'un point de vue sémantique, les deux listes de voisins sont globalement très similaires. On peut juste souligner que la variation touche légèrement plus les noms d'instrument que les noms d'agent. Ainsi, des noms comme *injecteur*, *poussoir*, *ouvre-boîte*, ou *tourne-à-gauche* laissent la place dans ce nouveau voisinage à d'autres noms d'instrument comme *charriot*, *boulon*, *cliquet*, *tensiomètre*, *treuil*, ou *électroaimants*, sans réelle incidence sur la représentation de l'agentivité décrite précédemment.

La comparaison de cette liste moyennée de voisins avec la liste obtenue sur le second modèle distributionnel (DSM_2) confirme ces résultats. On observe ainsi à nouveau une variation de 13 voisins. À titre de comparaison, les barycentres calculés dans DSM_1 et DSM_2 présentent 18 voisins différents. Dans tous les cas, les voisinages étudiés captent la même information sémantique, à savoir l'agentivité et l'instrumentalité.

Ces résultats suggèrent que la méthode Cos permet est une bonne moyenne des résultats des modèles distributionnels, dans la mesure où les résultats diffèrent de façon similaire avec les deux modèles, et qu'elle permet de stabiliser la représentation, les différences obtenues vis-à-vis des modèles étant moins importantes qu'entre les modèles individuels eux-mêmes. Cette première méthode considère le noyau du barycentre dans une approche quantitative, puisque son identification repose sur une moyenne – numérique par nature – des scores de proximité. Nous proposons ci-après deux méthodes plus qualitatives dans la mesure où le moyennage est appliqué en amont de la sélection des voisins.

Moyenne des modèles (Ave)

La deuxième méthode que nous nous proposons de tester repose sur le moyennage des modèles distributionnels eux-mêmes et non sur le moyennage des distances. L'idée est ici de proposer un modèle unique construit à partir de cinq modèles que l'on entraîne séparément. Le modèle distributionnel unique est obtenu en faisant la moyenne des vecteurs qui composent les cinq modèles. Ainsi, les coordonnées de chaque vecteur sont les moyennes de ses cordonnées dans les cinq modèles distributionnels. On se retrouve alors avec un modèle distributionnel abstrait de 100 dimensions, qui n'a pas été directement calculé à partir du corpus. Comme précédemment, nous utilisons ce modèle Ave pour calculer le barycentre des noms déverbaux en *-eur* à partir des 1 675 amorces disponibles, et pour observer les 100 plus proches voisins.

Malgré la similarité de la représentation sémantique proposée, à savoir

principalement des noms d'agent et des noms d'instrument, la comparaison des voisinages obtenus dans les deux modèles DSM_1 et DSM_2 avec celui obtenu dans le modèle AVE montre de fortes disparités. En effet, respectivement 27 et 35 voisins sur 100 diffèrent dans les deux modèles (contre 13 pour COS). On constate donc que l'utilisation d'un modèle moyenné augmente les différences par rapport à ce que l'on a entre DSM_1 et DSM_2 (différence de 18).

Le moyennage des modèles ne fournit donc pas une bonne généralisation des résultats. Les deux voisinages sont cependant comparables en termes de types sémantiques, puisque l'on retrouve dans tous les cas des noms d'agent et des noms d'instrument.

Au vu de ces résultats, l'utilisation de la moyenne de plusieurs modèles distributionnels ne présente pas d'intérêt. Cela peut s'expliquer par la méthode elle-même, qui rend le modèle distributionnel ainsi créé décorrélé du corpus. L'espace formé par la moyenne des vecteurs ne permet pas de conserver une géométrie pertinente. La méthode ne semble pas cependant produire de résultats complètement aberrants dans la mesure où les voisinages étudiés indiquaient tout de même que les barycentres encapsulaient les traits sémantiques ciblés. Le type instrumental semble néanmoins privilégié par rapport au type agentif. Tout cela suggère que ce n'est pas une méthode pertinente pour moyenner les résultats lorsque comparés aux DSM_1 et DSM_2. Cette méthode a par ailleurs un certain coût computationnel puisqu'elle implique de devoir manipuler des modèles distributionnels de façon non triviale.

Concaténation des modèles (CONCAT)

La dernière méthode que nous testons est similaire dans son intention, à savoir construire un modèle unique sur lequel expérimenter, mais diffère dans son opérationnalisation. Nous proposons ici de concaténer les vecteurs et leurs coordonnées au lieu de les moyenner. Cela revient à considérer chaque espace DSM_j ($1 \leq j \leq 5$) comme un sous-espace de l'espace vectoriel $\text{DSM}_{\text{CONCAT}}$. Nous créons donc à partir de cinq modèles à 100 dimensions un modèle distributionnel à 500 dimensions. Chaque plongement lexical voit ainsi ses coordonnées dans les cinq modèles différents concaténées en un unique vecteur. Une fois le modèle constitué, on peut calculer le barycentre des noms déverbaux en *-eur*, et en analyser ses 100 premiers voisins.

On constate tout de suite que l'ensemble du voisinage ainsi obtenu est identique à celui obtenu avec la méthode COS. On retrouve les mêmes 100 premiers voisins, présentés quasiment dans le même ordre. Seuls 11 voisins voient leur rang interverti avec un ou plusieurs autres voisins. On a donc ici la même information sémantique que celle encapsulée par les cinq barycentres distincts pour lesquels on récupère les 100 plus proches voisins communs.

On observe à ce titre les mêmes comportements vis à vis des voisinages des barycentres dans les deux modèles distributionnels DSM_1 et DSM_2. Ce sont ainsi à chaque fois 13 voisins qui diffèrent, et 27 voisins qui diffèrent avec le modèle AVE.

Cette méthode basée sur la concaténation des modèles distributionnels donne les mêmes résultats que celle basée sur la moyenne des scores de proximité COS, et semble à ce titre aussi pertinente. Elle a le mérite de produire un modèle unique et évite de reproduire plusieurs fois les mêmes traitements. Cette méthode implique cependant un peu de plus de calculs supplémentaires en amont du calcul des barycentres, et produit un modèle plus lourd, mais qui reste gérable car les modèles initiaux sont denses.

Discussion

Malgré les variations ponctuelles observées, l'information sémantique encapsulée par le barycentre des noms déverbaux en *-eur* semble stable, et ce quelles que soient les configurations testées. Les voisins eux-même tendent à varier davantage lorsque l'on sélectionne un nombre limité d'amorces, ou lorsque celles-ci présentent des spécificités sémantiques fortes. L'agentivité et l'instrumentalité restent cependant, tout au long de cette expérience, les traits sémantiques que l'on retrouve de façon explicite. Cela conforte à la fois la méthode des barycentres et l'hypothèse que cette catégorie sémantique est localisée dans une région spécifique dans les modèles distributionnels.

Nous avons par ailleurs pu mettre à l'épreuve plusieurs méthodes pour atténuer l'instabilité inhérente des modèles distributionnels, et avons retenu deux méthodes computationnellement distinctes, pouvant s'adapter à différents dispositifs expérimentaux, mais dont les résultats sont similaires. Au vu des résultats, nous faisons le choix dans la suite de ce travail (sauf exceptions discutées lorsque nécessaire) de ne pas opérer de sous-échantillonnage des amorces, et d'utiliser le moyennage des scores cosinus (chapitres 6, 7, 8 et 9), ou un modèle concaténé (chapitres 10 et 12).

Dans la suite de cette étude, nous appliquons la méthode développée dans ce chapitre à la comparaison des noms d'agent déverbaux féminins en *-euse* et *-rice*.

Chapitre 6

Différenciation sémantique des noms en *-euse* et *-rice*

L'étude des noms d'agent déverbaux féminins en *-euse* et *-rice* vise à explorer deux hypothèses principales : d'une part que l'agentivité ne s'exprime pas de la même façon pour les noms déverbaux masculins en *-eur* et pour les noms déverbaux en *-euse* et *-rice*, et d'autre part que les deux suffixes féminins forment des noms aux propriétés sémantiques distinctes.

Nous venons de voir que les traits agentif et instrumental partagés par les noms déverbaux en *-eur* émergeaient dans la distribution moyenne d'un large échantillon de noms déverbaux en *-eur*. Nous faisons l'hypothèse que ces composantes sémantiques devraient être aussi saillantes dans la distribution moyenne des noms en *-euse* et *-rice*. Nous montrons cependant que l'agentivité telle qu'exprimée par les barycentres des noms féminins diffère de celle des noms masculins dans la mesure où elle met l'accent sur le genre référentiel des agents dénotés. Nous avons par ailleurs vu dans la section 5.1.2 que les deux suffixes affichaient localement des propriétés sémantiques distinctes, notamment en lien avec la connotation associée aux agents dénotés par les noms qu'ils construisent. Nous pouvons ainsi faire l'hypothèse que la dimension agentive ne va pas se traduire de façon similaire dans la distribution des noms en *-euse* et *-rice*, et que leurs barycentres vont définir deux profils distributionnels distincts. L'examen de leurs voisinages montre ainsi une différence de connotation, le barycentre des noms en *-euse* intégrant des traits axiologiques péjoratifs et sexuels absents du barycentre des noms en *-rice*. Nous montrons par ailleurs l'impact de la lexicalisation sur le sens capté par ces barycentres, et notamment la disparition de la connotation, au profit d'une spécialisation ontologique des voisins.

Ce chapitre explore ces deux hypothèses. La section 6.1 présente la comparaison des noms en *-euse* et *-rice* au travers de l'analyse des voisinages de

leurs barycentres. Nous évaluons dans la section 6.2 l'impact de la lexicalisation sur les spécificités que nous faisons émerger pour chaque classe de noms. Nous questionnons dans la section 6.3 les résultats obtenus au regard de la lemmatisation problématique des noms féminins, et nous concluons dans la section 6.4 en évoquant d'autres approches méthodologiques de la différentiation des noms en *-euse* et *-rice* dans les espaces vectoriels.

6.1 Une représentation inégale

Nous analysons dans les deux sections qui suivent les 100 premiers voisins moyens des barycentres construits à partir des noms déverbaux féminins en *-euse* et *-rice*. Ces analyses reposent sur la comparaison des voisinages des barycentres des noms en *-euse* et *-rice* avec les voisins du barycentre des noms déverbaux en *-eur* d'une part, et sur la comparaison des voisinages féminins entre eux d'autre part.

Puisque nous sommes dans la même démarche expérimentale de représentation unifiée d'un groupe de mots donnés, nous reprenons la méthode présentée dans la section 5.2.2. Nous construisons donc un barycentre à partir des vecteurs des noms déverbaux en *-euse* d'une part, et de ceux des noms déverbaux en *-rice* d'autre part. Pour rendre les comparaisons pertinentes, nous basons notre analyse sur le même corpus que pour les expériences précédentes, c'est-à-dire le corpus *Wikipedia2018*. Pour calculer les représentations vectorielles, nous utilisons les mêmes paramètres par défaut que précédemment (voir sections 3.2.2 et 5.2.2). Cependant, sur la base des explorations méthodologiques de la section 5.3, nous choisissons de ne pas utiliser un seul modèle distributionnel, mais de travailler à partir de plusieurs modèles distributionnels afin de garantir la stabilité des résultats. Plus précisément, nous choisissons d'appliquer la méthode Cos présentée dans la section 5.3.2, qui consiste à construire les barycentres – un par suffixe – dans cinq modèles distincts, et à sélectionner les 100 plus proches voisins communs aux cinq modèles sur la base de leur score de proximité moyen. Nous réutilisons pour cela les cinq modèles construits pour les expériences présentées en section 5.3.2. Les amorces sont sélectionnées parmi les noms féminins présents dans les familles déverbales de Lexeur. Nous extrayons un total de 3 476 noms en *-euse* et 1 145 noms en *-rice* distincts dans Lexeur. De ces 4 621 noms féminins, seuls 379 disposent d'un vecteur dans le modèle, dont 302 noms en *-euse* et 77 noms en *-rice*.

Nous présentons dans un premier l'analyse des 100 plus proches voisins communs des barycentres construits à partir des noms en *-euse* (section 6.1.1), puis celle pour les noms en *-rice* (section **??**).

6.1.1 Suffixe *-euse*, le féminin sous toutes ses facettes

Parmi les 302 amorces utilisées pour construire le barycentre des noms d'agent en *-euse*, 22 se retrouvent dans le voisinage formé par les 100 voisins communs, soit 7.3% des amorces. Il s'agit à la fois de noms d'agent (*serveuse*, *boxeuse*), de noms d'instrument (*tondeuse*, *essoreuse*, *dameuse*), et de noms à double lecture agentive et instrumentale (*coiffeuse*, *chauffeuse*). Ce recouvrement est bien plus limité que dans le cas des noms d'agent en *-eur*, où le taux atteignait 45%. Ce résultat doit cependant être mise en regard du nombre initial d'amorces, puisque l'échantillon est ici relativement limité (voir section 5.3.1), ainsi qu'au regard de l'effet de la lemmatisation. Au regard du taux obtenu par le barycentre constitué à partir de 500 amorces en *-eur* qui était de 4%, on peut même considérer que ce taux de 7,3% est élevé. Cela semble suggérer une homogénéité distributionnelle relativement faible des noms en *-euse*, qui se confirme avec l'analyse sémantique des voisins.

À ces 22 noms déverbaux en *-euse* s'ajoutent par ailleurs cinq autres voisins suffixés en *-euse*, à base nominale notamment (*stripteaseuse*) ainsi que d'autres types morphologiques. Il y a ainsi des noms suffixés en *-ière* (*cafetière*), en *-oire* (*rôtissoire*), *-iste* (*modiste*), *-arde* (*fêtarde*), mais aussi des converts (*râpe*), des noms simples (*gitane*, *poupée*) et des emprunts (*covergirl*, *chapka*). Notons que contrairement au voisinage du barycentre des noms d'agent en *-eur*, tous les voisins ne sont pas des noms communs (*Tinnie*, *prénomée*).

Sur le plan sémantique, on constate que le voisinage des noms féminins en *-euse* diffère fortement de celui des noms masculins en *-eur*. On remarquera tout d'abord que ces deux voisinages ne partagent aucun voisin. Le voisinage des noms en *-euse* contient les mêmes types sémantiques que celui des noms en *-eur*, notamment des noms d'agent (*serveuse*, *pêcheuse*), des noms d'instrument (*essoreuse*, *mortaiseuse*), et des noms à double lecture agentive et instrumentale (*cafetière*, *perceuse*). On constate que la distribution de ces catégories n'est pas la même. En effet, les voisins non agentifs et non instrumentaux sont très marginaux autour du barycentre des noms déverbaux masculins, mais très fortement représentés pour *-euse*. Plus d'un tiers des voisins ne relèvent pas de ces catégories. Les voisins étant ici considérés comme des indicateurs des propriétés sémantiques du barycentre, les composantes agentives et instrumentales sont donc bien moins saillantes dans le voisinage des noms en *-euse* que pour leurs équivalents masculins. On peut faire l'hypothèse que l'effacement de ces deux traits dans la représentation du barycentre se fait au profit d'un autre trait sémantique. On peut aussi se demander si ce trait supplémentaire est uniquement lié au genre féminin, ou s'il contient en plus une spécificité qui serait associée à la suffixation en

-euse.

Un premier élément de réponse est apporté par l'observation des voisins ne correspondant pas à des noms d'agent ou d'instrument, puisqu'on y retrouve certaines régularités. On a ainsi un certain nombre de noms d'animaux (*chatte*, *hérissonne*, *ponette*, *tigresse*, *cochonne*) désignant spécifiquement la femelle de l'espèce concerné (à l'exception de *crevette* dont le genre grammatical est intrinsèquement féminin). On retrouve aussi des noms de vêtement et accessoires (*jupe-culotte*, *salopette*, *doudounne*, *gourmette*), ou d'aliments (*tartelette*). Enfin, certains voisins désignent des humains ne correspondant pas à des noms d'agent (*midinette*, *mémère*).

Globalement, les voisins dénotant des humains ou des agents désignent majoritairement, voire exclusivement, des référents féminins. Ainsi, la plupart des professions que l'on retrouve dans le voisinage relèvent stéréotypiquement de la sphère féminine (*coiffeuse*, *serveuse*, *ballerine*). Dans tous les cas, le genre féminin du référent est mis en lumière, comme dans le cas de *snowboardeuse* ou *hackeuse*, qui ne sont pas des métiers prototypiquement occupés par des femmes. On peut faire l'hypothèse que la présence de la variante féminine peut être liée à l'influence du genre grammatical évoqué plus tôt. Cependant, on constate que même lorsqu'il n'y a pas d'indice formel quant au genre référentiel, comme dans le cas de *modiste*, *fleuriste*, *standardiste*, *manucure* ou *dactylo* (qui peuvent dénoter des hommes ou des femmes lorsqu'il s'agit du nom d'agent), il existe des attentes socioculturelles quant au genre de l'humain dénoté, en l'occurrence féminin.

Notons de plus qu'une partie de ces noms se caractérisent par une valeur axiologique négative, principalement associée à des connotations peu valorisantes ou vulgaires : *chatte* et *doudoune* sont des termes d'argot qui peuvent désigner des attributs anatomiques féminins, *tigresse* et *cochonne* désignent des femmes sur la base de leur comportement sexuel, et *bimbo*, *brune* et *jolie* font référence aux femmes sur la base de leur apparence. Cette connotation s'étend même aux noms d'agent présents dans le voisinage, dont certains sont également fortement connotés négativement (*stripteaseuse*, *call-girl*, *cover-girl*).

Le trait du féminin associé à la suffixation en *-euse* se retrouve par le biais de la connotation, la référence, mais aussi au travers du genre grammatical. Il est ainsi intéressant de noter qu'une partie des voisins sont féminins uniquement sur le plan grammatical. C'est notamment le cas pour l'ensemble des noms d'instrument présents dans le voisinage (*tondeuse*, *dameuse*), dont le genre est intrinsèque féminin. Le rapprochement n'est donc pas tellement lié à des considérations sémantiques. On pourrait arguer, pour certains de ces voisins, que leur appartenance à des domaines typiquement occupés par la femme (du moins, au quotidien), tels que la cuisine (*râpe*, *cafetière*, *bour-*

riche) ou la mode et la beauté (*bigoudi*, *guêpière*) justifie leur association à un barycentre captant l'information sémantique du féminin. Cette explication trouve cependant très rapidement ses limites dès lors que l'on considère des voisins relevant de ces domaines mais n'étant pas particulièrement associées à la femme (*gourmette*, *parka*), voire qui tendent à être associés à des activités relevant stéréotypiquement de la sphère masculine (*tondeuse*, *dameuse*, *batteuse*). Si la présence de ces voisins permet de mettre l'accent sur la dimension instrumentale historique du suffixe *-euse* (voir section 5.1.1), cela suggère aussi et surtout que le rapprochement est lié au genre grammatical, et de fait aux suffixes qu'ils exhibent. En effet, le point commun d'une grande partie de ces mots est d'être construits à l'aide des suffixes *-euse* et *-ette*, principalement.

Il ressort de ce voisinage que l'agentivité construite sur le plan distributionnel par les noms déverbaux féminins en *-euse* diffère singulièrement de celle construite par les noms d'agent masculins. Le barycentre des noms en *-euse* montre ainsi une plus forte hétérogénéité qui semble s'expliquer par les multiples facettes sémantiques associées à cette suffixation. D'une certaine façon, le féminin est retranscrit ici sous les trois aspects du genres décrits dans la littérature (Schafroth 2001) que sont le genre référentiel, social et grammatical (section 5.1). Si les noms en *-euse* tendent à désigner leur référent sur la base du genre féminin, ils font aussi ressortir les sens connotés et d'autres projections socio-culturelles. Cela tend à montrer que les agents féminins sont décrits par rapport à leur corps et comportement, alors que les agents masculins sont utilisés de façon plus neutre pour décrire des professions ou des statuts. Enfin, il est indéniable que le genre grammatical influence particulièrement cette représentation. Un examen plus approfondi de l'impact de la lemmatisation sur la représentation de la classe des noms en *-euse* reste à mener.

On peut en tout cas se demander par contraste si l'agentivité s'exprime par le biais de ce même triptyque pour les noms d'agent féminins en *-rice*, ou s'il existe des traits spécifiques à *-rice*.

6.1.2 Suffixe *-rice*, l'agent au féminin ?

Comme précédemment, nous calculons dans chacun des 5 modèles distributionnels le barycentre des 77 noms déverbaux en *-rice*, et analysons les 100 plus proches voisins dont le score de proximité moyen au barycentre est le plus élevé.

Notons que seules 8 amorces (10%) se retrouvent dans le voisinage. Ce taux de recouvrement est dans l'absolu plus faible que le recouvrement observé pour les noms déverbaux en *-euse*, mais les amorces sont proportionnel-

lement davantage représentées. Il est par ailleurs du même ordre (voire plus élevé) que celui observé en section 5.3.1 sur un effectif d'amorces similaire, où le taux était de 3% pour un effectif de 50 amorces. On peut donc supposer qu'il existe une certaine homogénéité sémantique des noms déverbaux en *-rice*, qui va cependant au-delà du critère formel qu'est le suffixe *-rice*. Notre analyse sémantique confirme partiellement cette hypothèse.

Au total, le suffixe *-rice* est représenté dans 17 voisins. Il est notamment présent dans des noms complexes non déverbaux (*médiatrice*). On trouve aussi parmi les voisins des noms suffixés en *-euse* (*régisseuse*), *-eure* (*pasteure*), *-ienne* (*plasticienne*), *-ière* (*parolière*), ou encore *-iste* (*modiste*). Enfin, on retrouve des noms composés (*auteure-compositrice*), des noms simples (*agente*, *cheffe*) et des emprunts (*scripte*). On remarque par ailleurs, à l'image du barycentre des noms en *-euse*, qu'une partie des voisins ne sont pas des noms communs (*R-173*, *Réso-Liain*, *intrafusales*, $\Delta 3$).

La présence de ces voisins qui semblent particulièrement idiosyncratiques soulève la question de la fréquence. Comme évoqué en section 1.2.2, un nombre trop limité d'occurrences ne permet pas de garantir une bonne qualité de la représentation vectorielle, puisqu'il y a peu de contextes sur lequel l'algorithme peut s'entraîner. On se retrouve ainsi dans des zones de l'espace vectoriel sans réelle homogénéité sémantique, qui contiennent des vecteurs pour lesquels les contextes sont trop peu nombreux et informatifs pour permettre une représentation pertinente, et où les rapprochements sont idiosyncratiques. Nous avons fait le choix de fixer à 5 le seuil de fréquence, ce qui est relativement faible, mais nous permet d'utiliser un plus grand nombre d'amorces. On peut cependant faire l'hypothèse que le barycentre des noms en *-rice*, (et dans une même mesure, des noms d'agent en *-euse*) se trouve dans une zone moins bien dense et sémantiquement moins cohérente que celle du barycentre des noms en *-eur*, ce qui expliquerait la forte présence de ce type de voisins.

Sémantiquement, la représentation de l'agentivité et de l'instrumentalité semble là aussi fortement influencée par le genre. Ainsi, comme précédemment, on constate une absence de recouvrement des voisinages des noms en *-rice* et en *-eur*. L'agentivité ne s'exprime donc pas de la même façon pour les noms en *-rice* et en *-eur*, et on peut faire l'hypothèse que cet écart est notamment dû au genre référentiel. Si les voisins diffèrent complètement, on retrouve cependant les mêmes types sémantiques, à savoir des noms d'agent (*cofondatrice*, *écrivaine*), des noms d'instrument (*grenailleuse*, *thermopile*), des noms à double lecture agentive et instrumentale (*coiffeuse*, *imprimeuse*) et des voisins n'appartenant pas aux précédents types.

À ce stade, nous pouvons faire deux constats. Tout d'abord, les noms d'instrument et les noms à double lecture sont bien moins représentés que

pour les barycentres des noms d'agent en *-eur* et *-euse*, au profit des noms dénotant des agents. L'agentivité est donc à ce titre particulièrement saillante, bien plus que pour *-euse*. À l'inverse, l'instrumentalité est quasiment absente de ce barycentre. Par ailleurs, on note que presque la moitié des voisins ne sont pas des noms d'agent ou d'instrument. Ils sont aussi nombreux que ceux du barycentre des noms en *-euse*, mais diffèrent quant à leurs propriétés sémantiques. On ne retrouve ainsi aucun objet, et très peu de noms d'humain (à l'exception de *petite-nièce* et, possiblement, *canadienne*). On note cependant la forte présence de mots issus du domaine scientifique et médical comme des maladies (*tétraparésie*, *gliose*, *acanthose*), des items liés à l'anatomie (*synostose*, *intrafusales*, *paracrine*), et des molécules et substances (*flavoprotéine*, *neurohormone*, *thyronamine*). Si le degré d'agentivité des noms de molécules et de substances est débattu dans la littérature, notamment du fait des traits d'animéité et d'intentionnalité, ces voisins renforcent néanmoins la tendance à l'agentivité du barycentre.

Nous nous arrêtons un instant sur cette forte présence de termes ontologiquement liés à la médecine, pour laquelle nous pouvons envisager trois pistes, mais que nous n'explorons pas dans ce travail. Une première hypothèse permettant d'expliquer cela serait que ce domaine serait intrinsèquement surreprésenté parmi les amorces, comme nous l'avons vu dans la section 5.3.1. Une analyse des 77 amorces ne permet cependant pas de faire ressortir une quelconque orientation sémantique liée à la médecine. Une seconde explication pourrait être apportée par la question des fréquences. On peut faire l'hypothèse qu'à l'image des nombres pour le barycentre des mots aléatoires présenté en section 5.3.1, les voisins relatifs à la médecine constituent une des classes présentes dans la zone grise où se trouvent les noms en *-rice*, globalement peu fréquents. Enfin, on peut se demander dans quelle mesure la préférence de sélection du suffixe *-rice* pour des bases savantes, très présentes dans le domaine de la médecine, a un impact sur ce rapprochement observé dans les modèles distributionnels.

Si l'on se penche un peu plus précisément sur les voisins animés, on constate qu'ils désignent principalement des référents féminins, comme pour le barycentre des noms en *-euse*. On remarque cependant que ni les noms d'agent ni les rares noms d'humain ne font référence à l'apparence physique ou au comportement du référent. La seule exception concerne le voisin *transsexuelle*, qui est cependant dépourvu de connotation péjorative. La connotation de ces voisins est positive ou neutre, puisque les métiers ou référents dénotés par les noms correspondent très majoritairement à des positions au moins neutres (*rédactrice*, *traductrice*), sinon valorisées socio-culturellement (*directrice*, *ingénieure*, *cheffe*). La forte orientation axiologique repérée pour *-euse* se manifeste ici plus discrètement, et est orientée dans le sens positif.

Il est intéressant de noter que l'on trouve quelques voisins grammaticalement non marqués du point de vue du genre (*scripte*) ou masculins, comme *costumier* ou *esthéticien*. Leur analyse en corpus montre qu'ils apparaissent dans deux situations distinctes. Dans un cas, le nom d'agent masculin est en réalité un nom d'agent féminin lemmatisé comme masculin, comme nous l'évoquons dans la section 6.3. Dans le second cas, le nom d'agent est effectivement masculin, et son référent est un homme. Dans le cas de *esthéticien*, notons que le féminin *esthéticienne* est utilisé principalement pour désigner la femme qui travaille dans un institut de beauté, alors que la forme masculine désigne surtout (mais pas exclusivement) un théoricien d'art. Le fait que l'équivalent masculin plus valorisé se trouve dans le voisinage du barycentre des noms en *-rice* mais pas dans celui de *-euse* accentue la différence de connotation observée entre les deux suffixes *-euse* et *-rice*, à savoir que les noms animés sont davantage valorisés sur un plan socioculturel pour le barycentre des noms en *-rice* que pour le barycentre des noms en *-euse*. Il faudrait cependant évaluer pour le vecteur $\overrightarrow{esthéticien}$ dans quelle mesure les contextes de la forme masculine et de la forme féminine influencent sa représentation vectorielle afin de confirmer ou non l'explication de la présence de ce voisin parmi les plus proches voisins du barycentre des noms d'agent en *-rice*.

Les constats que nous venons de dresser suggèrent que les barycentres des noms en *-euse* et *-rice* n'ont pas les mêmes propriétés sémantiques. Cela est confirmé par le faible recouvrement des deux voisinages, puisque seuls 11 voisins sont partagés. Ces voisins sont donnés en (12). Ce sont tous des noms d'agent (*Youtubeuse*) ou des noms disposant d'une lecture agentive (*manucure* peut aussi dénoter le soin). La plupart de ces noms réfèrent sans ambiguité à un agent féminin, à l'exception de *manucure*, *modiste* et *standardiste*, qui sont neutres sur le plan grammatical et référentiel, mais qui sont fortement féminins sur le plan socio-culturel. Le contenu sémantique partagé par les deux barycentres semble donc inclure l'agentivité et le genre référentiel féminin. L'instrumentalité y est absente.

(12) *coiffeuse*, *manucure*, *ballerine*, *barmaid*, *modiste*, *standardiste*, *Youtubeuse*, *dactylo*, *call-girl*, *snowboardeuse*, *hackeuse*

Si nous avons fait le constat empirique et systématique d'une différence de représentation des noms en *-eur*, *-euse* et *-rice* dans l'espace vectoriel considéré, on peut se demander dans quelle mesure la nature du corpus a un impact sur le résultat. En effet, il a été montré que le lexique utilisé dans les pages Wikipedia dédiées aux femmes insistait davantage que dans celles dédiées aux hommes sur le statut familial, relationnel et social (Wagner *et al.* 2015). Il serait donc intéressant de reprendre cette expérieure à l'aune d'un

autre corpus, afin de vérifier la stabilité de ces résultats.

La comparaison de *-euse* et *-rice* doit cependant être analysée en tenant compte de la productivité des suffixes : le suffixe *-rice* n'étant au contraire de *-euse* plus productif, on peut faire l'hypothèse que l'effacement du suffixe *-rice* entraînera une neutralisation de la connotation du suffixe *-euse*. Il est de fait important de regarder les néologismes pour voir si la dimension axiologique demeure. Nous avons vu que la construction de la ressource Lexeur a principalement reposé sur des dictionnaires, dont on peut faire l'hypothèse que les noms ont pu subir une évolution sémantique les éloignant de leur sens construit (Corbin 1987). Du fait de l'écart entre le sens construit et le sens attesté (Corbin 1987), on peut se demander dans quelle mesure la spécialisation sémantique de la suffixation en *-euse* (à savoir sa tendance à la connotation et à une plus forte instrumentalité) est liée à la lexicalisation d'une partie des noms construits par la règle, et dans quelle mesure cela est lié aux propriétés sémantiques de la règle elle-même.

Nous nous proposons de répondre à ces questions en reprenant l'analyse réalisée au chapitre 6.1 sur la base de noms dont nous faisons l'hypothèse qu'ils ne sont pas lexicalisés, et dont le sens attesté est donc plus proche du sens construit (section 6.2).

6.2 Impact de la lexicalisation

La lexicalisation est définie comme « la fixation de dénominations dans le lexique »(Lombard et Huyghe à paraître : 4). Elle s'accompagne parfois pour un nom construit morphologiquement d'un glissement sémantique, c'est-à-dire de la « modification du contenu sémantique d'une forme, l'adjonction d'un nouveau sens »(Le Draoulec *et al.* 2014 : 123). Ce glissement sémantique crée un écart entre le sens construit (c'est-à-dire prédit) par la règle morphologique et le sens attesté (c'est-à-dire réel) tel que fixé dans le lexique (Corbin 1987). Cette distance sémantique qui se traduit en synchronie est souvent le fruit d'une évolution diachronique. Ce phénomène est notamment à l'œuvre dans des noms comme *entraîneuse*, dont le sens attesté n'est pas strictement équivalent à celui d'*entraîneur* dans la mesure où il a acquis une connotation négative, contrairement à ce que prédit la règle de construction.

Pour tenter d'évaluer l'impact de ce phénomène sur nos résultats, nous reprenons l'analyse des barycentres des noms en *-eur*, *-euse* et *-rice* à partir de noms considérés comme néologiques, dont on peut faire l'hypothèse que leur sens attesté est plus proche du sens construit, puisqu'ils ont subi un degré de lexicalisation moindre. Nous présentons dans un premier temps notre méthode de sélection de ces noms (section 6.2.1), puis nous présentons les

résultats (section 6.2.2).

6.2.1 Extraction de noms néologiques

De par leur origine, nous faisons l'hypothèse que les noms extraits de Lexeur sont des noms attestés et potentiellement lexicalisés. *A contrario*, les noms déverbaux en *-eur*, *-euse* ou *-rice* n'apparaissant pas dans Lexeur ont plus de chances d'être néologiques et donc non lexicalisés. Nous avons donc extrait les noms déverbaux en *-eur*, *-euse* et *-rice* absents de Lexeur, c'est-à-dire les noms porteurs des suffixes *-eur*, *-euse* et *-rice* pour lesquels il existe un verbe dans le corpus et qui n'apparaissent pas dans Lexeur.

Pour cela, nous récupérons dans un premier temps l'ensemble des paires $\{N_{suff}, V\}$ présentes dans Lexeur. Pour chaque paire, un programme apprend la règle de transformation formelle liant le dérivé N_{suff} à sa base V (Tanguy et Hathout 2007). Dans un second temps, le programme parcourt l'ensemble des mots du vocabulaire du modèle distributionnel considéré, traite chaque mot finissant par la chaîne *suff* comme un dérivé potentiel et lui attribue une ou plusieurs bases potentielles, 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*. 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 dans le corpus, soit parce qu'elle n'apparait pas du tout.

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 d'agent ou d'instrument construit en *suff*, et dont le second élément est bien le verbe à partir duquel est dérivé le nom considéré, à l'image de *meuleuse* et *meuler*[1]. Nous avons exclu 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*} et {*autrice*, *autre*} ;

1. Le caractère agentif, instrumental ou verbal d'un item est annoté sur la base du *Wiktionnaire*, du *Petit Robert de la langue française*, et du corpus *Wikipedia2018*. Est considéré comme nom d'agent ou d'instrument tout dérivé ayant au moins une acception ou une occurence agentive ou instrumentale dans une de ces trois ressources. De la même façon, est considéré comme un verbe toute base ayant au moins une acception ou une occurrence verbale dans ces ressources. Des noms traditionnellement considérés comme dénominaux mais pour lesquels on peut trouver une base verbale dans une de ces ressources sont dès lors considérés comme potentiellement déverbaux, et sont donc pris en compte. Le nom *stripteaseuse*, dont on considère communément qu'il dérive du nom *striptease*, est ici conservé puisque l'on trouve le verbe *stripteaser* dans le *Wiktionnaire*.

— Le dérivé correspond au féminin d'un adjectif en *-eux*. Cela exclut des paires comme {*niaiseuse*, *niaiser*} et {*nécessiteuse*, *nécessiter*}

— Le dérivé n'a pas d'emploi agentif dans le corpus. Cela exclut des paires comme {*sueur*, *suer*} et {*caprice*, *caper*} ;

— Il s'agit d'une variante orthographique d'une paire présente dans Lexeur. Cela exclut les paires {*co-producteur*, *co-produire*} ou {*entraineuse*, *entrainer*} puisque les paires {*coproducteur*, *coproduire*} et {*entraîneuse*, *entraîner*} 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*}, parce qu la préfixation apporte une contribution sémantique ;

— Le verbe et son dérivé ne sont pas liés sémantiquement. Cela exclut des paires comme {*primeur*, *primer*}, et {*hardeuse*, *harder*}

De la même façon que Lexeur, nous ne distinguons pas les noms d'agent des noms d'instrument dans cette sélection. Nous ne mettons pas non plus de contrainte sur le caractère humain des noms d'agent (*chélatrice*). Enfin, nous n'excluons pas les noms polysémiques (*médiatrice*), ainsi que ceux dont la forme est aussi celle d'un adjectif (*modélisatrice*).

Le tableau 6.1 donne les chiffres aux différentes étapes.

	-eur	*-euse*	*-rice*
Paires potentielles	16 648	2 549	1 649
Paires instanciées	847	204	148
Paires validées	176	28	17

TABLE 6.1 – Nombre de paires {N_{suff}, V} néologiques

Parmi ces 221 paires {N_{suff}, V} extraites, certaines contiennent des noms spécifiques au corpus (*wikificateur*, *clôtureur*). On trouve aussi, de façon relativement importante, des noms qui ne sont pas néologiques. Ces noms ne faisaient pas partie des amorces dans les sections 5.2.2 et 6.1 car ils sont soit absents de Lexeur (*mijoteuse*), soit enregistrés dans Lexeur comme n'ayant pas une base verbale (*hackeuse* n'a pas de base enregistrée, et *designeuse* est considéré comme dénominal). Enfin, on trouve des noms que l'on peut effectivement considérer comme néologiques (*cosplayeuse*, *trolleuse*).

Notons que les effectifs d'amorces sont très réduits pour les noms féminins. Par ailleurs, on remarque que les amorces tendent à sur-représenter certains domaines ontologiques. Ainsi, les noms néologiques en *-rice* désignent en grande partie des métiers de la télécommunication et des média (*co-animatrice*, *reportrice*, *co-réalisatrice*, *coprésentatrice*). La majorité des

noms néologiques en *-euse* désignent quant à eux des sportives et des artistes (*jammeuse*, *contreuse*, *performeuse*, *graffeuse*), ainsi que des machines (*surfaceuse*, *mortaiseuse*, *dégarnisseuse*).

Si nous avons vu en section 5.3.1 que cela ne changeait pas fondamentalement le contenu sémantique du barycentre construit, les amorces se caractérisent ici par leur effectif réduit et leur orientation sémantique. L'analyse des barycentres sera donc nécessairement à nuancer au regard de ces paramètres.

6.2.2 Analyse

Comme précédemment, nous construisons sur la base de ces 176, 28 et 17 amorces respectivement les barycentres des noms d'agent néologiques en *-eur*, *-euse* et *-rice*, dans les cinq modèles distributionnels, et nous calculons les 100 plus proches voisins communs sur la base de leur score de proximité moyen. Nous analysons dans ce qui suit un à un leurs voisinages moyens, notamment au regard des résultats obtenus en section 6.1 à partir des noms issus de Lexeur. Nous ne nous attendons pas à ce que les voisins soient strictement identiques à ceux obtenus précédemment, puisque nous ne partons pas des mêmes amorces – ce qui se traduit par une variation des voisins comme nous l'avons vu en section 5.3.1. Bien que nous prenions aussi en compte ce critère, nous nous intéressons ici plus spécifiquement aux types sémantiques des voisins plus qu'aux voisins distributionnels eux-mêmes. L'analyse pour les noms d'agent en *-eur* est donnée à titre indicatif.

Suffixe *-eur*

Le voisinage du barycentre des noms néologiques en *-eur* contient 7 amorces néologiques (13a), et partage neuf voisins avec le voisinage du barycentre calculé à partir des noms issus de Lexeur (13b).

(13) a. routeur, débogueur, résolveur, ordonnanceur, visualiseur, blogueur, skateur

b. programmeur, bloqueur, mixeur, opérateur, adaptateur, joystick, microphone, installateur, ordinateur

On constate que six des sept amorces que l'on retrouve dans le voisinage sont liées au domaine de l'informatique, à l'exception de *skateur*, et que cinq de ces sept amorces désignent des instruments ou objets, à l'exception de *blogueur* et *skateur*. Un constat similaire peut être fait pour les voisins partagés par les barycentres des noms lexicalisés et néologiques, puisqu'ils sont tous plus ou moins liés au domaine de l'informatique et des technologies, et

seul un voisin est un nom d'agent monosémique (*programmeur*), non porteur d'une interprétation instrumentale.

Plus largement, le voisinage du barycentre des noms néologiques se caractérise par la présence majoritaire de noms d'instrument ou d'objet (concret ou virtuel) liés au domaine de l'informatique, à l'image de *routeur*, *anti-spyware*, *encodeur*, *gratuiciel*, *vidéoprojecteur*, *widget*, et *anti-virus*. Certains voisins sont quant à eux des entités nommées (*BitTorrent*, *KaZaA*, *Skype*), désignant cependant des logiciels, que l'on peut considérer par extension comme des noms d'instrument ou d'objet virtuel à l'image de *encodeur* ou *anti-virus*. On trouve plus marginalement des noms d'agent, dont la plus grande partie sont liés à l'informatique et aux nouvelles technologies : *programmeur*, *youtubeur*, *hacker*, *streamer*, *opérateur*, *blogueur*. Les autres agents relèvent du sport (*skateur*) ou de la musique (*disc-jockey*).

Il ressort de l'analyse des voisins du barycentre des noms néologiques en *-eur* une plus grande tendance à l'instrumentalité d'une part, et une orientation ontologique dirigée vers l'informatique, et plus marginalement vers le sport. Cela semble être la conséquence de la constitution des amorces, qui regroupe un nombre important d'outils informatiques (*anonymiseur*, *valideur*, *multiplexeur*) et parmi les agents qui ne sont pas liés à l'informatique et aux nouvelles technologies, des noms de sportifs (*dunkeur*, *skateur*, *catcheur*).

Suffixe *-euse*

On retrouve parmi les 100 plus proches voisins du barycentre des noms néologiques en *-euse* neuf amorces (présentées en (14a)) et neuf voisins partagés avec le voisinage du barycentre construit à partir des noms extraits de Lexeur (donnés en (14b)). Ces voisins partagés sont tous des noms d'agent dont le référent est féminin.

(14) a. blogueuse, snowboardeuse, bloggeuse, hackeuse, stripteaseuse, réceptionneuse, contreuse

b. boxeuse, snowboardeuse, hackeuse, stripteaseuse, Youtubeuse, tireuse, starlette, barmaid, fêtarde

Plus globalement, le voisinage moyen du barycentre des noms néologiques en *-euse* contient une grande proportion de noms de sportifs. Si la plupart de ces noms de sportifs marque grammaticalement le genre féminin du référent (15a), certains sont non marqués (15c) ou masculins (15c). On retrouve ainsi l'orientation présente au sein des amorces évoquée en section 6.2.1 mais aussi plus largement des noms de profession, pour la plupart relevant des arts ou de la télécommunication (16), à l'exception de *entrepreneure*, *généticienne* et *astrophysicienne*. Seuls trois voisins caractérisent le référent – féminin –

sur la base de son comportement (*fêtarde*), ou de propriétés intrinsèques (*transsexuelle*, *cannadienne*).

(15) a. boxeuse, snowboardeuse, golfeuse, rameuse, bobeuse, contreuse, footballeuse, curleuse, pentathlonienne, marathonienne, pistarde

b. taekwondoïste, trampoliniste, fleurettiste, karatéka

c. volleyeur, bodybuilder

(16) blogueuse, bloggeuse, YouTubeuse, Youtubeuse, youtubeuse, chanteuse-compositrice, parolière, mannequine, pop-star, maquilleur

Contrairement à ce que l'on observait en section 6.1.1, le trait instrumental est quasiment absent du barycentre des noms néologiques en *-euse*. Les seuls voisins porteurs d'une composante sémantique instrumentale sont les noms *doubleuse*, *pointeuse* et *fondeuse*, qui ne sont pas des noms d'instrument monosémiques, mais des noms à la double lecture agentive et instrumentale.

De même, on observe l'apparition d'un grand nombre de prénoms (*Vendula*, *Janay*, *Katsiaryna*) et de noms de famille (*Taeldeman*, *Dvorovenko*, *Liepiņa*) dont le référent est majoritairement féminin – du moins dans le corpus *Wikipedia2018*. Ce type de voisins est absent du voisinage du barycentre des noms d'agent en *-euse* extraits de Lexeur.

Globalement, on constate la disparition des connotations péjoratives ou sexuelles observées en section 6.1.1. Seuls *stripteaseuse* et *showgirl* impliquent un certain degré de sexualisation, et *starlette* une connotation péjorative. La composante instrumentale disparaît, ainsi que la connotation, au profit d'une composante humaine, particulièrement orientée vers le genre féminin. Ce barycentre se rapproche en cela, dans une certaine mesure, du barycentre que l'on avait en section 6.1.2 pour les noms d'agent en *-rice*. Cela suggère que *-euse* est un suffixe agentif sans dimension additionnelle.

Suffixe *-rice*

Le barycentre des noms néologiques en *-rice* présente parmi ses 100 plus proches voisins six de ses amorces (données en (17a)). Ces voisins, au même titre que ceux que le barycentre partage avec le barycentre des noms en *-rice* extraits de Lexeur (17b), relèvent pour la plupart du domaine de la télévision et des média. Cela est une conséquence de la sur-représentation de ce domaine parmi les amorces absentes de Lexeur.

(17) a. coprésentatrice, co-présentatrice, co-réalisatrice, coanimatrice, co-animatrice, médiatrice

b. co-directrice, co-créatrice, standardiste, Directrice, cocréatrice, médiatrice

Une grande partie des voisins du barycentre dénotent des métiers relevant de ce domaine, qu'ils marquent le genre féminin (18a), masculin (18b) ou qu'il s'agisse de noms épicènes (18c).

(18) a. intervieweuse, coanimatrice, co-animatrice, Coanimatrice, chroniqueuse, assistante-réalisatrice

b. co-producteur, animateur, co-animateur, script-éditeur, présentateur

c. panéliste, voix-off, co-vedette, standardiste, recherchiste

À l'image de ce que l'on observe pour le barycentre des noms néologiques en *-euse*, on retrouve des prénoms (*Énora*, *Véronika*, *Jessyca*) et des noms de famille (*Malagré*, *Pulvar*, *Damidot*) dont le référent est majoritairement féminin dans le corpus. Notons qu'une grande partie de ces référents sont liés au domaine ontologique des média, de par leur profession notamment. Globalement, le domaine ontologique ressort au travers des noms de métiers, d'individus, mais aussi des autres voisins dénotant des chaînes télévisées, des émissions ou séries, ou des collectifs associés à la production télévisuelle.

(19) a. Infosport+, AwesomenessTV, TVSud, Vivolta, TV+, ABS-CBN

b. Téléshopping, Tactik, Cocoricocoboy

c. Palmashow, coco-girl

Comparativement au barycentre des noms en *-rice* extraits de Lexeur, le barycentre des noms néologiques en *-rice* n'inclut pas de composante scientifique. La composante agentive semble moins marquée, laissant davantage la place au domaine ontologique, indépendamment de la catégorie sémantique des voisins (agent, humain, objet immatériel). Cela est sans doute aussi le fruit de l'effectif réduit d'amorces et de leur forte orientation sémantique.

Notons que les barycentres des noms d'agent néologiques en *-euse* et *-rice* ne partagent que deux voisins, *Clairembourg* et *Àstrid*, respectivement un nom de famille et un prénom dont les référents sont féminins. Les deux barycentres sont donc rapprochés davantage sur la base du trait du genre que sur le trait d'agentivité. Cela peut s'expliquer par la manifestation bien distincte pour chaque barycentre de l'agentivité.

Notons cependant que ces résultats sont obtenus à partir d'un nombre très limité d'amorces (notamment pour *-euse* et *-rice*), et relativement orientées sémantiques, comme nous l'avons souligné en section 6.2.1. Ces amorces ne sont donc pas nécessairement très représentatives des noms d'agent en *-euse* et *-rice*. Il faudrait pouvoir se baser sur un nombre plus important de noms néologiques. Pour obtenir ces noms, il faudrait donc un corpus plus grand voire plus récent, ce qui permettrait de conforter ces résultats sur un plus grand nombre de données.

6.3 Des noms féminins mâle-lemmatisés

Nous revenons sur un point que nous n'avons pas exploré dans l'analyse des noms en *-euse* et *-rice* lexicalisés (section 6.1) à savoir la forte perte d'amorces liée à leur fréquence. En effet, pour rappel, nous avons initialement extrait un total de 3 476 noms en *-euse* et 1 145 noms en *-rice* distincts de Lexeur, mais seuls 302 et 77 d'entre eux sont représentés dans le modèle. Cela signifie que seuls 8% des noms féminins déverbaux de Lexeur ont une fréquence supérieure ou égale à 5 dans le corpus. La faible fréquence des noms féminins avait déjà été mise en avant dans le tableau 5.2, notamment par comparaison avec la fréquence de leurs équivalents masculins. Or, Zeller *et al.* (2014) suggèrent que la plus grande distance observée pour les dérivés féminins vis-à-vis de leur base pourrait s'expliquer, en plus de la différence de genre (voir section 5.1.1), par leur fréquence. La faible représentation des noms féminins dans les modèles distributionnels est néanmoins à mettre en regard d'une propriété du corpus sur lequel ils sont entraînés : son pré-traitement.

En effet, le corpus est lemmatisé. Si ce choix semble logique pour les raisons que nous avons données en section 3.2.2, la question de la lemmatisation du corpus avant l'apprentissage de représentations vectorielles a encore récemment été discutée dans la littérature. Ainsi, Gonen *et al.* (2019) remarquent que les noms co-occurent davantage avec des unités du même genre grammatical. Les noms féminins sont donc plus facilement rapprochés de noms féminins, et réciproquement les noms masculins de noms masculins. De fait, si le corpus n'est pas lemmatisé, des synonymes de genre grammatical opposé ne seront pas pleinement rapprochés du fait de leurs contextes grammaticalement distincts, participant du biais inhérent aux représentations vectorielles constaté dans la littérature. La lemmatisation est cependant critiquée par les auteurs car si elle permet effectivement de neutraliser le genre grammatical dans le contexte, elle amène souvent à l'effacement de noms cibles féminins, remplacés par leurs équivalents masculins. Les auteurs suggèrent de fait de ne procéder qu'à la lemmatisation du contexte mais pas à celui des cibles.

De fait, on peut se demander dans quelle mesure notre pré-traitement du corpus a un impact sur la représentation des noms féminins dans les modèles, d'autant plus au regard de la faible représentation des noms d'agent féminins dans nos modèles distributionnels. En effet, respectivement 9% et 7% des noms d'agent en *-euse* et *-rice* font l'objet d'un vecteur dans le modèle distributionnel, ce qui peut sembler particulièrement faible, notamment au regard de la représentation des noms d'agent en *-eur* dans la section 5.2.2 (qui était de 37%).

Afin d'évaluer cet impact, et la possible déperdition de noms féminins, nous projetons dans le corpus – brut et lemmatisé – les noms déverbaux en *-euse* et *-rice* extraits de Lexeur. Nous ciblons les formes fléchies au singulier et au pluriel des noms d'agent, dans l'idée que ces formes sont agrégées, à l'aide de l'étape de lemmatisation, et qu'elles sont donc toutes prises en compte pour le calcul des vecteurs. Nous projetons en plus de la forme au singulier extraite de Lexeur la forme au pluriel que nous inférons à partir de la forme au singulier à laquelle nous ajoutons un *s* final. Le calcul des types est fait sur la base des formes au singulier et au pluriel, auxquelles on retire le *s* final. Le tableau 6.2 présente respectivement le nombre de noms distincts dans Lexeur, le nombre total d'occurrences et de formes de ces noms dans le corpus brut, le nombre total de lemmes et de types dans le corpus lemmatisé, et enfin le nombre de formes lemmatisées faisant l'objet d'un vecteur, donc instanciées au moins cinq fois dans le corpus lemmatisé.

	Lexeur	Corpus brut		Corpus lemmatisé		Modèle
		Tokens	Types	Lemmes	Types	
-euse	3 476	168 974	995	35 261	670	302
-rice	1 145	231 353	489	41 288	230	77

TABLE 6.2 – Distribution des noms d'agent féminins dans le corpus Wikipedia2018

Le tableau 6.2 montre de fortes disparités, à la fois entre les suffixes et entre les versions des corpus. On constate tout d'abord que les noms d'agent en *-rice* sont moins nombreux dans Lexeur que les noms d'agent en *-euse*, mais que ces formes sont néanmoins plus fréquentes en corpus, comme le montre le nombre de tokens, le nombre de lemmes, et le nombre de vecteurs. Le nombre total de types indique à ce titre qu'une majorité des lemmes en *-rice* apparaissent moins de cinq fois dans le corpus. Nous donnons un aperçu plus précis des fréquences des lemmes dans le tableau 6.3. Le tableau 6.3 diffère du tableau 5.2 en cela qu'il intègre tous les lemmes, et pas seulement ceux dont la fréquence est supérieure ou égale à 5.

	min	q1	mediane	q3	max	moyenne
-euse	1	1	3	16	8 524	53
-rice	1	1	2	6	12 076	180

TABLE 6.3 – Fréquences des noms d'agent féminins dans le corpus *Wikipedia2018*

Le tableau 6.3 donne un aperçu des fréquences des lemmes des noms en

-euse dans le corpus. On observe quelques différences entre les suffixes *-euse* et *-rice*, mais les valeurs sont globalement du même ordre.

Le second type de disparités mis en évidence par le tableau 6.2 concerne les différences entre les deux corpus. Si la lemmatisation était totalement correcte, on s'attendrait à ce que le nombre de tokens dans le corpus brut et le nombre de lemmes coincident. Or, les deux valeurs diffèrent fortement. Ainsi, les lemmes que l'on trouve dans le corpus lemmatisé ne représentent que 21% des tokens dans le corpus brut pour le suffixe *-euse*, et 18% pour le suffixe *-rice*. De la même façon, seuls 67% et 47% des types sont réellement conservés entre le corpus brut et le corpus lemmatisé.

Une analyse plus détaillée de la lemmatisation permet d'expliquer le nombre relativement peu important de lemmes et de types, ainsi que les fréquences faibles. En effet, sur les 400 327 occurrences en *-euse* et *-rice* (singulier et pluriel confondus) extraites du corpus brut, plus de 80% des formes sont lemmatisées par leur équivalent masculin en *-eur* (*codétentrice* > *codétenteur*) ou en *-eux* (*avantageuse* > *avantageux*, *vétilleuse* > *vétilleux*).

Cela se traduit de plusieurs façons. Ainsi, certains noms ne sont pas dans le corpus lemmatisé car le lemme féminin n'est pas identifié. C'est notamment le cas de la forme *accélératrice*, que l'on retrouve 47 fois au singulier et 47 fois au pluriel. Or, aucune de ces formes n'est lemmatisée sous la forme *accélératrice*, mais sous la forme *accélérateur*. Cela a deux conséquences directes. D'une part, la forme *accélératrice* ne fait l'objet d'aucun vecteur, puisqu'elle est absente du corpus lemmatisé. D'autre part, le remplacement systématique de la forme *accélératrice* par *accélérateur* dans le corpus lemmatisé modifie la représentation vectorielle de *accélérateur* puisque la distribution des formes masculines et celle des formes féminines sont confondues.

Dans d'autres cas, le lemme – féminin ou masculin – n'est pas identifié, et la forme est donc conservée. C'est notamment le cas pour *danseuse* qui apparaît 8 121 fois au singulier, et 1 632 fois au pluriel. La forme *danseuse* n'est jamais lemmatisée en *danseuse*, mais uniquement au masculin sous la forme *danseur* (à hauteur de 92%). Dans les 8% des cas restants, le lemme n'est pas identifié, et la forme est conservée, ce qui nous permet de disposer d'un vecteur pour *danseuse*. Cependant, comme pour *accélératrice*, la représentation du nom masculin correspondant intègre des informations distributionnelles liées à *danseuse*, et que le vecteur de *danseuse* est fortement tronqué. Notons que parmi ces 8%, on trouve des formes au pluriel, qui sont donc elles aussi utilisées comme lemme. La forme *danseuses* fait donc elle aussi l'objet d'un vecteur, au même titre que *danseuse* et *danseur*. Ce phénomène est cependant marginal, puisqu'il concerne globalement moins de 1% de l'ensemble des occurrences en *-euse* et *-rice* confondues. Signalons néanmoins que cela amène à l'absence de certains noms comme *aboyeuse*.

Une partie des noms en *-euse* et *-rice* se voient ainsi associés au lemme féminin et au lemme masculin. Ce phénomène se fait cependant dans des proportions et avec des conséquences distributionnelles variables. Ainsi, dans le cas de *directrice*, que l'on retrouve 13 304 fois en corpus, la forme est lemmatisée et remplacée par *directrice* dans 89% des cas, et par le lemme masculin *directeur* dans les 11% restants. On peut donc penser que la représentation vectorielle de *directrice* est relativement complète (du moins qu'elle aggrère de façon représentative les contextes de la forme), et que celle de *directeur* n'est pas trop bruitée par les informations distributionnelles du féminin qu'elle aggrège. *A contrario*, seules 6 des 1 283 occurrences de *vendeuse* sont correctement lemmatisées, les 1 277 occurrences restantes étant remplacées par le masculin *vendeur*. De fait, le vecteur pour *vendeuse* n'est calculé qu'à partir d'un échantillon très restreint de contextes parmi ceux qui étaient originellement disponibles, et le vecteur pour *vendeur* est largement alimenté par les 1 277 occurrences féminines.

Enfin, certains noms sont présents dans le corpus lemmatisé uniquement sous la forme féminine. Cela peut être le résultat d'une bonne lemmatisation par le parseur, à l'image de *vendangeuse*, dont les 23 occurrences ont toutes été lemmatisées en *vendangeuse*.

On constate donc que la lemmatisation réalisée par Talismane amène à l'effacement d'un grand nombre de noms féminins dans notre modèle distributionnel. Une façon de pallier cette difficulté serait, comme le conseillent Gonen *et al.* (2019), de ne lemmatiser que les mots du contexte. Cela implique l'utilisation d'une version adaptée de Word2Vec comme celle proposée par Levy et Goldberg (2014a), qui permet un traitement différencié des mots cibles et du contexte dans la construction du modèle. Une autre solution aurait été d'utiliser un modèle construit sur le corpus non lemmatisé. Cependant, encore une fois, notre volonté n'est pas ici de neutraliser ce biais mais de l'étudier (s'il fait partie des propriétés sémantiques de ces noms). À ce stade de l'étude, nous conservons la lemmatisation réalisée afin et de conserver les modèles entraînés par Word2Vec afin de garantir la cohérence des analyses.

6.4 Discussion

L'analyse comparative des barycentres des noms d'agent en *-eur*, *-euse* et *-rice* que nous venons de présenter a permis de tirer deux conclusions. Tout d'abord, on remarque que sur le plan distributionnel, les noms déverbaux en *-eur*, *-euse* et *-rice* ne sont pas représentés de façon similaire. Une première raison à cela repose évidemment sur la différence de genre, à la fois gramma-

tical et référentiel, des noms étudiés. La question de la séparation de ces deux facteurs pour l'étude de ces noms se pose, notamment si l'on considère que les noms d'agent en *-eur*, *-euse* et *-rice* se trouvent dans une zone intermédiaire entre flexion et dérivation (Bonami et Boyé 2019). Malgré la lemmatisation, les noms masculins ne co-occurent pas avec les mêmes types de mots en corpus. C'est tout particulièrement le cas dans un corpus comme *Wikipedia2018*, pour lequel il a été montré que les femmes ne sont pas décrites de la même façon que les hommes (Wagner *et al.* 2015). Les noms féminins étant globalement moins fréquents (voir tableau 5.2), leur représentation vectorielle est moins précise et plus sujette aux variations idiosyncratiques. Ces idiosyncrasies n'empêchent cependant pas l'émergence de propriétés sémantiques, et le contrôle du critère de la fréquence pourrait amener à une perte dommageable de données déjà peu nombreuses.

Signalons au passage que nous avons aussi tenté d'approcher les différences entre ces trois suffixations via les vecteurs de différence (*offset vectors*), calculé en soustrayant au vecteur du dérivé le vecteur de sa base. On obtient ainsi le vecteur qui correspond théoriquement à l'apport sémantique de la suffixation pour la paire donnée. On fait alors l'hypothèse que le vecteur de différence moyen calculé à partir de l'ensemble des vecteurs de différence pour un procédé donné représente la suffixation liant la base au dérivé. Cette méthode a été appliquée pour l'étude de l'expression du genre en flexion et en dérivation (Mickus *et al.* 2019), mais aussi des connotations Bolukbasi *et al.* (2016). Nous avons utilisé ces vecteurs moyens de différence de deux façons. Nous avons dans un premier temps comparé les voisins de ces vecteurs pour les suffixes *-eur*, *-euse* et *-rice*. Les résultats préliminaires font apparaître que les vecteurs de différence se trouvent dans une zone grise de l'espace vectoriel, et suggèrent la présence d'une composante sémantique supplémentaire associée au genre, les voisinages contenant majoritairement des noms difficiles à classer, ou des patronymes genrés. Nous avons dans un second reconstruit un vecteur pour les dérivés à partir des vecteurs des bases et du vecteur moyen du suffixe. Nous avons alors calculé pour chaque suffixe les scores de proximité entre les vecteurs des dérivés attestés et ceux des dérivés reconstruits. Les scores moyens plus élvés obtenus pour les suffixes *-eur* et dans une certaine mesure *-rice* suggèrent que la suffixation en *-eur* est plus régulière que les suffixations en *-euse* et *-rice*, et au sein des suffixes féminins, que celle en *-rice* est plus régulière et prédictible que celle en *-euse*. Ces résultats restent cependant à confirmer.

Plus globalement, la confrontation des différentes configurations étudiées dans cette partie, tant sur le plan de la description que des outils, souligne l'importance d'un contrôle précis sur le dispositif expérimental. Ainsi, la sélection en amont des données linguistiques joue un rôle primordial dans l'ana-

lyse distributionnelle, et notre étude est limitée par l'assimilation de plusieurs catégories sémantiques. Par ailleurs, nos analyses ont fait émerger les limites de notre définition des noms d'agent, notamment relativement au trait humain et au lien avec le prédicat, pour la caractérisation des barycentres. Nous souhaitons donc à ce stade raffiner l'expérience, tant sur le plan méthodologique que théorique, ce que nous proposons de faire dans la partie IV, au travers de l'étude plus spécifique de la catégorie lexicale des noms d'agent.

Quatrième partie

Caractérisation de la catégorie lexicale des noms d'agent

La partie III était consacrée à la comparaison sémantique des noms déverbaux en *-eur*, *-euse* et *-rice*. Nous y avons mis en place une démarche visant à tester la pertinence et l'apport de la représentation par un vecteur moyen de ces trois ensembles de noms d'agent et d'instrument, dans une approche extensive. Cette méthode nous a permis d'explorer les propriétés sémantiques de classes de lexèmes définies formellement sur la base d'un large ensemble de données, ainsi que de caractériser à grande échelle les différences associées aux noms féminins en *-euse* et *-rice*.

Mais notre étude a mis en évidence certaines des limites de cette méthode, notamment au niveau de la variation de la représentation de l'agentivité, difficile à distinguer des traits d'instrumentalité et d'humanité.

Nous avons également constaté dans la section 5.2.2 que le caractère agentif de certains noms ou syntagmes nominaux est sujet à discussion en fonction des critères considérés pour la délimitation de la classe des noms d'agent. L'attribution d'un trait agentif au syntagme *chien de garde* (chien – ou personne, dans un emploi figuré – qui surveille quelque chose ou quelqu'un) dépendra par exemple de l'activation d'une contrainte relative au trait humain. De fait, la définition des noms d'agent n'est pas clairement posée dans la littérature. En tant que catégorie lexicale, nous n'avons pas de moyens linguistiques d'en définir les contours. Il n'existe notamment pas de tests linguistiques, et s'il est admis que les noms déverbaux en *-eur* constituent le cœur de cette catégorie, l'intégration de noms comme *émeutier*, *disquaire* ou *médecin* au sein de la classe des noms d'agent n'est pas clairement tranchée.

Le nom d'agent est traditionnellement compris comme un nom dénotant un référent relativement à l'action qu'il effectue. La définition des noms d'agent repose donc principalement sur la présence d'un trait actionnel dans le nom. Là où le critère morphologique des noms dérivés permettait de garantir la présence d'une composante actionnelle dans le nom, héritée du verbe, la question de son identification voire de sa simple présence dans les noms non déverbaux se pose. Pour surmonter cet obstacle, nous tirons profit des représentations sémantiques offertes par la sémantique distributionnelle, nous permettant de nous libérer de la contrainte formelle.

Nous caractérisons la catégorie sémantique des noms d'agent au travers de trois axes. Nous étudions d'abord les propriétés morphosémantiques de cette catégorie à l'aide de ses représentants les plus prototypiques, en nous interrogeant notamment sur la présence d'une dimension agentive dans les modèles distributionnels (chapitre 7). Nous montrons à ce titre, grâce à l'étude du voisinage d'un barycentre construit à partir d'un échantillon contrôlé de noms d'agent prototypiques, que les noms d'agent ne sont pas morphologiquement contraints, et que la catégorie recouvre des profils morphologiques variés. Nous établissons le caractère agentif des voisins distributionnels par

comparaison avec des barycentres de noms définis par leur non-agentivité, à savoir les noms d'humain généraux et phasiques, les noms relationnels et les gentilés.

Une fois la diversité morphologique des noms d'agent établie, nous nous attachons à comparer les différents procédés dérivationnels agentifs afin d'identifier les similarités et les différences des noms qu'ils construisent (chapitre 8). Nous comparons dans un premier temps les voisinages des barycentres construits à partir de noms d'agent en *-ant*, *-aire*, *-eur*, *-ien*, *-ier* et *-iste* afin de faire émerger les propriétés morphosémantiques et ontologiques de ces classes. Nous montrons que ces propriétés sont (du moins partiellement) corrélées aux préférences de ces suffixations comme la construction de noms de spécialistes sur des bases dénotant des domaines à l'aide des suffixes *-ien* et *-iste*, ou l'affinité du suffixe *-ier* avec les bases nominales dénotant des objets. Ces corrélations sont dans un second temps confortés par une étude *bottom-up* basée sur le clustering des noms d'agent.

Enfin, nous nous interrogeons sur l'homogénéité sémantique de la classe des noms d'agent en nous penchant sur leurs propriétés référentielles (chapitre 9). Nous explorons à ce titre la pertinence et les caractéristiques d'une tripartition des noms d'agent selon leur caractère institutionnalisé (pour les noms statutaires comme *coiffeur*), ponctuel (pour les noms occasionnels comme *agresseur*) ou habituel (pour les noms dispositionnels comme *bosseur*) du prédicat dénoté. En combinant le clustering des noms d'agent et l'analyse des barycentres des trois classes, nous montrons que les noms d'agent sont bien discriminés dans les modèles distributionnels sur la base des traits sémantiques liés au caractère statutaire, occasionnel ou dispositionnel, traits pour lesquels nous dressons des corrélations préliminaires avec des propriétés morphosémantiques, à l'image de la prévalence de bases dénotant des propriétés pour les noms dispositionnels.

Plus largement, nous montrons la diversité, tant formelle que sémantique, des noms d'agent, ainsi que l'apport de la sémantique distributionnelle – dans un contexte contrôlé – pour l'étude de phénomènes linguistiques fins. Cette étude s'inscrit dans une démarche combinatoire impliquant une méthodologie computationnelle contrôlée et des connaissances sémantiques fines. Le travail présenté dans cette partie a été réalisé en collaboration avec Richard Huyghe, et a fait l'objet de plusieurs publications (Huyghe et Wauquier 2020, Huyghe et Wauquier à paraître, Huyghe et Wauquier soumis). Les données présentées sont accessibles à l'adresse `https://github.com/mwauquier/PhdData/tree/main/Part4`.

Chapitre 7

Trait agentif

La catégorie lexicale des noms d'agent, c'est-à-dire les noms qui décrivent les entités qui réalisent des actions (*chanteur*, *illustrateur*), présente des contours flous quant à ses propriétés définitoires. Ainsi, il n'existe à notre connaissance pas de tests linguistiques permettant d'identifier les noms d'agent. La construction à partir d'une base verbale est ainsi souvent considérée comme un prérequis garantissant la dénotation de l'agent réalisant l'action dénotée par la base, mais des noms dénominaux sont aussi parfois étudiés à l'aune de leur agentivité (*charpentier*, *mécanicien*). La question se pose alors de garantir la dénotation d'une action, et de la référence à un agent, lorsque ces propriétés ne sont pas héritées de la base. Quelle que soit la position prise dans les différents travaux traitant des noms d'agent, la délimitation de la classe des noms d'agent n'est jamais clairement établie.

L'objectif de ce chapitre est de savoir si la catégorie lexicale des noms d'agent est définie voire contrainte sur un plan morphosémantique par la dérivation déverbale, et si non, d'évaluer son extension. Pour ce faire, nous nous proposons d'observer le cœur de la catégorie des noms d'agent selon la méthodologie présentée dans la partie III, afin d'analyser les propriétés morphosémantiques des voisins du barycentre de la classe des noms d'agent.

La classe lexicale des noms d'agent est ici approchée par le biais des noms d'agent prototypiques, à savoir les noms d'agent déverbaux monosémiques en *-eur*. Ce chapitre présente donc un raffinement de la méthode par rapport à la partie III dans la mesure où nous visons l'étude d'un sous-ensemble des noms déverbaux en *-eur*, tous n'étant pas des noms d'agent. Ce passage d'une classe définie par un suffixe à une classe définie par des propriétés sémantiques implique donc de revoir les données, tant lexicales que distributionnelles. Nous ne souhaitons en effet pas intégrer dans notre barycentre des vecteurs qui ne représentent pas uniquement un nom d'agent, que l'ambiguïté soit sémantique (*navigateur*) ou formelle (*enchanteur*). Une sélection

plus stricte des noms et le recours à un modèle distributionnels dont les vecteurs sont désambiguisés syntaxiquement sont mis en place pour remédier à ces difficultés.

L'examen du barycentre ainsi construit propose une description sémantique plus fine de la catégorie des noms d'agent, construite sur la proximité sémantique des mots du corpus par rapport à la représentation moyenne des noms d'agent. Cela nous permet d'explorer les propriétés morphosémantiques des noms d'agent, et notamment évaluer les procédés dérivationnels en œuvre pour leur formation, mais aussi certaines de leurs propriétés sémantiques. Il ressort de cette analyse heuristique que les noms d'agent sont de types morphologiques variés, le prédicat pouvant être hérité, ou être construit en lien avec la base (lorsque base il y a). Nous montrons par ailleurs que le trait humain est une propriété saillante des noms d'agent, mais qui n'est pas suffisante à la qualification en noms d'agent. La comparaison d'un barycentre de noms d'agent à des barycentres de noms d'humain non agentifs (noms généraux, phasiques, relationnels et gentilés) souligne ainsi la présence distincte d'un trait humain et d'un trait agentif au sein des noms d'agent. La proximité de noms aux barycentres de noms d'agent et de noms d'humain non agentifs permet alors être vue comme un indice de la présence du trait agentif.

Le chapitre s'organise comme suit. Nous commençons par présenter la reprise de la méthode en détail (section 7.2), puis nous présentons les résultats de l'analyse de la représentation des noms d'agent en *-eur* prototypiques (section 7.3). Enfin, nous testons l'extension de la catégorie lexicale des noms d'agent et la saillance du trait agentif dans la représentation distributionnelle de la catégorie en la confrontant avec des catégories prototypiquement non agentives (section 7.4).

7.1 De la définition lexicale des noms d'agent

La difficulté à circonscrire la catégorie lexicale des noms d'agent s'explique par la transposition lexicale du rôle sémantique d'agent, cette notion syntaxique étant elle-même floue. L'absence de consensus concernant les critères définitoires de l'agentivité syntaxique se retrouve ainsi dans la flou définitionnel de l'agentivité lexicale. Nous passons dans un premier temps en revue les principales difficultés relatives à la description du rôle sémantique d'agent (section 7.1.1), puis nous présentons les conséquences de sa transposition dans le domaine lexical (section 7.1.2) avant de poser la définition de l'agentivité que nous utilisons dans ce travail afin de rendre opérationnelle son étude (section 7.1.3).

7.1.1 L'agentivité

Les rôles sémantiques rendent compte des relations syntaxiques entre un verbe et ses arguments, et sont mobilisés à la fois dans l'analyse syntaxique et dans la description verbale. Ces rôles sont multiples, et diffèrent d'une typologie à l'autre en fonction des critères définitoires considérés. Kipper Schuler (2005) identifie par exemple 21 rôles, dont agent, instrument, patient, source, et destination. Dans l'exemple 20, les arguments *sœur* et *agresseur* du verbe *assommer* désignent respectivement son agent (l'entité qui réalise l'action dénotée par le verbe) et son patient (l'entité qui est affectée par l'action dénotée par le verbe).

(20) Ma sœur a assommé l'agresseur.

Parmi ces rôles, celui d'agent est toujours présent dans les typologies malgré l'instabilité de sa définition d'une typologie à l'autre (Dowty 1991). Les critères proposés pour identifier les agents sont nombreux, dont l'animéité, le contrôle et la responsabilité de l'agent, le caractère intentionnel et conscient de la réalisation de l'action, ou encore la dynamicité de l'action. Les positions divergent en revanche quant à la nécessité et à la réalisation de certains de ces critères, et tout particulièrement l'animéité et l'intentionnalité des agents. Gruber (1967) soutient par exemple le caractère animé et intentionnel de l'agent. À l'inverse, Cruse (1973) considère que l'animéité n'est pas une condition nécessaire à la qualification d'un participant en tant qu'agent. Schlesinger (1989) rejette pour sa part le trait d'intentionnalité.

Des approches plus souples de la définition du rôle d'agent ont été proposées, notamment par le biais d'une vision prototypique de l'agentivité (Lakoff 1977, DeLancey 1984, Dowty 1991, entre autres). L'identification du rôle d'agent repose sur des propriétés communes partagées par les agents (Kleiber 1988). L'ensemble des critères n'ont donc pas besoin d'être systématiquement satisfaits. Fillmore (1968) affirme ainsi que les agents sont typiquement des instigateurs animés, et Kipper Schuler (2005) qu'ils sont généralement des humains ou des animés, et habituellement volontaires. L'agentivité peut alors être envisagée comme une propriété scalaire, dont le degré varie en fonction des critères définitoires satisfaits. Certains arguments peuvent alors être analysés comme plus agentifs que d'autres (Grimm 2011).

7.1.2 De l'agent au nom d'agent

La transposition du rôle d'agent dans le domaine lexical des noms implique de se détacher des relations contextuelles entre le verbe et ses arguments. La catégorisation lexicale comme nom d'agent ne dépend en effet pas

du lien entretenu au niveau syntaxique par le nom et le verbe, mais de la description du référent comme l'agent d'une action intrinsèquement spécifiée par le nom. Si nous reprenons l'exemple (20), le nom *agresseur* peut être considéré comme un nom d'agent, puisque qu'il décrit l'agent de l'action spécifiée – ici héritée de la base verbale *agresser*. Malgré son rôle syntaxique de patient, il n'est pas considéré comme un nom de patient sur le plan lexical puisqu'il ne désigne pas le patient de l'action spécifiée. À l'inverse, *sœur* ne sera pas considéré comme un nom d'agent puisqu'il ne désigne pas l'agent d'une action spécifiée dans le nom, mais plutôt comme un nom relationnel (Barker 2008), c'est-à-dire un nom dénotant un humain en relation interpersonnelle avec d'autres humains.

De fait, la classification d'un nom comme nom d'agent dépend de deux conditions : la spécification d'une action et la description de l'agent correspondant. Limiter la classe des noms d'agent aux noms déverbaux dénotant l'argument agentif du verbe de base permet de garantir la référence à l'agent et la présence d'un composant actionnel dans la structure sémantique du nom par héritage du prédicat verbal. Cette définition contrôlée induit cependant une division de cette classe certes hétérogène sur le plan morphologique, mais homogène sur le plan sémantique. Des noms comme (21a), dérivés de verbes dynamiques et dénotant leur agent, seront considérés comme des noms d'agent, mais pas leurs quasi synonymes ou les noms non déverbaux sémantiquement très proches en (21b).

(21) a. sculpteur, guérisseur, rédacteur, manifestant, protestataire

b. artiste, médecin, scribe, gréviste, émeutier

Retirer la contrainte morphologique de la construction déverbale amène cependant à se reposer la question de la validation des deux conditions présentées précédemment, et donc par extension de la délimitation de la classe lexicale des noms d'agent : comment garantir la présence des traits actionnels et agentifs sans la garantie de leur héritage verbal ? Si le composant actionnel peut être hérité d'une base non verbale, comme dans le cas de *gréviste* ou *émeutier* en (21b), et dont les bases nominales sont intrinsèquement actionnelles puisqu'elles dénotent des événements, l'affaire s'avère plus compliquée pour des noms comme ceux donnés en (22).

(22) historien, bijoutier, juriste, prophète, tribun, disquaire, tyran, cardiologue

Un élément de réponse peut être apporté par les exemples donnés en 21, et plus précisément par leur comparaison. Si l'on admet que les noms en (21a) sont effectivement des noms d'agent, et que les noms en (21b) en sont sémantiquement très proches, on peut faire l'hypothèse que ces derniers

présentent les mêmes propriétés sémantiques, y compris les traits actionnel et agentif. La proximité sémantique, et par extension distributionnelle, à des noms d'agent peut ainsi être vue comme un indice d'agentivité. Par extension, nous faisons l'hypothèse que l'analyse du barycentre construit à partir de noms d'agent avérés pourrait permettre d'identifier les propriétés morphosémantiques de la classe des noms d'agent, par le biais des voisins dont on peut faire l'hypothèse qu'ils seront agentifs, qu'ils aient servi à la construction du barycentre ou non.

7.1.3 Notion opérationnelle de l'agentivité

Pour construire la représentation unifiée de la classe lexicale des noms d'agent, nous avons donc besoin de fixer la définition de l'agentivité, tant sur le plan syntaxique que lexical.

Afin de ne pas être *a priori* restrictif concernant les items lexicaux à considérer comme agentifs, nous nous basons sur une conception large des agents en tant que rôle sémantique. Nous définissons ici les agents comme des effectuateurs (c'est-à-dire des entités qui déploient de l'énergie pour réaliser des actions) qui sont prototypiquement, mais non nécessairement animés et intentionnels (Van Valin et LaPolla 1997 : 118).

Nous adoptons de fait une approche prototypique, mais qui diffère de celle de Dowty (1991). Contrairement à sa définition des proto-agents, nous considérons la dynamicité comme un critère définitoire des agents. Ce choix nous permet de distinguer les agents des cause et des expérienceurs (cafardeur), sur la base de la dynamicité des situations dans lesquelles ils sont impliqués.

Écarter l'animéité comme condition nécessaire affaiblit la distinction entre agents et instruments, puisque ces derniers sont souvent définis comme des entités respectivement animées et inanimées. Pour pallier ce flou, nous définissons les instruments comme des entités qui sont fondamentalement utilisées par d'autres entités pour réaliser des actions. Nous justifions par ailleurs le caractère facultatif de l'animéité par notre souhait de ne pas exclure de façon systématique les forces naturelles et les agents biologiques, chimiques ou abstraits qui réalisent des actions de façon autonome.

Dans la suite de ce travail, nous considérons donc comme nom d'agent tout nom dont l'entité décrite par le référent répond à cette définition du rôle sémantique d'agent.

7.2 Spécification de la méthode

Nous développons dans cette partie la même approche que dans la partie III, à savoir que l'on appréhende globalement un ensemble de mots. Pour cela, nous reprenons donc le principe des barycentres. Si les principes de la méthodologie sont calqués sur celle développée dans le chapitre 5, nous mettons en place un niveau de contrôle suplémentaire sur les données d'entrée (section 7.2.1) et sur le matériel distributionnel (section 7.2.2) afin de garantir une analyse plus précise.

7.2.1 Sélection de noms d'agent prototypiques

L'objectif est d'accéder au « cœur «de la catégorie des noms d'agent en passant par des représentants prototypiques de cette classe. En français, il est communément admis que ce cœur contient les noms d'agent déverbaux en *-eur* (voir chapitre 5). Cibler ces noms nous garantit donc l'accès aux propriétés prototypiques de l'agentivité. La ressource Lexeur nous permettait déjà de travailler à partir de ces noms, mais non sans certains écueils. Lexeur contient indistinctement des noms d'agent (*acheteur*), des noms d'instrument (*disjoncteur*), et des noms qui ont les deux sens (*navigateur*). Cela ne pose pas de problèmes pour la comparaison exploratoire des noms d'agent en *-eur*, *-euse* et *-rice* car il s'agissait de comparer les groupes définis par ces suffixes, sans distinguer ces propriétés lexicales. Un tri excluant les noms d'instrument ou les noms polysémiques n'aurait pas été envisageable, notamment pour les noms d'agent féminins en *-euse* et *-rice*, au vu des effectifs déjà très limités.

Or, nous ciblons ici précisément la catégorie lexicale des noms d'agent, et non celle des noms d'instrument. Ces deux catégories présentent des propriétés sémantiques différentes à un niveau de granularité fine. Nous souhaitons éviter que les propriétés sémantiques des noms d'agent soient « diluées », confondues avec celles d'autres catégories sémantiques. Pour limiter au maximum cette influence, nous avons opéré une sélection stricte de noms d'agent déverbaux monosémiques en *-eur*, *-euse* et *-rice*[1].

Suite aux discussions en section 7.1.3, nous définissons les noms d'agent comme des noms d'effectuateur, c'est-à-dire des noms qui dénotent des entités qui déploient de l'énergie pour réaliser des actions, et dont l'action réalisée est liée à la dénotation de la base. Nous ne posons pas de conditions sur le trait humain (*rongeur*), d'animéité (*chélateur* 'ion, atome ou molécule qui se lie à un autre ion pour former un composé chimique'), ou d'intentionnalité

1. Nous faisons le choix dans cette étude d'inclure les féminins en *-euse* et *-rice* dès lors qu'ils respectent les mêmes critères de sélection que le masculin en *-eur* car nous estimons que la variation sémantique liée au genre est orthogonale aux propriétés agentives.

(*ronfleur*). Nous extrayons les noms en *-eur*, *-euse* et *-rice* de la base de données Lexique3 (New *et al.* 2004). Nous optons pour cette ressource car elle se compose de plus de 3 000 noms en *-eur* du français contemporain, mais contrairement à Lexeur, ne se limite pas à un procédé dérivationnel précis. Si le présent chapitre se concentre plus précisément sur les noms d'agent déverbaux en *-eur*, nous comparons dans le chapitre 8 les noms d'agent construits avec d'autres suffixes (*-aire*, *-ant*, *-ien*, *-ier*, *-iste*). Lexeur ne nous permettant pas l'étude de ces noms, et afin de garantir la cohérence de notre méthodologie tout au long de cette partie IV, nous privilégions Lexique3.

Nous obtenons à l'issue de cette extraction un total de 2 215 noms, qui ne sont cependant pas tous des noms d'agent déverbaux prototypiques. Nous procédons à une sélection manuelle guidée par le principe de monosémie [2] d'une part, garantissant que le nom désigne exclusivement un agent, et le principe d'une construction déverbale motivée d'autre part, afin de garantir la présence d'une action dans la structure du nom. Nous excluons donc les noms suivant les critères suivants :

— les noms non dérivés (*peur*)

— les noms non analysables en synchronie comme étant liés à une base verbale (*ambassadeur*)

— les noms déverbaux dont au moins une acception n'est pas agentive, qu'ils dénotent un état (*clameur*), un instrument (*navigateur*), un possesseur (*détenteur*), un expérienceur (*cafardeur*), un stimulus (*inspirateur*), un bénéficiaire (*receveur*), etc.

— les noms dont au moins une acception est fortement désémantisée vis à vis de sa base verbale (*chauffeur* par rapport à *chauffer*)

— les noms en *-euse* ambigüs entre la forme féminine de *-eur* et le féminin des qualificatifs en *-eux* (*chatouilleuse* vis à vis de *chatouilleur* et *chatouilleux*)

Notons que malgré l'ambiguité syntaxique entre adjectif et nom pour un nombre important de lexèmes (*enchanteur*), nous choisisons de les conserver.

2. Nous désignons par le terme monosémique les noms réellement monosémiques, c'est-à-dire les noms qui n'ont qu'un seul sens lexical, agentif, comme *acheteur*, ainsi que les noms monotypés, c'est-à-dire qui ont plusieurs sens mais qui désignent tous un agent, comme *ouvreur*, dans la mesure où ce type de polysémie n'induit pas une dilution du trait agentif. La monosémie d'un nom est évaluée à partir de sa description dans trois ressources lexicographiques : *le Trésor de la Langue Française informatisé*, *le Petit Robert de la langue française*, et *Wiktionnaire*. Un nom est considéré comme monosémique si l'ensemble de ses définitions dans ces trois resssources présentent le référent comme un agent.

Nous justifions ce choix et le dispositif distributionnel que nous mettons en place pour pallier cette ambiguïté en section 7.2.2.

A l'issue de cette sélection, 1 121 noms sont conservés. Du fait de la méthodologie que nous présentons en section 7.2.2, seuls les noms dont la-fréquence [3] est supérieure ou égale à 5 sont conservés. Nous nous retrouvons donc avec une liste finale de 681 noms [4], dont 624 noms masculins en *-eur*, 49 noms féminins en *-euse* et 8 noms féminins en *-rice*. Ces noms sont majoritairement des noms d'humain à l'exception de quelques noms dénotant des animaux (*rongeur*) ou des agents chimiques (*inhibiteur*), et quelques noms sous-spécifiés par rapport à l'animéité (*catalyseur*).

Notons que la totalité des 681 noms sélectionnés, à l'exception de 8 d'entre eux (*motivateur*, *publieur*, *rappeur*, *squatteur*, *supporteur*, *catcheur*, *dictateur*, *dictatrice*) sont présents dans les amorces utilisées dans la partie III. L'absence initiale de ces noms s'explique par leur absence dans la ressource Lexeur (*motivateur*, *publieur*, *rappeur*, *squatteur*, *supporteur*) ou par l'absence de base verbale dans Lexeur (*catcheur* étant noté comme dénominal, *dictateur* et *dictatrice* n'ayant pas de base associée).

7.2.2 Adaptation du modèle distributionnel

Les 681 noms précédemment sélectionnés servent d'amorces dans la construction du barycentre des noms d'agent déverbaux prototypiques en *-eur*. Nous choisissons cependant de ne pas reprendre le modèle distributionnel utilisé dans les chapitres 5 et 6 qui ne permet pas de gérer l'ambiguité syntaxique de certaines formes comme *enchanteur* entre nom et adjectif. Ainsi, une forme comme *enchanteur* fait l'objet d'un unique vecteur aggrégeant les occurrences adjectivales et nominales.

Afin de distinguer les occurrences nominales et adjectivales, et ainsi calculer deux vecteurs distincts pour les lexèmes ENCHANTEUR$_{\text{N}}$ et ENCHANTEUR$_{\text{ADJ}}$, nous décidons d'entraîner notre modèle distributionnel sur un corpus étiqueté morphosyntaxiquement. Nous lemmatisons et étiquetons le corpus à l'aide de l'analyseur Talismane. Ce choix se traduit cependant par l'importance accordé à l'analyseur et à ses performances. L'étiquetage réalisé par Talismane reposant sur un apprentissage supervisé, le prétraitement du corpus introduit des erreurs de lemmatisation (comm vu dans le chapitre 6) et d'étiquetage. Ainsi, le taux de précision de l'étiquetage varie entre 93% et 97% en fonction des configurations du parseur et du corpus (Urieli 2013 : 141).

3. Fréquence calculée pour les lemmes étiquetés en tant que nom dans le corpus. Voir section 7.2.2

4. L'ensemble des données utilisées sont accessibles sur Github à l'adresse `https://github.com/french-agent-nouns/`.

Voisin	Score de proximité moyen
NC :plombier	0.769
NC :truand	0.751
NC :escroc	0.745
NC :rabatteur	0.730
NC :proxénète	0.722
NC :coiffeur	0.718
NC :prestidigitateur	0.717
NC :garagiste	0.704
NC :gangster	0.704
NC :malfrat	0.699

TABLE 7.1 – 10 plus proches voisins du barycentre des noms d'agent déverbaux prototypiques en *-eur*

Bien qu'il puisse induire du bruit, l'étiquetage morphosyntaxique nous permet de désambiguiser les noms qui nous intéressent pour une couverture plus large, comme nous le détaillons plus loin. À l'image de ce que nous faisions précédemment, la forme est conservée lorsque le lemme n'est pas identifié.

Pour pallier l'instabilité des modèles, nous reprenons la méthode COS présentée dans le chapitre 5 et utilisée dans le chapitre 6. Nous entraînons cinq modèles distributionnels sur le corpus *Wikipedia2018* et en utilisant les mêmes paramètres par défaut. Nous construisons le barycentre des noms d'agent déverbaux en *-eur* monosémiques en faisant la moyenne des vecteurs des 681 noms étiquetés NC dans le corpus. Nous identifions les 100 plus proches voisins moyens à partir des voisins dont le score de proximité moyen est le plus élevé, indépendemment de l'étiquette syntaxique de ce voisin. La section 7.3 présente l'analyse de ces 100 voisins. Un aperçu des 10 premiers voisins (avec leur score de proximité moyen aux barycentres) est donné dans le tableau 7.1.

7.3 Caractérisation de l'agentivité

L'analyse des 100 plus proches voisins du barycentre permet d'avoir un aperçu des propriétés sémantiques de la classe indépendamment de toute contrainte formelle. Nous présentons dans un premier temps ces propriétés à partir de l'analyse des 100 voisins moyens des barycentres des noms d'agent déverbaux prototypiques en *-eur* (section 7.3.1). Nous proposons pour cela une analyse plus fine et approfondie des voisins que celle proposée dans la partie III afin de faire émerger les propriétés prototypiques de l'agentivité.

Nous abordons cette analyse à l'aune des types morphologiques et des suffixes des voisins, et des propriétés grammaticales et sémantiques des bases. Nous évaluons ensuite brièvement dans la section 7.3.2 l'impact des contraintes sémantiques imposées sur la sélection des amorces sur la représentation vectorielle des noms déverbaux en *-eur* (et notamment sur l'agentivité), en comparant le barycentre des noms en *-eur* analysé dans le chapitre 5 avec ce nouveau barycentre.

7.3.1 Profil morphosémantique des noms d'agent

Comme l'illustrent les 10 plus proches voisins présentés dans le tableau 7.1, l'ensemble des 100 voisins sont des noms [5], et plus précisément des noms d'humain, à part *chien* au 61e rang. Notons que 27% des voisins sont des amorces. C'est moins que les 45% que l'on avait observé dans l'expérience du chapitre 5.2.2. On observe cependant là aussi une répartition plutôt uniforme parmi les 100 premiers voisins. Là encore, donc, ces amorces ne constituent pas à elles seules le cœur du barycentre et se trouvent parmi d'autres noms d'agent. Ceci constitue un premier indice, déjà vu dans la partie III que l'agentivité ne concerne *a priori* pas que les noms d'agent déverbaux monosémiques prototypiques en *-eur*.

Pour nous faire une idée de l'extension morphologique de la catégorie des noms d'agent, nous analysons les propriétés morphosémantiques des voisins du barycentre. Nous abordons cette analyse à l'aune des types morphologiques des voisins. Ces types recouvrent les procédés d'affixation (Haspelmath et Sims 2013), de conversion (Tribout 2010), de composition (Haspelmath et Sims 2013), de formation extragrammaticale (Fradin *et al.* 2009), mais aussi des noms simples (Huyghe *et al.* 2017), des noms complexes non construits [6], et des noms dont le type est indéterminé car une ambiguïté subsiste entre plusieurs types morphologiques. La distribution des types morphologiques est donné dans le tableau 7.2. Notons que l'analyse morphologique est fournie en synchronie et ne prend donc pas en compte la formation historique. En cas de formation multiple, l'annotation s'applique uniquement à la dernière opération morphologique. Un nom comme *couturier* est donc analysé comme un nom dénominal suffixé en *-ier* à partir du nom *couture*, *couture* étant

5. Pour rendre la lecture des exemples plus claire, nous retirons dans la suite de cette étude les indications grammaticales fournies par le parseur lorsque nous citerons les voisins. Sauf indication contraire, nous parlerons du voisin *plombier* pour désigner en réalité le voisin *NC :plombier*.

6. Corbin (1987 : 459) définit les mots complexes non construits comme des mots dont‹ la structure interne et le sens ne sont que partiellement superposables », et dont l'analysabilité morphosémantique n'est donc pas complète.

lui-même un nom déverbal dérivé de *coudre*.

Affixé	Convert	Composé	Simple	Complexe	Extragram.	Indéterminé
64	8	3	10	8	1	6

TABLE 7.2 – Types morphologiques des voisins du barycentre des noms d'agent en *-eur*

Le tableau 7.2 montre que les voisins sont très majoritairement affixés (*rabatteur*, *plombier*, *faussaire*), à hauteur de 64%. On retrouve aussi un nombre non négligeable de noms simples (*proxénète*, *héros*), de complexes non construits (*cuistot*, *ivrogne*) et de noms converts (*criminel*, *débauché*). Les six noms annotés comme indéterminés sont des noms en relation de conversion avec un lexème non nominal, comme *assassin* avec le verbe *assassiner* ou *voyou* avec l'adjectif *voyou*, mais dont la direction ne peut pas être identifiée (Tribout 2010). L'indétermination se joue donc ici entre les types morphologiques de noms converts et de noms simples. Parmi les trois voisins composés, on trouve deux composés néoclassiques (Amiot et Dal 2008) (*photographe*, *ventriloque*) et un composé V-N (*vaurien*). Enfin, l'unique nom construit par un procédé extragrammatical, *indic*, est le fruit d'une troncation[7] à part de *indicateur*, lui-même étant un nom suffixé déverbal en *-eur*. Cela suggère donc que les noms d'agent prototypiques ne sont pas nécessairement construits, et lorsqu'ils le sont, qu'ils ne sont pas tous construits à partir de verbes.

Notons que parmi les 64 voisins affixés, 63 sont suffixés, et seul un nom est préfixé, *contremaître*. La préfixation, et tout particulièrement le préfixe *contre-*, est assez peu représentée dans nos données, y compris dans les voisins présentées dans la partie III, et n'est pas un préfixe agentif, contrairement aux autres suffixes impliqués et présentés dans le tableau 7.3.

-eur	*-ier*	*-iste*	*-ard*	*-on*	*-aire*	*-ien*
33	16	7	3	2	1	1

TABLE 7.3 – Suffixes des voisins dérivés du barycentre des noms d'agent en *-eur*

7. La troncation, et plus largement les phénomènes de réduction sont au même titre que la réduplication ou la formation de mots-valises et d'acronymes (entre autres) sont communément considérés comme des procédés extragrammaticaux dans la mesure où, malgré leur régularité, ils enfreignent des principes de la grammaire (Fradin *et al.* (2009)).

Le tableau 7.3 montre que le suffixe *-eur* est le plus représenté, à raison de 33 voisins (soit 52% des voisins suffixés), dont 27 correspondent à des amorces du barycentre. Les 6 autres voisins en *-eur* présentent des caractéristiques différentes des amorces : ils ont une base nominale (*camionneur*, *farceur*) ou une base verbale liée (*prestidigitateur*, *délateur*, *imposteur*) [8], ou présentent une double lecture agentive et instrumentale (*rabatteur*) – bien que le sens agentif semble très majoritaire dans le corpus *Wikipedia2018*.

Les deux autres suffixes les plus représentés sont le suffixe *-ier* (*plombier*, *bijoutier*), à hauteur de 25%, et le suffixe *-iste* (*machiniste*, *marionnettiste*), à hauteur de 11%. On retrouve enfin de façon plus marginale les suffixes *-ard* (*motard*), *-on* (*forgeron*), *-aire* (*faussaire*) et *-ien* (*magicien*). Comme suggéré précédemment, ces résultats confirment que les noms d'agent, lorsqu'ils sont dérivés, peuvent être construits par une variété de procédés morphologiques, pas tous déverbaux. Cela pose la question de l'équivalence de ces différents procédés pour la population de la catégorie lexicale des noms d'agent, à laquelle nous nous proposons de répondre en section 8.

Nous évaluons dans un second temps l'importance de la présence d'une base verbale pour la construction des noms d'agent en analysant les catégories grammaticales des bases des 100 plus proches voisins du barycentre des noms d'agent en *-eur* (tableau 7.4). Lorsque la catégorie grammaticale de la base est ambigue, nous alignons l'analyse sur les schémas de formation les plus représentés, si cela est cohérent avec l'analyse sémantique de la construction morphologique. Par exemple, nous considérons que *fêtard* est dérivé du verbe *fêter* et pas du nom *fête* dans la mesure où il y a peu, voire aucun, noms en *-ard* dérivés de noms dénotant des actions ou événéments en français, contrairement aux noms déverbaux en *-ard* (*fuyard*, *vantard*, *pleurnichard*) (Dubois et Dubois-Charlier 1999).

Verbe	Nom	Adjectif
37	29	6

TABLE 7.4 – Catégories grammaticales de la base des voisins dérivés du barycentre des noms d'agent en *-eur*

8. Nous considérons pour la suite de l'analyse que ces noms sont dérivés car leur base peut être identifiée dans d'autres noms morphologiquement analysables. Ainsi, *prestidigitateur*, *délateur* et *imposteur* sont morphosémantiquement en accord avec les noms d'agent déverbaux en *-eur*, et leur base potentielle peut être identifiée dans les noms *prestidigitation*, *délation* et *imposture*, qui sont eux même morphosémantiquement en accord avec les noms d'action déverbaux en *-ion* et *-ure*. Ils ne sont cependant pas pris en tant qu'amorces du fait de leur caractère non prototypique.

Comme le montre le tableau 7.4, les bases des 72 voisins dérivés (c'est-à-dire des 64 noms affixés et des 8 noms converts) sont principalement verbales (à hauteur de 51%) et nominales (à hauteur de 40%). Les bases adjectivales sont relativement marginales, et se retrouvent exclusivement pour les voisins converts (*criminel*, *sadique*, *alcoolique*). Si les bases verbales sont majoritaires, les bases nominales sont donc néanmoins bien représentées, confirmant que les noms d'agent ne sont pas nécessairement déverbaux, et que la composante actionnelle présente dans les noms d'agent n'est pas nécessairement héritée d'un verbe.

Pour comprendre comment se construit l'agentivité indépendamment de la catégorie de la base, nous procédons à l'analyse sémantique des bases des voisins dérivés. Les adjectifs sont analysés comme dénotant des propriétés. Les verbes sont identifiés comme dénotant des actions ou des propriétés, en fonction de leur dynamicité ou de leur stativité [9]. Comme les noms relèvent de types sémantiques plus divers (action, objet, propriété, domaine [10], institution), leur annotation repose sur une variété de tests linguistiques tirés de la littérature (Godard et Jayez 1993, Flaux et Van de Velde 2000, Huyghe 2015, entre autres), que nous présentons ci-dessous.

— Les bases dénotant des **actions** peuvent être le sujet de *avoir lieu* ou *se produire*, ou comme l'objet des verbes *effectuer*, *accomplir*, ou *procéder*.

— Les bases dénotant des **objets** peuvent être le sujet de *se trouver* suivi d'un locatif spatial

— Les bases dénotant des **propriétés** pouvent être utilisées dans *être d'un grand N*, *état de N*, ou pouvent être l'objet de syntagmes verbaux comme *ressentir*, *éprouver*, ou *faire preuve de*.

— Les bases dénotant des **domaines** sont compatibles avec des verbes supports comme *faire du N*.

— Les bases dénotant des **institutions** peuvent être sujet de *être fondé* suivi d'un locatif temporel, ou être utilisées dans des expressions telles que *être nommé à la tête du N*.

Dans le cas où une base est polysémique ou homonymique, nous annotons le sens qui correspond le plus à celui du dérivé, tant que cela est en accord avec les propriétés sémantiques du procédé morphologique impliqué. Par exemple, nous considérons que *farceur* 'personne qui dit ou fait des choses bouffonnes'

9. Contrairement aux verbes dynamiques, les verbes statifs ne sont pas compatibles avec la forme progressive *est en train de* (Haas *et al.* 2008).

10. Nous définissons par domaine tout ensemble de connaissances, d'expertises et de pratiques relevant d'un champ disciplinaire donné qui répond à un besoin de théorisation (Namer et Villoing 2015)

dérive de *farce*$_1$ 'plaisanterie faite à quelqu'un' plutôt que de *farce*$_2$ 'hachis d'ingrédients' sur la base de la correspondance sémantique et sur le fait que les noms dénominaux en *-eur* peuvent être construit à partir de noms d'action, à l'image de *bienfaiteur*, ou *navetteur*. Le résultat de l'annotation pour les 72 voisins dérivés du barycentre des noms d'agent en *-eur* est donné dans le tableau 7.5.

Action	Objet	Propriété	Domaine	Institution
43	15	6	6	2

TABLE 7.5 – Types sémantiques de la base des 72 voisins dérivés du barycentre des noms d'agent déverbaux prototypiques en *-eur*

Le tableau 7.5 montre que les bases des voisins dérivés sont sémantiquement très hétérogènes. Elles dénotent majoritairement (60%) des actions, qu'il s'agisse de verbes (*soigner* → *soigneur*) ou de noms (*course* → *coursier*). On retrouve aussi à hauteur de 21% des bases dénotant des objets, comme dans le cas de *machiniste*, construit sur *machine*. Enfin, les bases dénotant des propriétés (*sadique* → *sadique*) ou des domaines (*couture* → *couturier*) sont plus marginales (8% pour chaque type), et les bases institutionnelles sont rares (*police* → *policier*).

Les données du tableau 7.5 confirment d'une part que le trait actionnel des noms d'agent peut être hérité d'un nom et non d'un verbe, et d'autre part, la construction de noms d'agent ne repose pas nécessairement sur des bases actionnelles. Ainsi, 40% des voisins dérivés ne sont pas dérivés d'une base dénotant une action. Dans le cas où la base dénote un objet, une propriété, un domaine ou une institution, la composante actionnelle se construit dans la relation prédicative entre l'action et le nom dénoté, que cette relation relève de la production ou de la manipulation pour l'objet, la pratique pour le domaine, ou l'appartenance pour l'institution.

7.3.2 Impact de la polysémie des noms d'agent

Nous cherchons à évaluer l'impact du traitement fin des données lexicales. Les amorces utilisées dans cette partie constituent un sous-ensemble des amorces utilisées dans la partie III, et ce sous-ensemble a fait l'objet d'un contrôle strict relativement à la monosémie des noms. La comparaison des barycentres obtenus à partir de ces deux ensembles d'amorces nous permet d'appréhender, du moins partiellement [11], l'impact de la polysémie des noms

11. La comparaison est à nuancer au regard des différences de constitution des barycentres. Le barycentre des noms d'agent pomysémiques sont calculés à partir d'un corpus

d'agent en *-eur*, et plus largement de la suffixation déverbale en *-eur*, sur notre représentation.

La comparaison des deux listes de voisins fait émerger 18 voisins communs (23). Ces voisins sont tous des noms d'agent monosémiques, à trois exceptions près (*chien*, *rabatteur*, *coursier*).

(23) plombier, machiniste, rabatteur, armurier, bricoleur, garagiste, déménageur, coiffeur, cuistot, chien, vendeur, coursier, cuisinier, contremaître

Une fois ces 18 voisins communs exclus, l'analyse du voisinage exclusif au barycentre construit à partir des noms polysémiques de la partie III met en avant le caractère fortement instrumental de ce voisinage. On y retrouve en effet un nombre important de noms d'instrument (*aspirateur*, *lance-pierre*), ou de noms polysémiques impliquant une lecture instrumentale (*nettoyeur*, *mouchard*). Globalement, on retrouve une plus grande proportion de noms d'agent polysémiques dans ce voisinage, alors que les voisins du barycentre construit à partir de noms monosémiques sont dans la quasi totalité des noms eux-même monosémiques agentifs (*vétérinaire*, *masseur*, *photographe*, *contrebandier*).

Une remarque supplémentaire peut être faite concernant les noms d'agent des deux listes. En effet, la comparaison des noms d'agent fait émerger une différence axiologique. Plus précisément, les actions réalisées par les agents désignés par les voisins du barycentre des noms d'agent monosémiques tendent à être plus négatives (24a), alors que les actions dénotées dans le cas des voisins de l'autre barycentre tendent à être plus neutres ((24b)). Il faudrait pouvoir évaluer la représentation de ce trait axiologique parmi les amorces monosémiques, mais on peut faire l'hypothèse que la suppression des noms d'instrument parmi les amorces du second barycentre révèle le trait axiologique négatif présent parmi un certain nombre de noms d'agent.

(24) a. pickpocket, charlatan, voleur, voyou, cambrioleur, receleur, assassin, criminel, drogué, contrebandier, faussaire, ivrogne, arnaqueur, meurtrier, dealer, fraudeur

b. plombier, machiniste, conducteur, mécano, technicien, installateur, élagueur, soudeur, gabier, magasinier, cuistot, vendeur, opérateur, mécanicien, coursier, cuisinier, contremaître

Cette analyse ne vaut donc qu'à titre indicatif, mais suggère néanmoins l'importance de la sélection des données. Cette comparaison mériterait d'être reprise et approfondie à partir de barycentres calculés dans un même modèle,

lemmatisé, alors que celui des noms d'agent monosémiques est calculé à partir d'un corpus lemmatisé et étiqueté, ce qui tend donc à réduire le nombre de contextes utilisés et les représentations.

à la fois pour conforter ce résultat que pour permettre de caractériser en creux les noms d'instrument relativement au noms d'agent monosémiques.

7.4 Limites et extension de la classe

Dans la section 7.3.1, nous avons caractérisé les propriétés morphosémantiques des noms d'agent sur la base de l'analyse des voisins du barycentre construit à partir des noms d'agent déverbaux monosémiques en *-eur*. Cette analyse repose cependant sur l'hypothèse que les voisins du barycentre des noms d'agent sont bien eux-mêmes des noms d'agent. On peut cependant se demander si tous les voisins sont bien des noms d'agent. Nous avons signalé en section 7.2.1 que la plupart des amorces utilisées pour construire le barycentre dénotaient des agents humains, qui combinaient donc les traits sémantiques [+agent] et [+humain]. On peut donc se demander si les voisins sont bien porteurs des deux traits, et si les voisins désignent bien leur référent sur la base de la réalisation d'une action. En d'autres termes, on peut s'interroger sur l'influence du trait humain présent dans les amorces sur les voisins.

Ainsi, un examen plus approfondi de la liste des voisins fait émerger des résultats inattendus. Par exemple, on trouve dans cette liste des converts de participe passé comme *drogué*, au rang 33, et *travesti* au rang 78. On s'attend plutôt à ce que de tels converts soient des patients plutôt que des noms d'agent, puisqu'ils correspondent traditionnellement à l'argument interne de verbes transitifs (*condamné*, *blessé*, *invité*). Leur présence dans le proche voisin du barycentre des noms d'agent en *-eur* est néanmoins sensé si on considère que les référents de *drogué* et *travesti* se définissent plutôt comme des effectuateurs, étant le sujet du verbe de base dans des constructions réflexives (exemples (25a) et (26a)) et non l'objet dans des constructions transivitives (exemples (25b) et (26b)). On peut néanmoins plus largement s'interroger sur la nature agentive des voisins du barycentre.

(25) a. Un drogué est quelqu'un qui se drogue

b. ?Un drogué est quelqu'un que l'on drogue

(26) a. Un travesti est quelqu'un qui se travestit

b. ?Un travesti est quelqu'un que l'on travestit

Plus généralement, dans quelle mesure la classe lexicale des noms d'agent est distributionnellement bien délimitée, et notamment distincte de celle des noms d'humain ? Nous tentons dans la section 7.4.1 la proximité distributionnelle entre les noms d'agent et les noms d'humain non agentifs. Du fait

de l'absence de trait agentif dans la structure des noms d'humain non agentifs, et sur la base du constat que les noms d'agent intègrent à la fois le trait agentif et le trait humain, nous faisons l'hypothèse qu'ils seront relativement distants des noms d'agent. Une fois la classe délimitée, nous explorons l'extension de la catégorie en évaluant sur le plan distributionnel le potentiel agentif de noms candidats par le biais de leur proximité aux noms d'agent et aux noms d'humain (section 7.4.2).

7.4.1 Influences du trait humain

Nous souhaitons évaluer la sensibilité au trait humain du barycentre des noms d'agent déverbaux prototypiques en *-eur*. Plus précisément, nous vérifions qu'il y a bien une différence distributionnelle entre les noms d'agent et les noms d'humain, et faisons ainsi émerger le trait agentif. Nous calculons pour cela la proximité du barycentre des noms d'agent avec les noms d'humain qui ne dénotent pas des agents, afin de montrer par constraste la prédominance du trait agentif chez les noms d'agent.

Nous nous basons pour cela sur trois grandes catégories de noms d'humain exempts du trait agentif, à savoir les noms phasiques et généraux (*personne*, *vieillard*), les noms relationnels (*fils*, *otage*) et les gentilés (*fidjien*, *genevois*). Les noms généraux et phasiques dénotent des êtres humains respectivement sans spécification supplémentaire (Halliday et Hasan 1976, Mahlberg 2005), ou avec uniquement une indication d'âge (Aleksandrova 2013). Certains d'entre eux peuvent être vus comme des hyperonymes d'autres noms d'humain : *homme* peut ainsi être vu comme l'hyperonyme du nom de phase *adolescent*, mais aussi du nom relationnel *frère*. Nous les groupons ensemble du fait de la frontière floue entre les deux groupes, principalement due à l'absence de critères pour délimiter la classe des noms généraux. Les noms relationnels dénotent quant à eux des êtres humains dans une relation interpersonnelle avec d'autres humains, et sont compatibles avec des génitifs dénotant des personnes liées (Vikner et Jensen 2002, Partee et Borschev 2003, Barker 2008, entre autres). Les gentilés dénotent des habitants et sont dérivés de noms de lieux. Nous compilons les listes de noms d'humain non agentifs à partir de deux ressources existantes : la base de données Humanymes [12] pour les noms généraux, phasiques et relationnels, et Prolexbase [13] pour les gentilés. Les listes sont filtrées manuellement afin de retirer les noms polysémiques [14], et de compléter avec une sélection de synonymes tirés du *Dic-*

12. https ://humanymes.u-strasbg.fr/.

13. https ://www.cnrtl.fr/lexiques/prolex/.

14. Est exclu tout nom ne relevant pas strictement d'une de ces trois catégories. À ce titre, les noms *demoiselle* et *femme* ne sont par exemple pas conservés car ils correspondent

tionnaire Electronique des Synonymes édité par le laboratoire CRISCO [15]. À l'issue de cette sélection, 46 noms généraux et phasiques, 84 noms relationnels et 18 080 gentilés sont conservés. Du fait du seuil de fréquence imposé par les modèles distributionnels utilisés, seuls 39, 65 et 380 de ces noms (484 au total) sont utilisés par la suite.

Une première façon d'évaluer l'impact des traits humain et agentif sur le barycentre des noms d'agent en *-eur* est d'estimer la proximité de noms d'humain à ce barycentre. Dans la mesure où nous faisons l'hypothèse que les noms d'humain ne possèdent pas le trait agentif, nous nous attendons à ce que les noms d'humain soient relativement distants du barycentre des noms d'agent. Nous calculons le rang et le score de proximité moyen des 484 noms d'humain sélectionnés précedemment parmi le voisinage du barycentre des noms d'agent en *-eur*. Les 10 premiers noms de chaque classe à apparaître dans le voisinage sont donnés dans les tableaux 7.6, 7.7 et 7.8. Sont renseignés avec les noms [16] les scores de proximité au barycentre ainsi que leur rang parmi les plus proches voisins.

Nom général/phasique	Proximité	Rang
homme	0.616	118
adolescent	0.577	240
gars	0.566	281
gens	0.551	359
bambin	0.550	365
quadragénaire	0.531	521
vieillard	0.520	604
sexagénaire	0.519	622
garçonnet	0.493	932
quinquagénaire	0.473	1194

TABLE 7.6 – Score de proximité et rang des dix noms d'humain généraux et phasique les plus proches du barycentre des noms d'agent monosémiques en *-eur*

Les tableaux 7.6, 7.7 et 7.8 montrent que les scores de proximité des

respectivement à un nom d'humain général et un animal, et à un nom d'humain général et un nom d'humain relationnel.

15. https ://crisco2.unicaen.fr/des/.

16. Pour des raisons de lisibilité, les noms ne sont pas fournis avec leur étiquette grammaticale. L'interrogation porte cependant sur les noms étiquetés comme nom commun. La proximité du nom *homme* au barycentre est donc évaluée à l'aide du vecteur de *NC :homme*.

Nom relationnel	Proximité	Rang
copain	0.591	191
amant	0.579	231
invité	0.549	376
compagnon	0.541	427
camarade	0.528	544
colocataire	0.526	563
fiancé	0.512	686
père	0.492	936
rejeton	0.486	1016
frère	0.485	1018

TABLE 7.7 – Score de proximité et rang des dix noms d'humain relationnels les plus proches du barycentre des noms d'agent monosémiques en *-eur*

10 premiers noms de chaque classe sont bien plus bas que ceux des 100 premiers voisins du barycentre, puisque le 100^{e} voisin du barycentre des noms d'agent affiche un score de proximité de 0.625. Les noms généraux, phasiques, relationnels et les gentilés apparaissent globalement loin dans le voisinage [17].

L'analyse plus globale des noms d'humain confirme leur grande distance au barycentre, avec des scores de proximité moyens respectifs de 0.388, 0.348 et 0.131 pour les noms généraux ou phasiques, les noms relationnels et les gentilés. Pour conforter cette observation, nous observons la distribution des noms d'humain dans le voisinage du barycentre. Plus précisément, nous quantifions la proportion de noms apparaissant après le 5 500^{e} rang, nombre que nous fixons arbitrairement pour être très élevé et ainsi traduire une forte distance au barycentre. On constate ainsi que 39% des noms généraux phasiques apparaîssent après le 5 500^{e} rang, contre 45% pour les noms relationnels, et 97% pour les gentilés. Ces derniers sont donc les noms d'humain les plus distants des noms d'agent. La proximité relativement plus élevée des noms d'humain généraux ou phasiques peut sans doute s'expliquer par leur usage en corpus comme des hyperonymes de noms d'agent.

Le fait que les noms d'humain non agentifs n'apparaissent pas parmi les 100 plus proches voisins du barycentre des noms d'agent en *-eur*, contrairement aux noms clairement agentifs (dans le cas des 27 amorces présentes dans

17. Le nom général *homme* est une exception, puisqu'il apparaît au 118e rang, avec un score moyen de proximité de 0.616. Cette proximité relative pourrait s'expliquer par le fait que *homme* est un hyperonyme de tous les autres noms d'humain et par le fait qu'il appartient à plusieurs expressions polylexicales qui dénotent des agents comme *homme de main* ou *homme d'entretien*.

Gentilé	Proximité	Rang
Cauchois	0.434	1981
Tourquennois	0.422	2314
béké	0.410	2732
Berlinois	0.397	3278
Asiatique	0.397	3300
Niçois	0.391	3568
Ardennais	0.369	4925
Véronais	0.367	5091
Lorientais	0.367	5093
Toulousain	0.363	5441

TABLE 7.8 – Score de proximité et rang des dix gentilés les plus proches du barycentre des noms d'agent monosémiques en *-eur*

les 100 premiers voisins), soutient l'hypothèse que le barycentre des noms d'agent déverbaux monosémiques en *-eur* est déterminé par le trait agentif, et non exclusivement humain, des amorces. Dès lors, nous pouvons envisager les voisins analysés dans la section 7.3 comme un échantillon représentatif de noms d'agent puisqu'ils sont particulièrement proches du barycentre. Nous en concluons donc, sur la base des analyses faites en section 7.3, que les noms d'agent français sont morphologiquement divers concernant leur dérivation, et la sélection des affixes et des bases. L'hétérogénéité morphologique des tout premiers voisins du barycentre des noms d'agent en *-eur* présentés dans le tableau 7.1 va dans le sens de cette conclusion.

7.4.2 Candidats à l'agentivité

Nous utilisons à présent le barycentre et ses voisins d'une autre manière, comme étalon pour mesurer l'agentivité d'un mot. Nous exploitons pour cela l'hypothèse que la proximité d'un nom au barycentre des noms d'agent déverbaux prototypiques en *-eur* permet d'estimer l'agentivité des noms candidats. Nous définissons comme noms candidats des noms qui ne sont pas des noms déverbaux d'agent prototypiques en *-eur*, mais pour lesquels nous conjecturons qu'il s'agit néanmoins de noms d'agent.

Nous partons du principe que la catégorie lexicale des noms d'agent, instanciée dans les modèles distributionnels par le barycentre des noms d'agent déverbaux prototypiques en *-eur*, semble distributionnellement bien définie, et du moins distincte de la catégorie des noms d'humain non agentifs. Après avoir établi en section 7.3.1 que les noms d'agent pouvaient être morpho-

logiquement variés, nous souhaitons tester l'extension de la catégorie, sur le plan morphologique, en examinant le profil distributionnel de noms d'humain pour lesquels nous avons des soupçons d'agentivité. Pour cela, nous évaluons leur position dans l'espace vectoriel, et plus précisément vis à vis des noms d'agent déverbaux prototypiques en *-eur*, mais aussi des noms d'humain précédemment étudiés. Au même titre que les noms d'agent prototypiques en *-eur*, nous choisissons de représenter les 3 catégories de noms d'humain précédemment étudiées par des barycentres construits selon la même approche. Cela nous permet de comparer la proximité des noms candidats à ces quatre barycentres, et ainsi d'établir le trait sémantique proéminent (agentif ou humain) de chacun de ces noms. Nous nous attendons ainsi à ce qu'un nom d'humain agentif, indépendamment de son profil morphologique, soit plus proche du barycentre des noms d'agent que des barycentres des noms d'humain, et réciproquement, à ce qu'un nom non agentif soit plus proche d'un des barycentres de noms d'humain que de celui des noms d'agent.

Nous envisageons cette méthode comme un indicateur d'agentivité, mais nous n'affirmons pas qu'elle a valeur de test pour diagnostiquer l'agentivité d'un item lexical donné. En analysant des échantillons d'items, nous visons des inférences générales concernant certains types morphologiques, et non une caractérisation individuelle des noms, dont nous avons vu dans la partie II les limites dans l'utilisation des espaces vectoriels. En effet, la plus grande proximité d'un mot donné au barycentre des noms d'agent par comparaison aux barycentres des noms d'humain est certainement un indice des propriétés sémantiques agentives du mot, mais pas une condition permettant la qualification en tant que nom d'agent, et réciproquement. Si la tendance générale indique clairement que la plupart des noms d'agent en *-eur* sont plus proches du barycentre agentif que des barycentres non agentifs, on observe quelques aberrations, à l'image de noms d'agent déverbaux en *-eur* comme *guetteur*, *randonneur*, et *sauveur*, qui sont plus proches des barycentres de noms d'humain que de celui des noms d'agent en *-eur*. Cela ne remet cependant pas leur agentivité en question. *A contrario*, certains noms non agentifs (*Ardennais*, *gus*, *zig*) sont plus proches du barycentre des noms d'agent en *-eur* que des barycentres de noms d'humain, sans que l'on puisse leur attribuer un trait agentif. Cette méthode nous permet néanmoins de faire ressortir des tendances à l'agentivité à l'échelle de types morphologiques.

Pour construire ces barycentres, nous utilisons les listes de noms d'humain non agentifs monosémiques présentées en section 7.4.1. Aussi extensive soient-elles, ces listes varient significativement en taille. La liste la plus petite, c'est-à-dire celle des noms généraux et phasiques, contient 39 items, alors que la liste initiale des noms d'agent en *-eur* en contient 681. Pour neutraliser une possible source de biais liée à ces disparités, nous sous-échantillonnons les

listes de façon à avoir le même nombre d'amorces pour chaque barycentre, basé par conséquence sur le plus petit groupe. Étant donné que les fréquences des noms d'humain varient aussi de façon considérable, nous divisons chacune des trois listes de noms d'humain en 39 quantiles par rapport aux fréquences, et nous sélectionnons aléatoirement un item de chaque groupe de sorte que les fréquences soient distribuées de façon similaire dans chaque liste d'amorces.

Nous sélectionnons ensuite aléatoirement, pour chaque type morphologique que nous testons, 20 candidats à l'agentivité. Les types morphologiques testés sont les suivants :

— noms dénominaux suffixés en *-eur* (*autostoppeur*, *basketteur*, *précepteur*)

— noms suffixés en *-aire* (*bibliothécaire*, *gestionnaire*, *plagiaire*)

— noms suffixés en *-iste* (*pianiste*, *éclairagiste*, *exorciste*)

— noms suffixés en *-ier* (*braconnier*, *caissier*, *luthier*)

— noms suffixés en *-ant* (*combattant*, *manifestant*, *surveillant*)

— noms en relation de conversion avec un verbe, mais indéterminés quant à la direction de la conversion (*arbitre*, *pèlerin*, *pilote*)

— noms morphologiquement simples (*médecin*, *architecte*, *scribe*)

Les candidats sont extraits de différentes sources : la base Lexeur pour les noms dénominaux en *-eur*, Lexique3 pour les noms en *-aire*, *-ant*, *-iste*, *-ien* et *-ier*, (Tribout 2010) pour les noms en relation de conversion avec un verbe, et (Tribout *et al.* 2014) pour les noms morphologiquement simples. Nous nous assurons que les noms sélectionnés sont des noms d'humain monosémiques et qu'ils ne font pas partie des 100 plus proches voisins du barycentre des noms d'agent en *-eur* précédemment analysés, afin de ne pas biaiser la comparaison [18]. Nous limitons notre sélection aléatoire à 20 items par type morphologique car nous ne visons pas l'exhaustivité, mais plutôt un aperçu comparatif qui requiert donc un nombre fixe de candidats pour chaque type. Tous les types morphologiques n'étant pas représentés de façon similaire dans le corpus, nous basons notre échantillon sur le type le moins représenté – ici les noms simples, pour lesquels les candidats monosémiques à l'agentivité avec une fréquence supérieure ou égale à 5 dans le corpus *Wikipedia2018* ne

18. Cette contrainte est ici pensée pour tester des noms dont nous n'avons pas déjà *a priori* d'information quant à l'agentivité. Par ailleurs, les noms d'humain arrivant assez loin parmi les voisins du barycentre des noms d'agent, l'absence du nom candidat des 100 premiers voisins du barycentre des noms d'agent n'exclut en rien qu'il soit plus proche du barycentre des noms d'agent que de celui des noms d'humain. Cette contrainte ne constitue donc pas à nos yeux un réel biais, d'autant que nous cherchons simplement à faire émerger des tendances.

Candidat	Ag	GP	Rel	Gent
basketteur	**0.336**	0.151	0.198	0.175
précepteur	0.412	0.444	**0.676**	0.044
bibliothécaire	**0.403**	0.260	0.341	0.168
gestionnaire	**0.366**	0.121	0.072	0.098
pianiste	0.332	0.244	**0.408**	0.048
exorciste	**0.490**	0.463	0.376	0.055
chirurgien	**0.559**	0.418	0.408	0.065
historien	**0.477**	0.265	0.294	0.027
braconnier	0.542	**0.566**	0.355	0.128
horloger	**0.518**	0.369	0.348	0.069
manifestant	0.332	**0.385**	0.087	0.059
combattant	**0.600**	0.428	0.256	0.330
arbitre	**0.483**	0.274	0.152	0.367
pilote	**0.601**	0.229	0.246	0.395
médecin	**0.561**	0.505	0.485	0.030
architecte	**0.387**	0.119	0.245	0.005

TABLE 7.9 – Scores de proximité de 16 noms candidats aux barycentres des noms d'agent en *-eur* (Ag), des noms d'humain généraux et phasiques (GP), des noms relationnels (Rel) et des gentilés (Gent)

sont pas nombreux. Le tableau 7.9 présente un échantillon des scores moyens de proximité calculés pour les candidats à l'agentivité avec les quatre barycentres. Le score de proximité le plus élevé pour chaque candidat (en ligne) est indiqué en gras.

La majorité des candidats présentés dans le tableau 7.9 sont plus proches du barycentre des noms d'agent que des autres barycentres, à quatre exceptions près. La proximité de *précepteur* au barycentre des noms d'humain relationnel peut s'expliquer par le fait que l'on est le précepteur de quelqu'un, et donc le référent est décrit en relation par rapport à quelqu'un. Dans le cas de *braconnier* et *manifestant*, les scores de proximité tendent à indiquer qu'ils sont quasi équidistants des noms agents et des noms d'humain généraux et phasiques. De même, la plus grande proximité des noms *médecin* et *exorciste* au barycentre des noms d'agent n'est pas très marquée. Cela semble confirmer que l'analyse d'items à l'aune de ces scores de proximité ne semble pas être le niveau de granularité le plus adapté (voir chapitre 4).

Nous favorisons l'analyse de ces proximités aux barycentres des noms d'agent et d'humain à l'échelle des types morphologiques. Les résultats globaux pour chaque type morphologique sont donnés dans le tableau 7.10. La

Type morphologique	Ag	GP	Rel	Gent
Dénominaux en *-eur*	**13 (65%)**	5 (25%)	-	2 (10%)
Suffixés en *-aire*	**19 (95%)**	-	1 (5%)	-
Suffixés en *-iste*	**16 (80%)**	3 (15%)	1 (5%)	-
Suffixés en *-ien*	**16 (80%)**	4 (20%)	-	-
Suffixés en *-ier*	**14 (70%)**	5 (25%)	1 (5%)	-
Suffixés en *-ant*	**13 (65%)**	4 (20%)	2 (10%)	1 (5%)
Conversion non marquée	**9 (45%)**	8 (40%)	3 (15%)	-
Simples	**10 (50%)**	9 (45%)	1 (5%)	-

TABLE 7.10 – Propension des noms candidats à l'agentivité en fonction de leur proximité aux barycentres des noms d'agent en *-eur* (Ag), des noms d'humain généraux et phasiques (GP), des noms relationnels (Rel) et des gentilés (Gent)

proximité à un barycentre donné la plus représentée pour chaque type morphologique (en ligne) est indiquée en gras.

Il apparaît dans le tableau 7.10 que les noms candidats testés tendent à être plus proches du barycentre des noms d'agent que des autres barycentres, et ce quel que soit leur type morphologique. On observe cependant des disparités en fonction des types morphologiques. Ainsi, la totalité des noms en *-aire* rentrent dans ce schéma, à l'exception de *commanditaire* qui est plus proche du barycentre des noms d'humain relationnels. *A contrario*, seuls 9 des 20 noms converts testés sont plus proches du barycentre des noms d'agent que des autres barycentres, ce qui représente certes la plus grande proportion, mais pas la majorité. Globalement, le deuxième type le plus représenté est celui des noms d'agent généraux et phasiques.

Ces tendances renforcent l'affirmation qu'une grande variété de constructions morphologiques peuvent être la base des noms d'agent, et que d'une façon générale, les noms d'agent en français ne sont pas limités aux noms déverbaux se finissant en *-eur*. Les résultats suggèrent cependant que certains types morphologiques présentent un degré d'agentivité variable. Les noms converts et les noms simples semblent afficher une plus faible propension à l'agentivité, alors que les noms suffixés en *-aire*, *-iste* et *-ien* affichent une plus grande propension à l'agentivité.

Les résultats présentés dans ce chapitre corroborent ceux suggérés avec l'expérience de la partie III, à savoir que la catégorie lexicale des noms d'agent intègre des profils morphologiques très variés, qu'il s'agisse de noms construits morphologiquement par affixation, par conversion ou composition,

de noms complexes non construits ou encore de noms simples. Lorsque les noms d'agent sont construits, il n'y a pas de contrainte particulière sur la catégorie grammaticale ou le type sémantique de la base. Cela suggère que la prédication n'est pas pas toujours héritée de la base, et qu'elle se construit au sein même de la structure du nom d'agent. Une contribution majeure de cette expérience n'est pas tant la confirmation de cette diversité que son évaluation quantitative. Ces variations – tant de types morphologiques que de types grammaticaux ou sémantiques de bases – s'instancient de façon non négligeable dans le voisinage du barycentre des noms d'agent prototypiques en *-eur*, ce qui suggère que cette diversité n'est pas marginale, et qu'elle contribue largement au peuplement de la catégorie des noms d'agent.

La diversité morphologique des noms d'agent pose cependant la question de leur équivalence. On peut ainsi se demander si la classe des noms d'agent est homogène sémantiquement, et dans quelle mesure la coexistence de plusieurs procédés de formation des noms d'agent contribue ou non à son hétérogénéité. Nous explorons cette question dans le chapitre 8.

Chapitre 8

Conccurence affixale

Nous avons vu dans la section 7.4 que la catégorie lexicale des noms d'agent en français ne se limite pas aux noms déverbaux en *-eur*, mais qu'elle contient d'autres types de noms. Cela inclut des noms dérivés – déverbaux (*arriviste*) ou dénominaux (*garagiste*) –, des noms composés (*aide-soignant*), des noms en relation de conversion avec un verbe (*juge*) ou un adjectif (*criminel*), des noms tronqués (*indic*), et/ou réduplicatifs (*nounou*), des mots-valises (*entreprenaute*, à partir d'*entrepreneur* et *internaute*), des acronymes (*dircab*), des complexes non construits (*ivrogne*), et des noms morphologiquement simples (*médecin*). Cette diversité morphologique se retrouve aussi au sein des types morphologiques, et notamment au sein des noms suffixés, pour lesquels plusieurs suffixes co-existent, dont *-aire*, *-ant*, *-eur*, *-ien*, *-ier*, *-iste*, et plus marginalement *-on* et *-ard*.

On peut se demander si cette diversité morphologique est corrélée à des différences sémantiques fines. Plus précisément, on peut faire l'hypothèse que l'existence de tous ces suffixes, dont certains sont concurrents dans la mesure où ils exploitent les mêmes bases, produit des noms d'agent qui vont présenter des spécificités sémantiques et/ou ontologiques, expliquant l'existence simultanée de ces suffixes, et qui va dessiner une organisation particulière des noms d'agent. Réciproquement, on peut se demander si l'existence de différents types de noms d'agent est corrélée à ces différentes suffixations. La conccurrence entre les suffixes dits agentifs, c'est-à-dire permettant la formation de noms d'agent, a notamment été abordée au regard de critères formels et phonologiques (Roché 1997, Lignon 2000) ou sémantiques (Cartoni *et al.* 2015), mais à notre connaissance pas dans une approche extensive.

Nous travaillons cette question à l'échelle des suffixations dont on peut faire l'hypothèse qu'elles définissent des catégories particulières. Plus précisément, on s'intéresse aux suffixes *-ant*, *-aire*, *-ien*, *-ier*, *-iste*, en comparaison

avec le suffixe *-eur*[1]. Nous évaluons dans quelle mesure elles sont concurrentes, et dans quelle mesure leurs spécificités sémantiques dessinent des catégories lexicales et/ou ontologiques distinctes.

La comparaison des voisinages des barycentres construits à partir des différentes suffixations fait émerger des propriétés distinctives, qui permettent de rapprocher sur des bases morphosémantiques certains suffixes, à l'image de *-ien* et *-iste* qui favorisent des bases dénotant des domaines, ou *-aire*, *-ant* et *-eur* qui favorisent des bases dénotant des actions. Nous montrons que ces propriétés morphosémantiques sont corrélées à une caractérisation ontologique des noms d'agent.

Pour ce faire, nous commençons par présenter la sélection et les caractéristiques des noms d'agent que nous allons considérer (section 8.2), puis nous comparons le profil distributionnel des différentes suffixations pour en faire émerger les différences et les similarités (section 8.3). Enfin, nous abordons la question des différences sémantiques entre ces suffixations dans une approche bottom-up à l'aide du clustering (section 8.4).

8.1 Diversité des suffixes agentifs

L'existence de plusieurs affixes permettant la formation de noms d'agent s'observe dans de nombreuses langues, et remonterait aux langues indo-européennes, pour lesquelles Benveniste (1975) identifie deux constructions morphologiques des noms d'agent, les suffixes **-ter* et **-tor*. Benveniste (1975) argue que ces constructions se distinguent par la nature du lien entretenu entre l'agent dénoté et l'action (définitoire d'une part pour les noms d'agent dits statutaires, et ponctuel d'autre part pour les noms d'agent dits occasionnels[2]). Si la relation directe entre ces suffixes (conservés uniquement sous la forme *-eur* en français) et ces deux sous-types sémantiques semble avoir disparu dans les langues contemporaines, on observe toujours une grande diversité de suffixes agentifs.

Plusieurs suffixes agentifs ont été identifiés pour le français dans la littérature, dont les plus notables sont *-eur*, *-iste*, *-aire*, *-ien* et *-ier* (Dubois 1962, Anscombre 2001, Fradin et Kerleroux 2003, Roché 2003 2011, Lignon 2000, Roy et Soare 2012, Schnedecker et Aleksandrova 2016, entre autres). Le statut

1. Nous incluons ici dans la catégorie des noms en *-eur* les noms déverbaux et dénominaux car nous ne séparons pas les noms formés par les autres suffixes sur la base de la catégorie grammaticale de leur input. Les chiffres que nous présentons par la suite varient donc légèrement des chiffres présentés en dans le chapitre 7.

2. Nous discutons plus en détail cette distinction dans le chapitre 9

agentif du suffixe [3] *-ant* ne fait quant à lui pas consensus, des auteurs comme Lerat (1984), Rosenberg (2008) considérant qu'il forme des noms d'agent, contrairement à des auteurs comme Winther (1975), Anscombre (2003), Roy et Soare (2012) qui rejettent cette hypothèse. L'existence de ces différents suffixes pose inévitablement la question de leur possible concurrence [4] entre ces suffixes, comme l'illustre l'existence de doublons comme *anecdotier* et *anecdotiste*, *bibelotier* et *bibeloteur*, *théologien* et *théologiste*, *boursicoteur* et *boursicotier*, ou *exécutaire* et *exécutant*.

Des premières différences peuvent cependant déjà être mises en avant. Ainsi, si tous ces suffixes sont polysémiques (ou sont du moins associés à des règles de formation polysémiques), les types sémantiques à l'œuvre ne sont pas les mêmes. Le suffixe *-aire* forme par exemple, en plus de noms d'agent, des noms de bénéficiaire comme *allocataire* (Schnedecker et Aleksandrova 2016), le suffixe *-ant* des noms de moyen comme *désherbant* (Knittel 2017), le suffixe *-eur* des noms d'instrument comme *réfrigérateur* (Fradin et Kerleroux 2003), le suffixe *-ien* des gentilés comme *Malien* (Lignon 2000), le suffixe *-ier* des noms d'arbres fruitiers comme *cerisier* (Corbin et Corbin 1991), et le suffixe *-iste* des noms axiologiques comme *marxiste* (Roché 2011). De ce fait la concurrence entre ces suffixes n'est que partielle. La question de leur distinction subsiste néanmoins pour leur acception agentive, sur laquelle nous nous concentrons désormais.

Notons que la concurrence ne concerne pas strictement l'ensemble des 6 suffixes de façon équivalente, puisqu'ils ne sélectionnent pas tous les mêmes bases. Les suffixes *-aire*, *-eur*, *-ier* et *-iste* peuvent sélectionner des bases verbales et nominales (même si *-ier* et *-iste* favorisent les bases nominales), alors que le suffixe *-ant* n'admet que des base verbales, et le suffixe *-ien* des bases nominales. La concurrence entre certains suffixes est donc parfois inexistante (à l'image des suffixes *-ant* et *-ier*) ou partielle, comme dans le cas de *-eur* et *-iste*, puisqu'elle n'est envisageable sur sur les constructions dénominales. Ainsi, des noms comme *bouquineur* et *bouquiniste* ne peuvent pas vraiment être considérés comme concurrents puisque le premier désigne

3. La construction morphologique des noms en *-ant* a elle-même fait l'objet de discussion. Ces noms peuvent être analysés soit comme des noms converts déverbaux à partir du participe présent, soit comme des noms suffixés. L'existence de noms féminins en *-ante*, désinence absente des paradigmes flexionnels verbaux, tend à corroborer l'hypothèse d'une construction suffixale des noms en *-ant*. Nous considérons de fait dans cette étude les noms en *-ant* et -ante comme des noms construits avec le suffixe *-ant*.

4. La concurrence affixale est ici définie comme un cas de compétition morphologique impliquant deux affixes ou plus associés à des règles de construction similaires. Cette définition de la concurrence peut cependant être discutée au regard de la définition de la similarité de règles, et du degré de granularité de la comparaison. Cette discussion dépasse le cadre de ce travail. Voir Huyghe et Wauquier (umis) pour plus de détails.

le lecteur, et se construit sur le verbe *bouquiner*, alors que le second, désigne le revendeur de livres, se construit sur le nom *bouquin*.

Malgré tout, les cas de concurrence effective invitent à s'interroger sur les spécificités de chaque suffixe. Différents facteurs ont été envisagés, incluant des aspects phonologiques, diachroniques, syntaxiques et sémantiques. L'argument phonologique a ainsi été avancé par Roché (1997) dans le blocage des formes pourtant attendues **camionnier* et **avionnier*, au profit de la construction des formes *camionneur* et *avionneur*, ou par Lignon (2000) dans le blocage de **mathématiciste* au profit de *mathématicien*. Lignon (2007) met aussi en avant le critère de productivité en diachronie, montrant qu'une partie des noms de spécialiste en *-ien* ont été formés alors que le suffixe *-ien* était plus productif que le suffixe *-iste*. Sur le plan syntaxique, Roy et Soare (2012) soutiennent par exemple que la compatibilité des suffixes avec des verbes inaccusatifs distingue le suffixe *-ant*, compatible avec un verbe comme *arriver* (*arrivant*), du suffixe *-eur*, non compatible avec de tels verbes (**arriveur*).

Sur le plan sémantique enfin, des différences ont été suggérées relativement aux domaines ontologiques. Dans leur étude contrastive des noms d'agent en français et en italien, Cartoni *et al.* (2015) montrent par exemple que les suffixes *-ien* et *-iste* tendent à dénoter des spécialistes, et que les suffixes *-eur* et *-ier* privilégient les activités manuelles et commerciales. Une comparaison systématique et à grande échelle des suffixes français *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste* reste cependant à faire pour évaluer les propriétés morphosémantiques de chaque suffixe, et leur possible recouvrement. Dans la lignée de Cartoni *et al.* (2015), on peut faire l'hypothèse que les suffixes présenteront des spécificités sémantiques permettant de dessiner plusieurs sous-groupes de noms d'agent.

8.2 Noms d'agent en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste*

Pour comparer morphosémantiquement les dérivés construits par les suffixations en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste*, nous nous proposons de reprendre la méthode comparative utilisée dans le chapitre 6 pour comparer les noms d'agent et d'instrument en *-euse* et *-rice*, mais adaptée au niveau de granularité mis en place dans le chapitre 7. L'objectif est de construire des barycentres pour les catégories définies par chaque suffixe, et de comparer les voisinages de ces barycentres sur la base de leurs propriétés morphosémantiques et ontologiques.

Les suffixes *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste* permettent de construire

des noms d'agent, mais tous les noms construits par ces suffixes ne dénotent pas des agents. Nous procédons donc dans un premier temps à la sélection des noms d'agent qui serviront d'amorces (section 8.2.1) avant de proposer une première comparaison de ces noms d'agent (section 8.2.2).

8.2.1 Sélection des noms d'agent

Les amorces sont sélectionnées parmi les noms suffixés en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste*[5], en reprenant la définition ainsi que les critères utilisés en section 7.2.1, que nous complétons pour intégrer les contraintes morphologiques des divers suffixes. Notre principal critère reste la monosémie. Nous étendons cependant la condition morphologique, puisque les noms ne sont plus nécessairement dérivés de verbes. Nous conservons ainsi plus largement les noms d'agent construits morphologiquement et dont l'agentivité est issue de la suffixation, en lien avec la base. On exclut donc les noms lexicalement définis comme agentifs, mais dont l'agentivité n'est pas construite par la dérivation (*ex-bibliothécaire*, *maire*). Nous considérons ici comme des noms d'agent les noms qui dénotent une entité réalisant une action en lien avec les éléments décrits par la base, que ce lien soit basé sur l'accomplissement de l'action dénotée par la base, la production ou la manipulation de l'objet dénoté par la base, la pratique ou l'étude du domaine dénoté par la base, ou encore que la base localise l'action réalisée par l'agent dénoté.

Nous reprenons la ressource et les critères utilisés dans le chapitre 7. Les noms sont extraits de Lexique3, et complétés par les dénominaux de Lexeur pour le suffixe *-eur*. Nous filtrons manuellement les candidats afin d'exclure :

— les noms non dérivés (*peur*)

— les noms pour lesquels le lien à une base par une relation d'agentivité n'est pas perçu en synchronie (*cordonnier*)

— les noms dont la dernière étape de construction n'est pas la suffixation (*aide-cuisinier*)

— les noms en *-euse* ambigüs entre féminins de *-eur* et de *-eux* (*chatouilleuse*)

— les noms dont une acception au moins n'est pas agentive, qu'ils dénotent un état (*clameur*), un habitant (*jurassien*), un membre de communauté (*mariste*), un instrument (*navigateur*), un expérienceur (*croyant*), etc.

5. À l'image des suffixes *-euse* et *-rice* dans la section 7, nous intégrons les variantes féminines *-ante*, *-ienne* et *-ière* à notre étude. Nous référons par la suite aux seuls suffixes masculins pour des raisons de lisibilité.

— les noms multitypés ou sous-déterminés, permettant les interprétations agentives et non agentives (ex. agent/adepte, *impressionniste*)

Ne sont pas exclus les noms d'agent de même forme qu'un adjectif (*militant*) à condition que ceux-ci puissent être analysés comme des noms suffixés, c'est-à-dire s'il est établi sans équivoque que leur suffixe forme effectivement des noms d'agent (à l'image de *assistant*, issu de *assister*, qui n'est pas identifié comme un adjectif dans les ressources dictionnaires mais comme un nom d'agent), et qu'il n'existe pas d'indication morphologique que les noms d'agent examinés soient dérivés d'adjectifs.

Du fait des conditions d'analyse computationnelle présentée en section 7.2.2 que nous reprenons ici, seuls les noms ayant au moins cinq occurrences dans le corpus *Wikipedia2018* sont conservés. À l'issue de cette sélection, notre échantillon contient 1 252 noms d'agent, dont 27 *-aire*, 86 *-ant* (dont 3 féminins), 717 *-eur* (dont 54 féminins en *-euse* et 8 féminins en *-rice*), 54 *-ien* (dont 6 féminins), 161 *-ier* (dont 14 féminins) et 207 *-iste*.

8.2.2 Caractérisation des amorces

Nous proposons un premier diagnostic des similarités et différences entre les noms d'agent construits par ces différentes suffixations sur la base de l'analyse des amorces que nous venons de sélectionner. Nous nous penchons pour cela sur leurs propriétés morphosémantiques, puis sur leur comportement distributionnel au niveau individuel. Nous intégrons cette première étape d'analyse, absente de l'expérience présentée dans la partie III, car nous estimons qu'elle apporte un premier éclairage sur les propriétés notamment sémantiques des noms et de leurs bases, au vu de leur grande diversité. Cela nous permet de nous interroger dans un second temps sur leur maintien à l'échelle de la classe sémantique formée, et donc parmi les voisins des barycentres, indépendamment de la construction morphologique des voisins.

Une première différence entre les noms d'agent construits par les suffixes *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste* concerne la catégorie grammaticale des bases des amorces. Celle-ci est donnée dans le tableau 8.1. Le pourcentage par suffixe (en ligne) est donnée entre parenthèses. Signalons que la base analysée est celle intervenant dans la formation du nom d'agent, indépendamment des éventuelles constructions non agentives des suffixes. Dans le cas où le nom d'agent est sémantiquement dérivé d'une expression polylexicale, seul un des mots sert à construire la forme – généralement l'unité la plus discriminante sémantiquement (*feu d'artifice -> artificier*). Nous considérons alors que la catégorie de la base est celle de la tête de l'expression polylexicale, suivant Roché (2003) et**?**. Ainsi, les bases en apparence adjectivales de noms d'agent

tels que *fiscaliste* et *plasticien* sont réanalysées comme des bases nominales : *droit fiscal* et *arts plastiques*.

	Verbe	Nom
-aire	6 (22%)	**21 (78%)**
-ant	**86 (100%)**	-
-eur	**681 (95%)**	36 (5%)
-ien	-	**54 (100%)**
-ier	6 (4%)	**155 (96%)**
-iste	7 (3%)	**200 (97%)**

TABLE 8.1 – Catégories grammaticales des bases des amorces en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste*

Le tableau 8.1 montre que tous les suffixes sélectionnent à la fois des bases verbales et nominales, à l'exception de *-ant* (qui est strictement déverbal) et de *-ien* (qui est strictement dénominal). Néanmoins, même lorsque les suffixes sélectionnent les deux types de bases, ils montrent une très nette préférence pour une des deux catégories : *-ier*, *-iste*, et dans une moindre mesure *-aire*, pour les bases nominales, et *-eur* pour les bases verbales. Des premiers rapprochements peuvent être ébauchés sur la base de la catégorie grammaticale, à savoir que les suffixes *-ant* et *-eur* sont sélectionnés quand la base est verbale, et les suffixes *-ier*, *-iste* et dans une certaine mesure *-ien* quand elle est nominale.

Nous réutilisons les tests décrits en section 7.4 pour annoter le type sémantique des bases. L'annotation est basée sur les cinq catégories précédemment identifiées, à savoir action (*bricoler -> bricoleur*), objet (*machine -> machiniste*), domaine (*histoire -> historien*), propriété (*équilibre -> équilibriste*), institution (*poste -> postier*), auxquels nous rajoutons une sixième catégorie, à savoir les objets cognitifs (*préface -> préfacier*), qui inclut les noms qui ont une facette sémantique d'objet physique (*roman*). Ces noms sont annotés ici comme dénotant des objets cognitifs et non des objets physiques en raison du caractère distinctif et prépondérant de la facette cognitive. L'identification des bases dénotant objets cognitifs repose sur leur capacité à être objet du verbe *écrire*, ou suivis d'une relative de la forme *selon lequel P*. La distribution des types sémantiques des bases en fonction des suffixes est donnée dans le tableau 8.2. Le pourcentage par suffixe (en ligne) est donné entre parenthèses. Les valeurs les plus élevées par suffixe (en ligne) sont indiquées en gras.

Le tableau 8.2 fait apparaître que les suffixes *-ant*, *-eur*, et dans une moindre mesure *-aire*, privilégient les bases actionnelles. Les suffixes *-ien* et

	Action	Objet	Domaine	Propriété	Institution	Obj. cog.
-aire	**19 (70%)**	6 (22%)	-	-	-	2 (7%)
-ant	**86 (100%)**	-	-	-	-	-
-eur	**686 (96%)**	12 (2%)	18 (3%)	-	-	1 (0.1%)
-ien	4 (7%)	2 (4%)	**46 (85%)**	1 (2%)	1 (2%)	-
-ier	21 (13%)	**127 (79%)**	4 (3%)	-	3 (2%)	6 (4%)
-iste	32 (16%)	42 (20%)	**105 (51%)**	3 (1%)	-	25 (12%)

TABLE 8.2 – Type sémantique des bases des amorces en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste*

-ier favorisent quant à eux les bases dénotant respectivement des domaines et des objets. Le suffixe *-iste* est le plus hétérogène, avec néanmoins une préférence pour les bases qui dénotent des domaines. Cela tend à confirmer certains des rapprochements faits sur la base des catégories grammaticales des bases, notamment celui entre *-ant* et *-eur* (corrélé à la dénotation exclusivement actionnelle des bases verbales, Croft (1991)), et celui entre les suffixes *-ien* et *-iste*, bien que cela soit moins marqué puisque seuls 50% des noms en *-iste* sont construits sur une base dénotant un domaine.

Nous opérons une première comparaison sur le plan distributionnel par le biais des amorces en examinant pour chaque suffixe la distribution des noms d'agent échantillonnés selon leur proximité moyenne avec les autres amorces porteuses du même suffixe. Cela nous donne un aperçu de la densité des noms d'agent dans l'espace vectoriel, c'est-à-dire leur tendance à former un cluster localisé ou non. Cela nous donne un indice de leur cohésion sémantique. Nous pouvons ainsi faire l'hypothèse que plus les membres d'une classe forment un cluster dense dans l'espace vectoriel, plus la classe est sémantiquement homogène. La distribution des scores de proximité en fonction des suffixes est donnée dans la figure 8.1.

La figure 8.1 montre que les amorces en *-ien*, *-ier* et *-iste* tendent à être plus proches entre elles que ne le sont les amorces en *-aire*, *-ant* et *-eur*. La médiane est comprise entre 0.275 et 0.35 pour les premiers alors qu'elle est comprise entre 0.175 et 0.218 pour les seconds. Cela suggère que les noms d'agent en *-ien*, *-ier* et *-iste* sont plus similaires entre eux que ne le sont les autres noms d'agent. En d'autres termes, les suffixes *-aire*, *-ant* et *-eur* forment des noms d'agent sémantiquement plus variés que les suffixes *-ien*, *-ier* et *-iste*. De ce point de vue, le suffixe *-eur* forme la catégorie la plus hétérogène, alors que *-ien* forme la catégorie la plus homogène. Il faut cependant noter que la variation affichée par les noms en *-aire*, *-ant* et *-eur* est moindre que celle affichée par les noms en *-ien*, *-ier* et *-iste*. Bien qu'ils

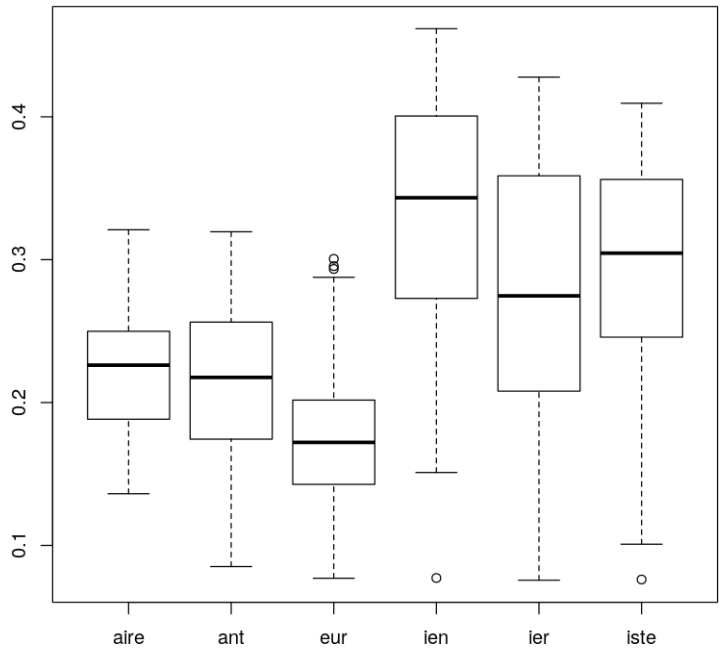

FIGURE 8.1 – Densité distributionnelle des amorces en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste*

soient en moyenne plus proches, les noms d'agent en *-ien*, *-ier* et *-iste* affichent donc de plus grandes disparités interindividuelles. Une explication serait que des sous-groupes sémantiques ou des séries idiosyncrasiques existent dans ces trois classes. Ces données permettent une nouvelle fois de faire émerger les deux groupes déjà mis en avant précédemment par le biais les préférences de sélection en termes de catégorie grammaticale et de type sémantique des bases, à savoir les suffixes *-ant* et *-eur* d'une part, auxquels s'ajoute le suffixe *-aire*, et les suffixes *-ien*, *-ier* et *-iste* d'autre part.

8.3 Profil distributionnel des suffixations

Notre objectif est d'étudier les catégories formées par ces suffixes, pour les comparer, et non pas rester au niveau des amorces. Nous souhaitons voir si les tendances que nous avons fait émerger à l'échelle des amorces se retrouvent à l'échelle de la classe sémantique formée par ces noms, et pas simplement

au niveau des amorces. Nous évaluons cela par le biais des barycentres. Nous commençons par quantifier les propriétés des catégories par le biais des profils distributionnels (section 8.3.1), puis nous les discutons (section 8.3.2).

8.3.1 Quantification morphosémantique

Pour étudier les catégories formées par les noms en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste*, nous construisons pour chacune un barycentre en reprenant la méthode présentée dans le chapitre 7. Un aperçu des 10 premiers voisins de chaque barycentre est donné dans le tableau 8.3.

-aire	*-ant*	*-eur*	*-ien*	*-ier*	*-iste*
médecin	détenu	plombier	physicien	cordonnier	ophtalmologue
comptable	citoyen	truand	neurologue	boulanger	photographe
collaborateur	gens	escroc	linguiste	armurier	pédagogue
bibliothécaire	délinquant	proxénète	psychiatre	maréchal-ferrant	neurologue
juriste	personne	rabatteur	mathématicien	bijoutier	clarinettiste
financier	villageois	prestidigitateur	informaticien	serrurier	pédiatre
informateur	employé	coiffeur	statisticien	épicier	violoniste
avocat	travailleur	garagiste	biologiste	perruquier	mycologue
savant	wikipédiens	gangster	pédiatre	charpentier	naturopathe
lobbyiste	comploteur	cuisinier	anthropologue	forgeron	ornithologue

TABLE 8.3 – 10 plus proches voisins des barycentres des noms d'agent en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste*

Comme l'illustre le tableau 8.3, l'ensemble des 100 voisins obtenus pour chacun des six voisinages sont des noms, à l'exception de deux voisins que l'on trouve dans le voisinage du barycentre des noms en *-ier*, *fermier* et *chaudronnier*, qui sont étiquetés comme adjectifs. Ces deux cas, et tout particulièrement *chaudronnier*, mettent en évidence le biais introduit par l'analyseur syntaxique, qui est à l'origine d'erreurs d'étiquetage. Cette source de bruit est néanmoins inévitable dans la mesure où nous souhaitons combiner une approche extensive et un contrôle strict sur les données lexicales, et s'avère dans les faits marginale au sein des voisinages étudiés. Malgré cette erreur d'étiquetage, on constate néanmoins que l'ensemble des 100 voisins de chaque barycentre dénotent des humains.

À l'image de ce que l'on observait dans la partie III, on retrouve parmi les voisins des barycentres des noms ayant servi d'amorces. Ce nombre varie cependant d'un barycentre à l'autre. Ainsi, ce taux de recouvrement est de 40% pour *-ier*, 37% pour *-iste* et 28% pour *-eur*, mais se révèle plus faible pour les barycentres des noms en *-ien* (13%), *-ant* (11%) et *-aire* (5%). Ces recouvrements tendent à corroborer certaines observations faites sur la base de la densité des amorces (figure 8.1, et notamment l'homogénéité des classes

sémantiques formée par les noms en *-ier* et *-iste*, et la plus grande hétérogénéité des classes formées par *-aire* et *-ant*.

De fait, on constate que les voisins sont très divers morphologiquement. Tout comme pour le voisinage de *-eur* dans la section 7.3, les voisins peuvent être des noms affixés (*collaborateur*), des converts (*émigré*), des composés (*maréchal-ferrant*), des noms morphologiquement simples (*personne*), des noms complexes non construits (*écrivain*), et des noms construits par des procédés extragrammaticaux comme la troncation (*admins*). Les propriétés morphologiques des voisins en fonction du suffixe ciblé sont donnés dans le tableau 8.5.

	Affixé	Convert	Composé	Simple	Complexe	Extragram.	Indeterminé
-aire	57	16	6	12	8	-	1
-ant	48	30	-	14	7	1	10
-eur	66	7	3	10	8	1	5
-ien	81	-	14	2	2	-	1
-ier	78	5	3	1	11	-	2
-iste	87	-	10	-	3	-	-

TABLE 8.4 – Type morphologique des voisins des barycentres des noms d'agent en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste*

Le tableau 8.5 montre que les voisins tendent globalement à être des noms affixés. C'est particulièrement le cas pour les barycentres des suffixes *-ien*, *-ier* et *-iste*, pour lesquels au moins 78% des voisins sont affixés. Ces tendances sont plus faibles notamment pour *-aire* (57%) et *-ant* (48%). Les voisinages des différents barycentres diffèrent cependant au regard des types morphologiques représentés. Ainsi, les types morphologiques des voisins de *-iste*, *-ien* et *-ier* sont moins variés que ceux des voisins de *-aire*, *-ant* et *-eur*. Par ailleurs, les types morphologiques ne se retrouvent pas dans les mêmes proportions dans les voisinages des différents barycentres. Les composés se trouvent ainsi principalement dans les voisinages des barycentres des noms d'agent en *-ien* et *-iste*, alors que les noms simples se trouvent surtout dans le voisinage des barycentres des noms d'agent en *-aire*, *-ant* et *-eur*. Les noms converts sont quant à eux surtout présents parmi les voisins des barycentres des noms d'agent en *-aire* et *-ant*. Les noms dont le type morphologique est indéterminé (c'est-à-dire les noms possiblement converts ou simples, voir section 7.3 pour l'explication) se retrouvent majoritairement parmi les voisins du barycentre des noms d'agent en *-ant*.

Les suffixes des voisins de chaque barycentre qui sont dérivés sont donnés dans le tableau 8.5. Pour les mêmes raisons que celles évoquées dans le chapitre 7, les préfixes – très marginaux puisque seuls 10 voisins sur 600 sont

préfixés, dont *haut-fonctionnaire*, *contremaître*, *parapsychologue* – ne sont pas pris en compte. La valeur la plus élevée pour chaque barycentre (en colonne) est indiquée en gras.

	-aire	*-ant*	*-eur*	*-ien*	*-ier*	*-iste*
-ain	-	1 (2%)	-	-	-	-
-aire	7 (12%)	1 (2%)	1 (2%)	1 (1%)	-	1 (1%)
-ant	4 (7%)	**16 (34%)**	-	-	2 (3%)	-
-ard	-	-	3 (4.6%)	-	-	-
-eur	**19 (33%)**	12 (26%)	**33 (51%)**	3 (4%)	17 (22%)	4(5%)
-ien	7 (12%)	7 (15%)	2 (3%)	17 (22%)	1 (1%)	10 (12%)
-ier	8 (14%)	6 (13%)	17 (26%)	-	**52 (66%)**	-
-iste	13 (22%)	3 (6%)	7 (11%)	28 (35%)	5 (6%)	**45 (52%)**
-logue	-	-	-	**30 (38%)**	-	26 (30.2%)
-ois	-	1 (2%)	-	-	-	-
-on	-	-	2 (3%)	-	2 (3%)	-

TABLE 8.5 – Suffixes des voisins dérivés des barycentres des noms d'agent en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste*

Le tableau 8.5 montre qu'un grand nombre de suffixes sont impliqués dans la construction des voisins dérivés, dont les 6 suffixes initialement ciblés ainsi que les suffixes *-ain*, *-ard*, *-logue*[6], *-oir*, et *-on*. Tous les suffixes ne sont cependant pas également représentés. Ainsi, les suffixes *-ain*, *-ois*, *-on* et *-ard* apparaîssent chacun moins de 5 fois parmi nos 600 voisins. Il est intéressant de souligner qu'une portion significative de noms en *-logue* sont voisins des barycentres de *-ien* et *-iste*. Cela s'explique par la propension des noms en *-logue* à dénoter des spécialistes (Amiot et Dal 2007, Villoing et Fiammetta 2014, Lasserre et Montermini 2014), les noms en *-ien* et *-iste* tendant eux-aussi à dénoter ce type d'agent. Dans nos échantillons, ils représentent respectivement 38% et 30% des voisins de *-ien* et *-iste*.

Du fait de la présence d'amorces dans les voisinages, on aurait pu s'attendre à ce que les suffixes ciblés soient majoritaires dans certains voisinages. C'est notamment le cas pour les barycentres des noms en *-eur*, *-ier* et *-iste*, pour lesquels respectivement 51%, 66% et 52% des voisins dérivés sont construits par la suffixation ciblée. Cela renforce encore une fois les observations quant à l'homogénéité des noms en *-iste* et *-ier*. Cette redondance affixale est en partie due aux amorces donc, mais pas seulement, puisque l'on

6. Le caractère suffixal vs élément de la chaîne *-logue* est discuté dans la littérature. Nous faisons le choix de le considérer ici comme un suffixe, suivant Amiot et Dal (2007), Villoing et Fiammetta (2014) puisque des bases en tant que lexème peuvent être identifiées.

retrouve aussi des noms qui n'ont pas été inclus dans les amorces, soit car ils ne valident pas nos critères (de monosémie pour *rabatteur*, construction prototypique pour *délateur*), soit parce qu'ils sont absents de Lexique3 (*neurophysiologiste*). Bien que dans des proportions moindres, le suffixe *-ant* est lui aussi majoritaire parmi les voisins dérivés du barycentre correspondant, à raison de 34%, ce qui est plus remarquable, puisque le recouvrement initial avec les amorces était relativement faible (11%). Cela s'explique par la présence parmi les voisins dérivés en *-ant* de noms dont une acception au moins n'est pas agentive (*croyant* dénotant un expérienceur). Le suffixe ciblé n'est cependant pas toujours majoritaire au sein du voisinage des barycentres, notamment pour les noms en *-aire* et *-ien*, où les suffixes respectifs se retrouvent à hauteur de 12%, et 22%. Notons que dans le cadre du barycentres des noms en *-aire*, le suffixe majoritaire est *-eur*, et pour *-ien*, *-logue*.

Nous présentons dans le tableau 8.6 la catégorie grammaticale des bases des voisins dérivés en fonction du barycentre. Le pourcentage par suffixe (en ligne) est donné entre parenthèses. La valeur la plus élevée par suffixe (en ligne) est signalée typographiquement.

	Verbe	Nom	Adjectif
-aire	28 (38%)	**37 (51%)**	8 (11%)
-ant	**43 (63%)**	19 (28%)	6 (9%)
-eur	**38 (52%)**	30 (41%)	5 (7%)
-ien	3 (4%)	**78 (96%)**	-
-ier	26 (31%)	**55 (66%)**	2 (2%)
-iste	4 (5%)	**82 (94%)**	1 (1%)

TABLE 8.6 – Catégories grammaticales des bases des voisins dérivés des barycentres des noms d'agent en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste*

Le tableau 8.6 montre que les voisins dérivés sont principalement dénominaux pour les suffixes *-aire* (*financier*), *-ien* (*chimiste*), *-ier* (*aubergiste*) et *-iste* (contrebassiste), et déverbaux pour les suffixes *-ant* (*travailleur*) et *-eur* (*arnaqueur*). Les bases adjectivales sont quant à elles marginales, et se trouvent principalement dans les voisinages de *-aire*, *-ant* et *-eur*. La répartition entre voisins déverbaux et dénominaux varie en fonction du barycentre, avec des tendances claires. La distinction est ainsi particulièrement sensible pour les suffixes *-ien* et *-ier*, où les voisins déverbaux sont marginaux. *A contrario*, la différence est moins marquée dans le cas de *-eur*, où les bases verbales et nominales se distribuent de façon relativement équilibrées (52% et 41%). Le voisinage intégrant la plus grande proportion de base verbale est celui du barycentre des noms d'agent en *-ant*. Il est intéressant de noter

que l'on observe les mêmes tendances, en termes de catégories grammaticales, que celles constatées pour les amorces dans le tableau 8.1, si ce n'est l'ajout de quelques bases adjectivales (qui restent néanmoins minoritaires). Cela montre que les propriétés observées à l'échelle des amorces tendent à se confirmer à l'échelle de la catégorie sémantique, indépendamment du type morphologique des membres de la catégorie.

Pour évaluer dans quelle mesure ces tendances syntaxiques se traduisent sur le plan sémantique (Croft 1991), nous annotons le type sémantique des bases des voisins dérivés selon les mêmes tests et catégories que pour les amorces. Le résultat de l'annotation est donné dans le tableau 8.7.

	Action	Objet	Domaine	Propriété	Institution	Obj. cog.
-aire	**33 (45%)**	12 (16%)	12 (16.4%)	9 (12%)	3 (4%)	4 (6%)
-ant	**41 (60%)**	8 (12%)	-	13 (19.1%)	4 (6%)	2 (3%)
-eur	**45 (62%)**	16 (22%)	5 (6.8%)	5 (7%)	2 (3%)	-
-ien	4 (5%)	26 (32%)	**45 (56%)**	3 (4%)	3 (4%)	-
-ier	28 (34%)	**49 (59%)**	3 (4%)	2 (2%)	1 (1%)	-
-iste	5 (6%)	26 (30%)	**47 (54%)**	3 (3%)	2 (2%)	4 (5%)

TABLE 8.7 – Types sémantiques des bases des voisins des barycentres des noms d'agent en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste*

Le tableau 8.7 indique que les bases dénotant des actions, des objets et des domaines sont prédominantes. Les bases dénotant des propriétés, des institutions et des objets cognitifs sont quant à elles plutôt marginales – excepté les propriétés dans le voisinage du barycentre des noms d'agent en *-ant* et dans une moindre mesure celui des noms d'agent en *-aire*. Là encore, d'importantes disparités peuvent être observées entre les suffixes. Certains barycentres favorisent nettement un type de base, comme dans le cas des barycentres des noms d'agent en *-eur* avec action, en *-ien* et *-iste* avec domaine, et en *-ier* avec objet. *A contrario*, les autres barycentres sont plus hétérogènes, notamment le suffixe *-aire*. Par ailleurs, une distribution relativement similaire des types sémantiques de bases peut être observée entre *-ier* et *-iste*, bien que seuls les derniers incluent des mots dérivés de bases dénotant des objets cognitifs. À l'image des catégories grammaticales des bases, on notera que les analyses sur les types sémantiques faites à partir des voisins dessinent les mêmes tendances que celles faites à partir des amorces (tableau 8.2). Cela permet une nouvelle fois de confirmer que les propriétés observées à l'échelle des amorces se confirment à l'échelle des catégories sémantiques.

Une grande partie des observations menées sur les types morphologiques et les suffixes des voisins, ainsi que sur les catégories grammaticales et types

sémantiques des bases, suggèrent que certains barycentres présentent des propriétés morphosémantiques communes. Nous avons ainsi notamment rapproché à plusieurs reprises les suffixes *-ien* et *-iste*. Nous reprenons en détails ces propriétés dans la section 8.3.2, mais pour compléter l'exploration de ces similarités, nous comparons directement les barycentres en analysant le recouvrement de leur voisinage et la proximité cosinus de ces barycentres. Le nombre de voisins partagés par les différents barycentres est donné dans le tableau 8.8.

	-aire	*-ant*	*-eur*	*-ien*	*-ier*	*-iste*
-aire	–	16	15	12	12	4
-ant	16	–	5	0	4	0
-eur	15	5	–	4	23	3
-ien	12	0	4	–	1	62
-ier	12	4	23	1	–	1
-iste	4	0	3	62	1	–

TABLE 8.8 – Nombre de voisins partagés par les barycentres des noms d'agent en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste*

Le tableau 8.8 montre que le taux de recouvrement des voisins est variable en fonction des barycentres. Par exemple, les barycentres des noms d'agent en *-ien* et *-iste* partagent 62 voisins sur 100, ce qui révèle une forte proximité entre les deux barycentres, et par extension entre les deux classes sémantiques formées par les noms d'agent construits par ces suffixes. Dans une moindre mesure, les 23 voisins partagés par les barycentres des noms d'agent en *-eur* et *-ier* suggèrent des similarités distributionnelles entre ces suffixes. D'un autre côté, concernant *-aire*, le nombre de voisins partagés est à peu près équivalent pour tous les barycentres, ce qui suggère qu'il n'est pas plus proche d'un des autres barycentres en particulier.

Le tableau 8.9 donne les scores de proximité moyens entre les barycentres des noms d'agent en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste*. Le score le plus élevé par barycentre (en ligne) est en gras.

Ces scores confirment les rapprochements ébauchés dans le tableau 8.8 et sur la base de nos observations. Ainsi, la très forte proximité observée entre les barycentres des noms d'agent en *-ien* et *-iste* (0.935) comparée à celle plus faible avec les autres barycentres confirment la plus forte similarité distributionnelle des dérivés en *-ien* et *-iste*. Un constat relativement similaire peut être fait pour les suffixes *-ier* et *-eur* d'une part, et *-ant* et *-eur* d'autre part. *A contrario*, le barycentre des noms d'agent en *-aire* affiche des scores de proximité similaires, bien que relativement élevés, avec l'ensemble des autres barycentres, ce qui suggère qu'il est équidistant des autres barycentres.

	-aire	*-ant*	*-eur*	*-ien*	*-ier*	*-iste*
-aire	-	0.762	**0.780**	0.732	0.744	0.715
-ant	0.762	-	**0.772**	0.534	0.656	0.502
-eur	0.780	0.772	-	0.745	**0.842**	0.773
-ien	0.732	0.534	0.745	-	0.677	**0.935**
-ier	0.744	0.656	**0.842**	0.677	-	0.716
-iste	0.715	0.502	0.773	**0.935**	0.716	-

TABLE 8.9 – Scores de proximité des barycentres des noms d'agent en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste*

La comparaison directe des barycentres, sur la base de leurs voisins partagés et de leurs scores de proximité, corrobore les observations faites à partir des différentes propriétés morphosémantiques des voisins. Ces observations vont par ailleurs dans le sens des tendances mises au jour pour les amorces. Cela montre que ces propriétés ne sont pas spécifiques aux amorces elles-mêmes, mais sont communes aux classes sémantiques qu'elles forment. Ces classes sont formellement plus variées, et subsument dans certains cas plusieurs suffixations, à l'image des noms en *-ien* et *-iste*.

8.3.2 Discussion

Les analyses présentées en section 8.3.1 sur les amorces et voisins des barycentres des noms d'agent en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste* permettent de faire émerger des fonctionnements sémantiques plus ou moins distinctifs des classes sémantiques formées, et plus largement des types de noms d'agent qui les constituent.

Les suffixes semblent ainsi former deux groupes aux profils distributionnels distincts. Dans le cas des suffixes *-ien*, *-ier* et *-iste*, des corrélations claires peuvent être faites entre la distribution des suffixes et des sous-types agentifs. Dans le cas des suffixes *-aire*, *-ant* et *-eur*, les spécificités sont plus complexes, bien que des tendances se dessinent. Nous détaillons les observations ci-dessous.

Convergence des noms d'agent en *-ien* et *-iste*

On observe à la fois une forte proximité et une forte homogénéité des suffixations en *-ien* et *-iste*, qui confirment les observations faites précédemment (Dubois 1962, Lignon 2007, voir section 8.1). Les deux suffixes ont des profils distributionnels similaires, comme le révèle la forte proximité de leurs barycentres, le taux élevé de voisins partagés, et les profils morphoséman-

tiques similaires similaire de leurs amorces et voisins. Les bases nominales prédominent à la fois parmi les amorces et les voisins dérivés, bases majoritairement construites à partir de noms de domaine – cette spécificité étant plus marquée pour le suffixe *-ien* que pour *-iste*. Les noms d'agent en *-ien* et *-iste* et les voisins de leurs barycentres respectifs dénotent principalement des experts de domaines spécifiques et des agents réalisant des activités intellectuelles, ce qui est confirmé par leur proximité avec les noms en *-logue*. La forte proximité interne des noms en *-ien* et *-iste* souligne leur homogénéité sémantique et, combiné à la similarité distributionnelle entre les deux suffixes, corrobore leur spécificité sémantique par rapport aux autres suffixes agentifs.

Les classes définies par les suffixes *-ien* et *-iste* sont donc à la fois très homogènes et très proches. Comme indiqué en section 8.1, les suffixes se distinguent par des aspects morphologiques et diachroniques, mais on peut se demander s'ils possèdent aussi des spécificités sémantiques. Un examen plus approfondi des amorces et voisins montre que le suffixe *-iste* a la capacité plus spécifique de former des noms dénotant des artistes, un certain nombre d'amorces et de voisins exclusifs au barycentre des noms d'agent en *-iste* étant des noms tels que ceux respectivement en (27a) et (27b).

(27) a. artiste, violoniste, clarinettiste, marionnettiste, nouvelliste, soliste

b. compositeur, cinéaste, illustrateur, peintre, dessinateur, chorégraphe

Cette distinction s'observait déjà dans les données présentées par Cartoni *et al.* (2015) dans leur étude constrastive des noms d'agent français et italiens. D'une manière générale, le suffixe *-iste* semble être légèrement plus diversifié que le suffixe *-ien*, avec davantage de cas marginaux (*automobiliste*, *gréviste*, *perchiste*, *projectionniste*), ainsi que la dénotation d'un plus grand nombre de sous-types agentifs, comme les noms de partisans et les gentilés. *A contrario*, le suffixe *-ien* semble plus homogène, avec une variation sémantique moindre en termes de sélection des bases, avec une densité distributionnelle plus élevée, et des voisins exclusifs dénotant uniquement des noms de scientifiques. On peut noter à cet égard que sur les 24 voisins exclusifs qui ne sont pas des noms en *-ien*, dix sont en *-logue*, et deux en *-graphe*.

Préférence du suffixe *-ier* pour les activités manuelles et commerçantes

Le suffixe *-ier* partage un certain nombre de propriétés avec les suffixes *-ien* et *-iste*, incluant notamment une forte proximité des amorces entre elles et une prédilection pour les bases nominales. Mais on note une prédilection tout particulière, qui n'est pas observée chez les autres suffixes, pour les bases

dénotant des objets, tant chez les amorces que les voisins dérivés. Les noms en *-ier* diffèrent des autres noms d'agent dans la mesure où ils dénotent principalement des professions relatives à l'artisanat ou au commerce, et corrélativement dérivent souvent de noms qui dénotent des artefacts, comme le souligne déjà Roché (2004). Des noms comme ceux en (28) sont uniquement présents dans le voisinage du barycentre des noms d'agent en *-ier*.

(28) maréchal-ferrant, tanneur, aubergiste, tailleur, tisserand, palefrenier, marchand, orfèvre, apothicaire, bûcheron

L'homogénéité morphosémantique des voisins et la forte proportion d'amorces dans le voisinage du barycentre confirment à la fois la spécificité de ce suffixe. Certaines connections peuvent être cependant faites avec le suffixe *-eur*, notamment au travers de la proximité de leurs barycentres. Cela pourrait être dû au fait que certains noms d'agent déverbaux en *-eur* dénotent des travailleurs manuels ou des commerçants (*carreleur*, *ferrailleur*, *fraiseur*, *vendeur*), comme souligné par Cartoni *et al.* (2015). Cette proximité se cristallise dans des paires synonymiques ou co-hyponymiques impliquant les deux suffixes comme *costumier* et *habilleur*, et *cafetier* et *restaurateur*. Il est intéressant de noter à ce titre que ces paires peuvent être complétées d'un troisième membre, respectivement *styliste* et *aubergiste*, ce qui tend à confirmer la polyvalence plus importante de *-iste* par rapport à *-ien*, en le rapprochant de *-ier* et de *-eur*.

Polyvalence des suffixes *-aire*, *-ant* et *-eur*

Les suffixes *-aire*, *-ant* et *-eur* sont sémantiquement plus versatiles que les suffixes *-iste*, *-ien* et *-ier*. Les noms d'agent montrent une plus grande dispersion dans l'espace vectoriel, correspondant à de plus grandes disparités sémantiques. Les voisins des barycentres sont plus divers relativement à leur type morphologique, ainsi qu'à leur suffixe lorsqu'ils sont dérivés. On trouve un plus grand nombre de noms simples dans leurs voisinages que dans ceux des autres suffixes, et dans la mesure où les noms simples ne sont pas contraints morphosémantiquement, ils peuvent afficher une plus grande variété de types agentifs. Par ailleurs, les voisins dérivés sont construits sur des bases sémantiquement grammaticalement plus hétérogènes, ce qui est cohérent avec une plus grande dispersion sémantique des agents dénotés si l'on part du postulat que le type sémantique et la catégorie grammaticale de la base participent du type de nom d'agent formé.

Une autre propriété distinctive des suffixes *-aire*, *-ant* et *-eur* est la prédominence de bases dénotant des actions dans les amorces et les voisins dérivés des barycentres. On pourrait faire l'hypothèse que cette propriété contribue à

la dispersion sémantique des noms formés, les bases dénotant des actions permettant la formation de noms d'agent plus variés. En effet, les noms d'agent construits sur des mots dénotant des actions peuvent dénoter des agent d'un événement particulier (*agresseur*) et pas exclusivement des agents avec un statut professionnel (*vendeur*).

Malgré ces propriétés communes, on observe des différences entre les trois suffixes.

Le suffixe *-ant* semble avoir un comportement plus singulier encore que *-aire* et *-eur*, comme le suggèrent plusieurs faits. Tout d'abord, le barycentre des noms d'agent en *-ant* est plus distant des barycentres des noms d'agent en *-ien*, *-ier* et *-iste* que ne le sont les barycentres des noms d'agent en *-aire* et *-eur*. Ce barycentre partage par ailleurs moins de voisins avec les autres barycentres que ne le font les autres.

On peut aussi noter que le barycentre de *-ant* inclut davantage de noms non agentifs dans son voisinage que les autres barycentres, à l'image des noms *personne*, *villageois*, *proche*, *patient*, *jeune*, *chrétien*, *enfant*, *nécessiteux*, et *déporté*, que l'on ne retrouve que dans le voisinage du barycentre des noms d'agent en *-ant*. Cela suggère que les noms d'agent en *-ant* ne dénotent pas des agents prototypiques. On observe en tout cas ici une situation paradoxale : le seul suffixe qui produit exclusivement des noms d'agent déverbaux apparaît comme le moins prototypiquement agentifs des suffixes formateurs de noms d'agent. Cela pourrait être un argument supplémentaire en faveur de la disqualification de *-ant* comme suffixe agentif (Anscombre 2001).

Les suffixes *-eur* et *-aire* présentent moins de spécificités que *-ant*. Là où le suffixe *-eur* semble être le suffixe agentif le plus productif, comme en atteste le nombre plus élevé d'amorces par rapport aux autres suffixes, mais aussi le plus versatile, comme en témoigne la plus forte dispersion, les noms en *-aire* forment aussi un ensemble plus restreint, et comme mentionné en section 8.1, le suffixe semble assez peu productif pour la formation de noms d'agent. Le suffixe *-aire* se distingue néanmoins par sa préférence pour les bases nominales, qui sont sémantiquement plus hétérogènes que pour *-ant* et *-eur*. Le type actionnel n'y est ainsi pas aussi dominant. Au même titre que les suffixes *-iste*, *-ien* et *-ier*, le suffixe *-aire* semble former des noms de profession, mais sans aucune prédilection pour un certain type ontologique (artistes et spécialistes pour *-iste*, spécialistes pour *-ien*). De fait, la propriété distinctive du suffixe *-aire* est de tendre à la caractérisation de statuts non spécialisés en sélectionnant des noms dénotant des actions comme bases.

8.4 Clustering

Nous avons vu dans la section 8.3 que les barycentres construits à partir des différents suffixes semblaient être associés à des spécificités sémantiques, notamment liées aux domaines d'activité dans le cas de *-ien*, *-ier* et *-iste*.Nous avons aussi vu que les voisins de ces barycentres étaient morphologiquement divers. On peut donc à ce titre se demander dans quelle mesure les propriétés sémantiques associées aux barycentres sont spécifiques aux suffixes ciblés par lesdits barycentres, et si elles sont corrélées à des propriétés morphologiques.

Pour répondre à cette question, nous adoptons une approche *bottom-up* basée sur le clustering. Nous nous proposons de regrouper les vecteurs des noms d'agent en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste* sur la base de leur profil distributionnel, pour ensuite étudier les propriétés morphosémantiques des clusters ainsi formés. Cette méthode nous permet d'identifier les noms d'agent à partir de leurs propriétés sémantiques et non plus formelles, et de déterminer les caractéristiques morphologiques des classes distributionnelles mises au jour.

Nous présentons en section 8.4.1 la méthode que nous déployons. Nous proposons ensuite une caractérisation sémantique des clusters (section 8.4.2) avant d'examiner l'interaction des caractéristiques sémantiques mises au jour avec des propriétés morphosémantiques (section 8.4.3).

8.4.1 Méthode

Nous souhaitons faire émerger des classes sémantiques au sein des noms d'agent étudiés, indépendamment de leur construction. Pour cela, nous appliquons un algorithme de clustering aux vecteurs des noms d'agent, qui visent donc à grouper les noms sur la base de leur propriétés distributionnelles.

Nous réalisons le clustering des vecteurs des 1 252 noms d'agent (section 8.2.1) en utilisant l'algorithme *spherical k-means*[7]. Cet algorithme d'apprentissage non supervisé cherche à partitionner les items en k clusters de sorte que les membres d'un cluster soient plus similaires entre eux qu'avec les membres des autres clusters. Pour cela, l'algorithme cherche à minimiser la distance cosinus entre les membres de chaque cluster donné.

Le *spherical kmeans* est un algorithme non supervisé en cela qu'il regroupe des items en clusters sans annotation préalable. On doit cependant en amont indiquer le nombre de clusters à former. Nous choisissons arbitrairement de fixer k à 6, sur la base du nombre de suffixes étudiés, en l'absence de toute présomption quant au nombre de sous-types agentifs pertinents.

7. Nous utilisons le package `skmeans` de R (Buchta *et al.* 2012), méthode `pclust`.

Comme expliqué à plusieurs reprises jusque-là, l'espace vectoriel et les représentations vectorielles varient d'un modèle à l'autre du fait de l'instabilité intrinsèque des modèles. Cela a donc un effet sur les proximités – effet que nous avons essayé de limiter en moyennant nos résultats sur cinq modèles – mais cela aura donc aussi potentiellement un effet sur le clustering, puisqu'il repose sur les représentations vectorielles. De ce fait, nous réalisons un clustering par modèle, et nous évaluons la stabilité du clustering à l'aide du Rand Index (Rand 1971), qui permet de quantifier l'accord entre plusieurs clusters sur une échelle de 0 (pas d'accord) à 1 (accord parfait). L'accord moyen [8] entre les cinq modèles est de 0.845, ce qui indique que le clustering est relativement stable. En d'autres termes, les noms d'agent sont regroupés de façon relativement stables dans les cinq modèles. Nous choisissons donc de n'analyser le clustering que d'un seul modèle, en considérant qu'il sera représentatif.

8.4.2 Caractérisation sémantique des clusters

Du fait des fondements théoriques de la sémantique distributionnelle, nous faisons l'hypothèse que le clustering des noms d'agent reposera sur les distinctions sémantiques précédemment étudiées, et plus particulièrement sur les types ontologiques (noms de spécialistes, d'artisans, etc). Nous nous attendons à ce que les clusters ainsi formés correspondent, dans une certaine mesure, à des sous-types agentifs. L'analyse qualitative des clusters permet dans une certaine mesure de faire émerger les types sémantiques précédemment évoqués, mais fait aussi émerger d'autres regroupements.

La répartition des 1 252 noms d'agent en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste* dans les six clusters (désormais C1 à C6) est donnée dans le tableau 8.10. On constate que la distribution des noms d'agent au sein des clusters est relativement homogène, les clusters agrégeant entre 12 et 21% des noms, même si les clusters 2, 3 et 4 tendent à être légèrement plus peuplés que les autres.

C1	C2	C3	C4	C5	C6
149	254	263	255	179	152

TABLE 8.10 – Répartition des 1 252 noms d'agent en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste* dans les six clusters

Nous commençons dans un premier temps par analyser les clusters au

8. L'accord moyen du clustering est calculé à l'aide de la fonction `rand.index()` du package `fossil` de R.

regard des types sémantiques d'agent qu'ils contiennent. Pour cela, nous analysons manuellement les noms d'agent, et faisons émerger les régularités que l'on observe en des termes ontologiques, sans appliquer une typologie préalable. Cette description des clusters ne se veut pas exhaustive dans le sens où elle ne permet pas de caractériser l'ensemble des membres des clusters, mais offre néanmoins un aperçu de certaines de leurs propriétés sémantiques.

Ainsi, l'examen détaillé des noms inclus dans le cluster 1 montre qu'ils dénotent principalement des agents impliqués dans des actions non physiques, en particulier en lien avec des opérations financières, comme dans (29a). Plus largement, le cluster 1 inclut de nombreux noms dénotant des hommes d'affaire (29b).

(29) a. acheteur, assureur, bienfaiteur, cotisant, donateur, enchérisseur, emprunteur, épargnant, gestionnaire, investisseur, parieur, payeur, prêteur, souscripteur, revendeur, spéculateur

b. affairiste, annonceur, assureur, avionneur, décideur, détaillant fournisseur, exportateur, promoteur, repreneur

Le cluster 2 est principalement constitué de noms d'agent tels que (30a) dénotant des artisans et des noms de professions manuelles. Le cluster 2 contient également des noms faisant référence à des commerçants (30b), appartenant parfois à la classe précédemment évoquée des professions manuelles.

(30) a. bagagiste, bricoleur, cardeur, carreleur, caissier, couvreur, cueilleur, déménageur, élagueur, ferblantier, fraiseur, machiniste, manutentionnaire, réparateur, rempailleur

b. antiquaire, armurier, barbier, bouquiniste, boutiquier, brocanteur, cafetier, caviste, droguiste, épicier, fleuriste, libraire, opticien, pharmacien, poissonnier, quincailler, tavernier

Les noms regroupés dans le cluster 3 dénotent majoritairement des agents engagés dans des activités intellectuelles, génériques dans le cas de *chercheur*, *concepteur*, *inventeur*, ou *penseur*, mais plus souvent incluant une spécialisation dans un domaine donné. Ces noms dénotent principalement :

— des scientifiques (*astrophysicienne*, *bactériologiste*, *ethnobotaniste*)

— des médecins (*clinicien*, *chirurgien*, *neurologiste*)

— des artistes graphiques (*aquarelliste*, *caricaturiste*, *maquettiste*)

— des musiciens (*bassiste*, *clarinettiste*, *concertiste*)

— des écrivains ou des journalistes (*apologiste*, *intervieweur*, *nouvelliste*)

Étonnamment, le cluster 3 regroupe aussi des noms de sportifs (*alpiniste*, *basketteur*, *boxeuse*), qui ne se caractérisent pas par une pratique intellectuelle mais bien physique.

Les noms présents dans le cluster 4 dénotent des agents qui réalisent des actions qui impliquent ou affectent directement d'autres êtres humains. Ces actions peuvent relever des relations sociales (31a) ou avoir une visée psychologique (31b). Le cluster 4 comprend également des noms dénotant des agents impliqués dans des actions communicationnelles (31c).Notons que ce cluster inclut de nombreux noms d'agent faisant référence à des actions négatives voire illégales tels que ((31d)).

(31) a. entremetteur, intercesseur, marieur, pacificateur

b. amuseur, consolateur, envoûteur, suborneur, tentateur, trompeur

c. causeur, contradicteur, diseur, écoutant, insulteur, parleur, prêcheur, questionneur, raconteur

d. abuseur, agresseur, écorcheur, empoisonneur, éventreur, kidnappeur, meurtrier, oppresseur, racketteur, ravisseur, tortionnaire, tourmenteur, tueur, violeur

Un des traits les plus distinctifs du cluster 5 est le fait qu'il contienne une grande proportion des amorces féminines (42 des 85 disponibles). On peut se demander si la présence des noms féminins est correlée à la forte présence dans le même cluster de noms masculins connotés négativement, et liés à la séduction (*baratineur*, *cavaleur*, *enjôleur*), à des activités sexuelles (*embrasseur*, *baiseur*, *fécondateur*, *trousseur*, *fouteur*, *enculeur*), et possiblement répréhensibles (*mateur*, *tripoteur*). À l'image de ce que l'on observait dans le cluster 4, on retrouve dans le cluster 5 la présence d'un trait axiologique négatif, relevant ici davantage de qualités morales que d'actions répréhensibles (*arnaqueuse*, *bluffeur*, *carotteur*, *crâneur*, *embrouilleur*, *feinteur*, *frimeur*, *lâcheur*).

Le cluster 6 regroupe quant à lui des noms d'agent dénotant des combattants (32a), des militaires (32b), et des sportifs (32c) (que l'on retrouvait déjà partiellement dans le cluster 3, avec *danseur*, *handballeuse*, ou *kitesurfeur*), avec des recouvrements possibles entre ces différents groupes.

(32) a. attaquant, bagarreur, cogneur, combattant, frappeur, flagellateur, jouteur, lutteur

b. archer, carabinier, commandant, démineur, missilier, mitrailleur, officier, piquier, tankiste

c. boxeur, catcheur, cycliste, épéiste, footballeur, marathonien, patineur, planchiste, pugiliste, skieur, surfeur

La classe d'agent dénotés dans le cluster 6 s'étend aux sportifs caractérisés par des mouvements techniques spécifiques (*buteur*, *descendeur*, *dribbleur*, *plaqueur*, *pagayeur*) et plus largement à des agents définis par une manière de déplacement (*coureur*, *fonceur*, *grimpeur*, *marcheur*, *nageur*).

Ces différents sous-clusters définissent des sous-types ontologiques d'agent, dont nous nous demandons s'ils sont corrélés à des propriétés morphosémantiques distinctives.

8.4.3 Analyse quantitative des clusters

L'analyse qualitative des clusters a permis de mettre au jour des tendances sémantiques. Bien que l'on ne s'y soit pas attardé, quelques remarques formelles pouvaient être faites sur la base des exemples donnés en section 8.4.2. Ainsi, les noms d'agent en *-ant* semblaient plus spécifiques au cluster 4, les noms d'agent en *-ien* et *-iste* au cluster 3, ou les noms en *-ier* et *-aire* au cluster 2. Nous systématisons ici cette analyse pour déterminer si les propriétés sémantiques sont corrélées à des propriétés morphologiques. Nous commençons donc par analyser l'appartenance des noms d'agent aux clusters au regard de leurs propriétés morphosémantiques précédemment étudiées (à savoir les suffixes des noms d'agent, ainsi que la catégorie grammaticale et le type sémantique de leur base), puis nous évaluons sur le plan statistique la spécificité de ces propriétés.

Nous donnons dans le tableau 8.11 le profil morphologique de chaque cluster au regard des suffixes des noms d'agent qui les composent. Le pourcentage par suffixe (en colonne) est donné entre parenthèses. Nous indiquons en gras pour chaque suffixe (en colonne) le cluster où il est le plus représenté.

	-aire	*-ant*	*-eur*	*-ien*	*-ier*	*-iste*
C1	6 (22%)	**49 (57%)**	77 (11%)	2 (4%)	9 (6%)	6 (3%)
C2	5 (19%)	6 (7%)	115 (16%)	5 (9%)	**86 (53%)**	37 (18%)
C3	5 (19%)	2 (2%)	84 (12%)	**34 (63%)**	11 (7%)	**127 (61%)**
C4	**7 (26%)**	17 (20%)	**198 (28%)**	6 (11%)	18 (11%)	9 (4%)
C5	-	2 (2%)	145 (20%)	2 (4%)	22 (14%)	8 (4%)
C6	4 (15%)	10 (12%)	98 (14%)	5 (9.3%)	15 (9%)	20 (10%)

TABLE 8.11 – Répartition des suffixes des noms d'agent dans les six clusters

Le tableau 8.11 montre que la distribution des suffixes dans les 6 clusters varie. Les noms d'agent en *-ant*, *-ien*, *-ier* et *-iste* sont principalement regroupés dans un cluster donné, alors que les noms d'agent en *-aire* et *-eur* se distribuent davantage au sein des différents clusters. La majorité des noms d'agent en *-ien* et *-iste* se trouvent dans le même cluster (C3), ce qui confirme

la forte proximité de ces deux suffixes. On peut cependant remarquer qu'une portion non négligeable de noms d'agent en *-iste* se trouvent dans le cluster 2. Ce regroupement s'explique en cela que ces noms se révèlent des noms d'agent en *-iste* non canoniques dans la mesure où ils ne dénotent pas des spécialistes mais des professions manuelles (*métallurgiste*, *garagiste*, *chauffagiste*) ou des commerçants (*aubergiste*, *bouquiniste*, *caviste*), en accord avec la caractérisation sémantique générale du cluster 2 (8.4.2).

Plus de la moitié des noms d'agent en *-ant* et *-ier* se regroupent respectivement dans les clusters 1 et 2. En se basant sur le postulat que les clusters se caractérisent par leur homogénéité distributionnelle, ces résultats suggèrent que les noms d'agent en *-ant*, *-ien*, *-ier*, et *-iste* présentent à la fois une certaine homogénéité sémantique, puisqu'ils sont en grande partie groupés dans un même cluster, et une spécialisation sémantique, puisqu'ils se mélangent assez peu aux autres noms d'agent suffixés. *A contrario*, la répartition plus uniforme des noms d'agent en *-eur*, et dans une moindre mesure de ceux en *-aire*, confirme la versatilité de ces deux suffixes. Notons néanmoins que la plus grande portion de noms d'agent en *-aire* et *-eur* se trouvent dans le même cluster (C4), ce qui suggère que ces suffixes tendent à former le même sous-type de noms d'agent.

Enfin, on constate que les clusters 5 et 6 ne sont pas spécialement privilégiés par un suffixe donné. Deux explications sont possibles. Tout d'abord, on peut envisager que l'hétérogénéité sémantique du cluster ait pour conséquence qu'aucun suffixe ne s'y reconnaît particulièrement. Par ailleurs, le suffixe n'est sans doute pas une variable explicative suffisante pour rendre compte de la cohérence sémantique de ces clusters, ce qui souligne le fait que si l'on observe des tendances, aucun cluster ne se caractérise par la présence exclusive d'un suffixe, et inversement, aucun suffixe n'occupe qu'un seul cluster. Ces résultats sont cependant à mettre en regard avec le déséquilibre des effectifs, le suffixe *-eur* étant sur-représenté. Les noms d'agent en *-eur* ont donc plus de chances de se répartir dans l'ensemble des clusters, et donc de diluer un peu l'isolement des autres suffixes. Une reprise du clustering à partir d'échantillons similaires serait nécessaire pour stabiliser les observations, mais avec le risque d'un biais induit par la sélection des noms d'agent. Des tendances émergent néanmoins, dont on peut se demander si elles interagissent avec d'autres critères.

Nous étudions dans un deuxième temps l'impact de la catégorie grammaticale des bases des noms d'agent comme critère explicatif de l'appartenance aux clusters. Le tableau 8.12 donne la distribution des catégories grammaticales des bases au sein des clusters. Le pourcentage par catégorie grammaticale (en colonne) est donné entre parenthèses.

Le tableau 8.12 nous montre que les bases nominales se trouvent princi-

	Verbe	Nom
C1	128 (16%)	21 (5%)
C2	122 (16%)	132 (28%)
C3	78 (10%)	185 (40%)
C4	214 (27%)	41 (9%)
C5	140 (18%)	39 (8%)
C6	104 (13%)	48 (10%)

TABLE 8.12 – Catégories grammaticales des bases des noms d'agent dans les six clusters

palement dans les clusters 2 et 3, alors que les bases verbales se répartissent de façon plus uniforme au sein des clusters, malgré une légère concentration dans le cluster 4. Les contrastes les plus importants sont observés au sein des clusters eux-mêmes. La distribution pondérée par le nombre global de bases verbales et nominales fait émerger d'importantes disparités. Les clusters 1 et 4 privilégient les bases verbales (respectivement 79% et 76%), alors que le cluster 3 inclut principalement des noms d'agent dérivés de bases nominales (80%). Des différences moins marquées sont observées dans les clusters 2 (65% de bases nominales, 35% de bases verbales) et 5 (68% de bases verbales, 32% de bases nominales). Le cluster 6 est le cluster le plus équilibré relativement à la distribution des catégories grammaticales des bases (56% de bases verbales, 43% de bases nominales).

Le clustering des noms d'agent en fonction du type sémantique de leur base morphologique est présenté dans le tableau 8.13. Le pourcentage par type sémantique (en colonne) est donné entre parenthèses, et la valeur la plus élevée par type sémantique (en colonne) est indiquée en gras.

	Action	Objet	Domaine	Propriété	Institution	Obj. cog.
C1	139 (16%)	8 (4.2%)	1 (1%)	-	1 (25%)	-
C2	129 (15%)	**104 (55%)**	17 (10%)	-	**2 (50%)**	2 (6%)
C3	92 (11%)	20 (10%)	**121 (70%)**	**2 (50%)**	-	**28 (82%)**
C4	**230 (27%)**	11 (6%	11 (6%)	-	-	3 (9%)
C5	144 (17%)	25 (13%)	6 (4%)	**2 (50%)**	1 (25%)	1 (3%)
C6	114 (13%)	21 (11%)	17 (10%)	-	-	-

TABLE 8.13 – Types sémantiques des bases des noms d'agent dans les six clusters

Les données du tableau 8.13 montrent que certains noms d'agent favorisent des clusters précis en fonction du type sémantique de leur base. C'est particulièrement le cas pour les noms d'agent dérivés de bases dénotant des

objets (principalement groupés dans le cluster 2), et pour les noms d'agent dérivés de bases dénotant des domaines et des objets cognifits (principalement groupés dans le cluster 3). Le regroupement au sein d'un même cluster de ces deux types de bases est lié au fait que le cluster 3 regroupe des noms d'agent dénotant des spécialistes et des artistes. *A contrario*, les bases dénotant des actions se répartissent de façon plus uniforme au sein des clusters, bien qu'une partie non négligeable soit présente dans le cluster 4 [9]. Concernant les bases dénotant des propriétés et des institutions, elles sont trop peu nombreuses pour tirer des conclusions au sujet de leur distribution. On peut aussi noter que certains clusters favorisent un type de base, notamment dans le cas des clusters 1 et 4, qui regroupent principalement des noms d'agent dérivés de bases actionnelles (à hauteur respectivement de 93% et de 90%, contre 80% et 75% pour les clusters 5 et 6, 51% pour le cluster 2, et 35% pour le cluster 3).

Certaines des observations présentées dans la section 8.4.2 et dans les tableaux 8.11, 8.12 et 8.13 convergent. Par exemple, la dénotation d'agents occasionnels, la concentration de noms d'agent en *-ant* et la sélection de bases verbales dynamiques dans le cluster 1 sont morphosémantiquement cohérentes. La même chose peut être dite pour la dénotation d'activités intellectuelles, la concentration de noms d'agent en *-ien* et *-iste*, et la prédilection pour les bases nominales dénotant des domaines dans le cluster 3.

Pour approfondir l'interaction de ces propriétés avec le clustering des noms d'agent, nous entraînons des modèles d'arbre d'inférence conditionnel [10]. Le but de ces modèles est de prédire le cluster le plus probable pour un nom d'agent sur la base de ses propriétés morphosémantiques. Cela permet d'identifier les propriétés qui influencent notre variable, ici l'appartenance à un cluster donné. Les arbres d'inférence conditionnels divisent de façon récursive les données en deux sous-ensembles significatifs, et ne s'arrêtent que lorsque les prédicteurs ne permettent plus aucun partitionnement significatif des données. L'arbre présenté en figure 8.2 a été entraîné avec le suffixe, la catégorie grammaticale et le type sémantique de la base comme prédicteurs, et le clustering comme variable de réponse. Chaque nœud représente un point de division des données. Le facteur prédictif est spécifié à chaque nœud, et les branches indiquent les valeurs pertinentes. Les histogrammes finaux présentent la distribution de la variable de réponse (ici les clusters).

L'arbre d'inférence conditionnel présenté en figure 8.2 confirme l'impor-

9. De la même manière que pour le suffixe *-eur* dans le tableau 8.11, une remarque peut être faite sur la sur-représentation des bases dénotant des actions. Ces dernières étant très majoritaires dans nos données, elles sont davantage représentées dans les clusters.

10. L'arbre d'inférence conditionnel est entraîné à l'aide de la fonction `ctree()` du package `party` de R

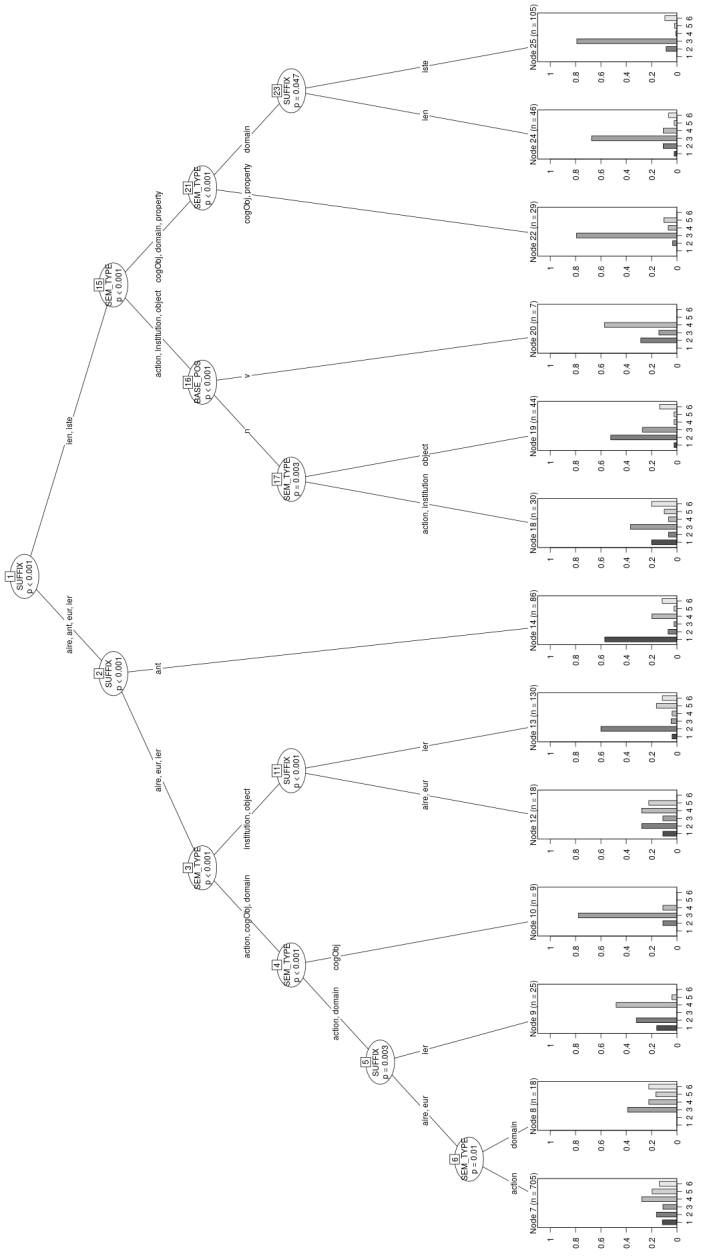

FIGURE 8.2 – Arbre d'inférence conditionnel

tance de certaines propriétés morphosémantiques dans la prédiction du clustering des noms d'agent, notamment les suffixes et les types sémantiques des bases.

On constate par exemple que les suffixes *-ien* et *-iste* d'une part, et *-aire* et *-eur* d'autre part, sont analysés comme des prédicteurs conjoints, accentuant la forte proximité au sein de ces deux paires de suffixes. Les suffixes *-ien* et *-iste* sont ainsi séparés des autres suffixes dès le nœud 1, et ne sont distingués qu'à partir du nœud 23, quand ils dérivent des noms à partir de bases dénotant des domaines. Cette distinction n'est cependant pas aussi significative que les autres distinctions, avec une p-value relativement élevée de 0.047. Les suffixes *-aire* et *-eur* sont distingués des autres suffixes aux noeuds 5 et 11, mais ne sont par la suite pas séparés, faisant de ces deux suffixes des prédicteurs similaires. Les noms d'agent en *-aire* et *-eur* ne sont à ce titre pas distingués sur la base de leur suffixe, mais sur la base du type sémantique de leur base morphologique (nœuds 3 et 6). Le suffixe *-ant* est remarquable dans la mesure où il est isolé à partir du nœud 2, et donne indépendamment la meilleure prédiction pour le cluster 1. Concernant le suffixe *-ier*, il est aussi analysé comme un prédicteur, mais seulement quand il est combiné à d'autres propriétés, à savoir le type sémantique de la base (nœuds 5 et 11).

Le type sémantique de la base est à ce titre un prédicteur fortement mobilisé. Par exemple, les bases dénotant des objets sont distinguées aux nœuds 3 et 17, celles dénotant des domaines aux nœuds 6 et 21, et celles dénotant des objets cognitifs aux nœuds 4 et 21. *A contrario*, le critère de la catégorie grammaticale n'est utilisé qu'une seule fois, lors du clustering des noms d'agent en *-ien* et *-iste* dérivés de bases dénotant des actions, des institutions ou des objets (nœud 16).

Le suffixe et le type sémantique de la base d'un nom d'agent sont souvent combinés pour la prédiction de certains clusters. Ainsi, les meilleurs prédicteurs pour le cluster 2 sont la présence du suffixe *-ier* combiné à l'utilisation de bases dénotant des institutions ou des objets (nœud 13), et les meilleurs prédicteurs du cluster 3 sont la présence des suffixes *-ien* ou *-iste* combinés à des bases dénotant des domaines ou des objets cognitifs (nœuds 22, 24 et 25). Tous les clusters ne sont cependant pas bien expliqués par les différents prédicteurs. C'est tout particulièrement le cas des clusters 5 et 6, pour lesquels l'arbre ne fait aucune prédiction forte. En d'autres termes, aucune de nos variables n'est clairement corrélée aux distinctions sémantiques caractérisant ces deux clusters. Cet arbre illustre le fait que les propriétés morphosémantiques étudiées ne permettent pas de rendre compte pleinement des distinctions à moyen grain pertinentes sur le plan distributionnel.

Pour déterminer précisément quels sont les facteurs prédictifs les plus significatifs dans le clustering, nous quantifions l'importance de nos prédic-

teurs. Pour cela, nous entraînons un modèle de random forest [11], qui calcule un grand nombre d'arbres d'inférences conditionnels, et ce afin d'approximer des mesures d'importance des variables [12]. Ces mesures quantifient l'impact des différents prédicteurs impliqués. Nous présentons ainsi l'importance de nos prédicteurs que sont le suffixe, la catégorie grammaticale et le type sémantique de la base dans la figure 8.3.

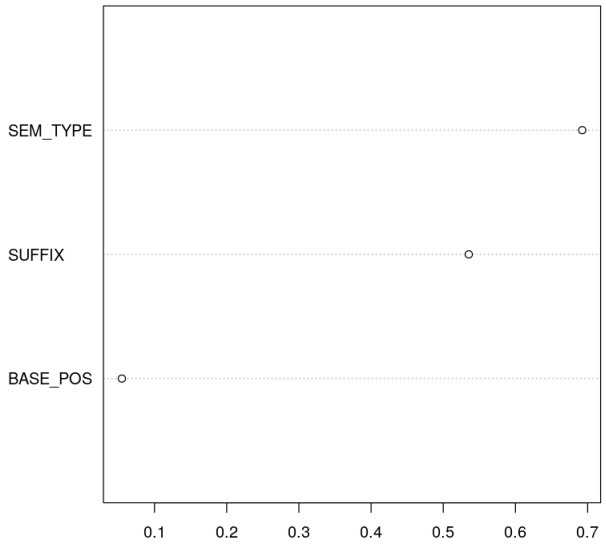

FIGURE 8.3 – Importance conditionnelle des variables

La figure 8.3 confirme que le type sémantique de la base et le suffixe sont les deux principales variables utilisées par l'algorithme pour prédire le clustering, le premier prévalant sur le second. Cela signifie que le clustering des noms d'agent est déterminé en grande partie par le type sémantique de leurs bases. En d'autres termes, le type sémantique des bases participe de la proximité des noms d'agent par le biais des sous-types ontologiques dénotés. On pourrait extrapoler de ces résultats que les noms d'agent dérivant de bases d'un même type sémantique ont plus de chances d'être sémantiquement proches que les noms d'agent dérivant de bases d'une même catégorie grammaticale – et donc qu'ils ont plus de chances d'être des concurrents, si l'on considère que la concurrence affixale se traduit par la formation de noms

11. Le modèle est entraîné à l'aide de la fonction `cforest()` du package `party` de R.
12. Nous mesurons l'importance des variables avec la fonction `varimp()` du package `party` de R.

sémantiquement proches. Par exemple, les suffixes *-eur* et *-aire* sont davantage en concurrence que les suffixes *-ier* et *-iste*. Contrairement à *-ier* et *-iste*, les suffixes *-eur* et *-aire* ne sélectionnent pas des bases de la même catégorie grammaticale, mais ils sélectionnent principalement des bases du même type sémantique. Si l'on comprend la concurrence des suffixes agentifs comme dépendant fondamentalement de l'output sémantique, on peut faire l'hypothèse que la concurrence entre suffixes agentifs du français dépend davantage du type sémantique de la base que de sa catégorie grammaticale.

Nous avons pu montrer dans ce chapitre que les noms d'agent formés par les suffixes *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste* se distinguaient morphosémantiquement. Que l'on parte des catégories formelles à l'aide des barycentres ou des catégories sémantiques à l'aide du clustering, nous avons fait émerger plusieurs groupes de noms d'agent se caractérisant par des propriétés morphologiques et sémantiques distinctes. Plus particulièrement, nous avons mis en évidence la convergence de certains suffixes comme *-ien* et *-iste* d'une part, et *-aire*, *-ant* et *-eur* d'autre part, mais aussi leur différenciation, relativement au type ontologique pour *-ien* et *-iste*, et à la sélection des bases pour *-ier*. Nous avons montré par ailleurs l'interaction de ces propriétés, et notamment l'impact du suffixe et du type sémantique de la base dans la construction de l'agentivité, et donc sur le type de noms d'agent construits par un procédé.

Evidemment, d'autres critères sont à prendre en compte (phonotactiques, diachroniques, etc). Un autre aspect sémantique, que nous n'avons pas approfondi jusque-là, émerge cependant des données, à savoir le type référentiel des agents dénotés. En effet, une autre façon d'envisager l'organisation interne de la catégorie lexicale des noms d'agent repose sur une distinction des noms dits statutaires, occasionnels et dispositionnels (Huyghe et Tribout 2015). Cette typologie distingue les noms d'agent selon qu'ils dénotent un statut (*coiffeur*), un événement particulier (*agresseur*), ou un trait comportemental (*farceur*).

L'examen des noms d'agent en *-aire*, *-ant*, *-eur*, *-ien*, *-ier* et *-iste* à l'aune de cette typologie montre qu'elle contribue à leur différentiation, comme nous l'ébauchons dans la section 8.4.2. Il apparaît par exemple que les noms en *-ant* dénotent de façon privilégiée des agents occasionnels (*assaillant*, *plaignant*, *poursuivant*), les noms en *-eur* des agents dispositionnels (*séducteur*, *gaffeur*, *bagarreur*), et les noms en *-aire* et *-ier* des agents statutaires (*bibliothécaire*, *bijoutier*). Cette propriété ressort aussi dans le clustering comme une caractéristique distinctive de certains clusters. Le cluster 1 agrège ainsi un grand nombre de noms occasionnels (*contributeur*, *participant*, *signataire*), et les clusters 4 et 5 des noms dispositionnels (*boudeur*, *glandeur*, *rêveur*).

Si cette distinction permettait d'appuyer ou d'expliquer les différences

entre certains suffixes, notamment pour les suffixes *-eur* et *-ant*, on peut se demander dans quelle mesure cette propriété est pertinente sur le plan distributionnel, et dans quelle mesure il s'agit ou non d'une observation idiosyncratique. En d'autres termes, on se demande si la différence entre noms statutaires, occasionnels et dispositionnels est perceptible dans les modèles distributionnels, et si oui, si ce trait sémantique est corrélé à des propriétés morphosémantiques des noms d'agent. Le chapitre 9 est consacré à cette question.

Chapitre 9

Sous-typologie agentive

Comme nous l'avons évoqué en fin de chapitre 8, plusieurs types référentiels de noms d'agent, statutaires, occasionnels et dispositionnels, ont été identifiés dans la littérature (Huyghe et Tribout 2015). Cette distinction, souvent envisagée d'un point de vue syntaxique, a été définie sur le plan lexical, bien que sa délimitation ne soit pas si nette. En particulier, le nombre de types impliqués, entre deux et trois catégories, est notamment discuté, et sa possible extension au delà des noms d'agent prototypiques en *-eur* n'est pas clairement établie.

Nous avons vu dans le chapitre 8 que la distinction entre les noms statutaires, occasionnels et dispositionnels semblaient contribuer à la description des classes de noms d'agent, autant par le biais des barycentres que des clusters. On peut donc se demander dans quelle mesure c'est un artefact dans nos analyses, ou c'est effectivement un trait sémantique explicatif sur le plan distributionnel. Cela pose donc la question de la pertinence de cette distinction dans les modèles distributionnels d'une part, et de l'extension de cette typologie, tant sur le plan lexical (en termes de nombre de sous-types) que formel (en termes de propriétés morphosémantiques).

Nous nous proposons de creuser ces questions dans ce chapitre. Nous testons l'hypothèse d'une tripartition sémantique des noms d'agent sur cette notion de statut, d'occasionnalité et de disposition, en étudiant sur le plan distributionnel ces trois catégories. Nous présentons dans un premier temps la constitution des trois groupes qui servent de base à cette étude (section 9.2). Nous proposons ensuite d'apporter des preuves distributionnelles de cette tripartition (section 9.3). Nous étudions enfin l'extension de la catégorie (section 9.4).

9.1 Homogénéité sémantique de la classe des noms d'agent

La distinction de trois sous-types d'agent au sein de la catégorie lexicale des noms d'agent repose sur la relation entretenue par l'agent dénoté vis-à-vis de l'action qu'il réalise. Cette distinction a d'abord été envisagée d'un point de vue sémantique (section 9.1.1), mais a aussi été décrite d'un point de vue syntaxique (section 9.1.2). Ces conceptions ont mené à la description de trois sous-types dont nous reprenons les principales caractéristiques en section 9.1.3.

9.1.1 Perspective lexicale

Comme évoqué dans la section 8.1, Benveniste (1975) établit pour les langues indo-européennes la distinction entre deux types de noms d'agent : ceux qui dénotent l'agent d'une fonction, suffixés en *-*ter*, et ceux qui dénotent l'auteur d'une action, suffixés en *-*tor*. Les premiers intègrent l'agentivité comme une propriété constitutive (intrinsèque dans la terminologie de Anscombre 2003) du référent, caractérisée par l'action dénotée, qu'elle soit actualisée ou non, alors qu'elle correspond à une propriété ajoutée (extrinsèque dans la terminologie de Anscombre 2003) pour les seconds, l'individu actualisant l'action donnée à un moment précis, sous la forme d'un événement particulier. Si Benveniste (1975) indique que cette distincton n'est désormais plus formellement marquée, elle se maintiendrait dans le lexique, et serait perceptible en français dans certains cas d'allomorphie notamment, comme pour les noms *sauveteur* et *sauveur*. Le référent du premier est caractérisé par sa fonction de sauveteur, quand bien même ce référent n'a jamais sauvé personne, alors que qu'un référent ne sera désigné par le nom sauveur que sur la base de la réalisation d'un sauvetage.

Benveniste (1975) précise cette distinction entre propriétés constitutives et ajoutées, illustrée par *sauveteur* et *sauveur*, par une seconde distinction sémantique, à savoir la notion d'objectivité et de subjectivité. Affichant un certain recouvrement avec la première distinction, cette opposition permet de caractériser un usage référentiel (ou objectif) et un usage attributif (ou subjectif) du nom. La caractérisation sera objective si elle décrit ce que fait un individu, sans pour autant altérer l'identité propre de l'individu. Cela est illustré par *sauveur*, dont le caractère ponctuel l'action ne vient pas remettre en question l'identité du référent. À l'inverse, la caractérisation sera subjective si elle définit l'individu, si cela abolit son identité à l'image de *sauveteur*, dont la fonction sert à identifier l'individu. De fait, dans cette perspective,

les propriétés ajoutées (ou extrinsèques) tendent à être objectives, et les propriétés constitutives (ou intrinsèques) subjectives.

Anscombre (2003) ébauche un troisième type de noms d'agent en complétant l'opposition en propriétés intrinsèque et extrinsèque par une opposition entre propriétés essentielles (partagées par tous les individus) et propriétés accidentelles (que certains ont, mais que tout le monde ne partage pas). La combinaison de ces deux séries de propriétés permet d'expliciter les différences sémantiques entre paires de mots en apparence concurrents. Ainsi, les adjectifs *maladif* et *malade* présentent des propriétés accidentelles (qui ne sont pas partagées par tout le monde) respectivement intrinsèques et extrinsèques. Autrement dit, l'adjectif *maladif* fait référence à une caractéristique constitutive de l'individu dont il est question, alors que l'adjectif *malade* dénote un trait temporaire. Notons que dans la mesure où l'explicitation des propriétés intrinsèques essentielles comme l'ouïe et la vision, par nature universelles, n'est pas pertinente (Anscombre 2001), les noms d'agent porteurs de ces propriétés comme *voyeur*, *entendeur* ou *coureur* ne désignent pas l'agent d'une fonction ('celui qui voit' pour voyeur) ou l'acteur d'une action ('celui qui a vu') mais l'agent qui présente un comportement spécifique en lien avec cette action ('personne qui aime regarder les gens').

9.1.2 Perspective syntaxique

Si la distinction de différents types de noms d'agent a d'abord émergé sur un plan lexical, elle a trouvé des échos sur le plan syntaxique. De façon similaire à ce que l'on observe pour le français, Rappaport Hovav et Levin (1992) séparent les noms d'agent en *-er* en deux groupes, les noms événementiels (*saver of life* 'sauver') et non-événementiels (*lifesaver* 'sauveteur'), sur la base de la référence à un événément actualisé. Les auteurs affirment que cette distinction est corrélée à la disponibilité de la structure argumentale. Selon les auteurs, seuls les noms événementiels ont des compléments qui dénotent les participants de l'action dénotée.

Alexiadou et Schäfer (2010) contestent cette hypothèse, affirmant au contraire que les deux types de noms ont les mêmes types de structures argumentales, et que la différence est d'ordre aspectuel. Les noms événementiels seraient caractérisés par leur aspect épisodique (c'est-à-dire occurrentiel), et les noms non-événementiels par leur aspect dispositionnel. Roy et Soare (2012) proposent une analyse similaire pour les noms d'agent en *-eur* du français, et soutiennent que la différence entre les noms d'agent en *-eur* épisodiques et dispositionnels relèveraient de l'interprétation spécifique ou générique du complément. Le nom *vendeur* serait ainsi épisodique dans *le vendeur de ce bien immobilier*, mais dispositionnel dans *un vendeur de*

journaux.

Si la distinction entre agents épisodiques et dispositionnels permet ici de rendre compte d'un point de vue syntaxique la formation des noms d'agent, elle coincide avec la perspective lexicale d'une distinction fondée sur la réalisation ou non d'un événement spécifique, qui se traduit donc par des propriétés sémantiques distinctes. Par ailleurs, notons que dans l'exemple proposé par Roy et Soare (2012), la différence entre la lecture épisodique ou dispositionnelle du nom *vendeur* est relative au contexte. Quelle que soit la perspective selon laquelle elle est étudiée, cette distinction se traduit en termes de distribution. Nous pouvons donc faire l'hypothèse qu'elle sera perceptible dans les modèles distributionnels.

9.1.3 Tripartition des noms d'agent

Huyghe et Tribout (2015) reprennent les différentes distinctions mises en évidence, et délimitent ainsi trois groupes de noms d'agent : les noms statutaires, les noms d'agents occasionnels et les noms dispositionnels.

Les noms statutaires correspondent aux noms caractérisés par leur propriété constitutive dans le travail de Benveniste (1975), aux noms non-événementiels chez Rappaport Hovav et Levin (1992), et aux noms dispositionnels chez Alexiadou et Schäfer (2010) et Roy et Soare (2012). De fait, les noms statutaires dénotent les agents d'une fonction, définis par l'absence de réalisation événementielle particulière, et par une interprétation habituelle, générique ou définitionnelle. Cela se traduit par certains emplois, illustrés dans les exemples 33 (tirés de Huyghe et Tribout 2015), comme l'effacement possible du déterminant (33a), la non compatibilité avec des arguments spécifiques (33b), la possibilité de la présence d'un complément du nom sous une forme générique (33c) et l'inscription dans des syntagmes nominaux génériques, indéfinis existentiels (33d).

(33) a. Pierre est brocanteur.

b. ?le déménageur de ces meubles

c. un carreleur de piscine

d. Les serveurs sont parfois distraits.

Les noms d'agents occasionnels tels que définis par Huyghe et Tribout (2015) ne désignent pas un statut, mais l'instanciation occasionnelle de l'action décrite par le verbe de base. Cette sous-classe recouvre donc les noms aux propriétés ajoutées dans l'approche de Benveniste (1975), les noms événementiels de Rappaport Hovav et Levin (1992) et les noms épisodiques chez Alexiadou et Schäfer (2010) et Roy et Soare (2012). Ils ne permettent en

cela pas la désignation d'un référent précis, à l'image de (34a) et (34c), et admettent, voire nécessitent, l'utilisation de compléments spécifiques (35b), comme le montrent les exemples en 34 (tirés de Huyghe et Tribout 2015).

(34) a. *Pierre est agresseur.
b. l'agresseur de Pierre
c. ? ?les dénicheurs sont parfois arrogants.

Enfin, Huyghe et Tribout (2015) identifient des noms dispositionnels, qui sont des noms d'agent en *-eur* dont l'interprétation n'est ni statutaire ni occassionnelle. Ils semblent se situer entre les deux interprétations, puisqu'une lecture habituelle (et non occurrentielle) est possible, mais sans qu'il s'agisse d'une position institutionnelle. Ces noms coincident du moins partiellement avec la catégorie identifiée par Anscombre (2001) au travers des noms aux propriétés intrinsèques essentielles. Les compléments qu'ils acceptent doivent avoir une forme générique ((35a) et (35c)), et le déterminant ne peut s'effacer (35b). Enfin, l'emploi d'adjectifs de taille amène à une lecture fréquentielle, intensive ou habituelle de l'action décrite ((35d)) (tirés de Huyghe et Tribout 2015). Ces noms peuvent par ailleurs facilement former des adjectifs par conversion.

(35) a. Un séducteur de jeunes filles
b. *Pierre est séducteur.
c. ? ?le séducteur de Sophie
d. Pierre est un grand séducteur.

Précisons que ces trois catégories ne sont pas exclusives, des noms pouvant être porteurs de plusieurs lectures, à l'image de *inventeur*.

La définition de ces caractéristiques permet d'expliquer certains phénomènes quant à l'association du nom d'agent avec d'autres éléments, comme des adjectifs ou des compléments du nom, ou encore des phénomènes liés à leur fonctionnalité (l'admission d'une reprise anaphorique par exemple) comme le montre Anscombre (2001). De fait, on peut faire l'hypothèse que ces différences se traduiront dans les espaces vectoriels. Pour cela, nous évaluons à l'aide de barycentres les profils distributionnels des trois classes afin de voir quelles sont leurs propriétés et spécificités.

9.2 Identification des trois sous-classes

Suivant Huyghe et Tribout (2015), nous partons du postulat qu'il existe une tripartition selon les noms d'agent statutaires, occasionnels et dispositionnels. Nous considérons ici ces trois groupes de noms comme trois classes

sémantiques distinctes, dont il va s'agir de voir si elles sont aussi distributionnellement et morphosémantiquement distinctes. Pour ce faire, nous devons partir de noms prototypiquement statutaires, occasionnels et dispositionnels qui soient représentatifs de ces trois groupes. Pour cette raison, nous procédons à une première étape de sélection des noms.

Pour leur identification, nous nous basons les conditions 1 à 4 ci-dessous, qui sont suffisantes pour la catégorisation respective en tant que statutaire, occasionnel et dispositionnel.

— Condition 1 : le nom peut être utilisé comme un prédicat nu sans aucun complément (*X est N* → *Pierre est coiffeur*).

— Condition 2 : le nom est compatible avec un complément spécifique qui dénote un participant dans un événement particulier (*le N de x* → *l'agresseur de Pierre*).

— Condition 3 : le nom est compatible avec un adjectif de taille dans une lecture non intersective, c'est-à-dire ne qualifiant pas sur la taille de l'agent mais la fréquence ou l'intensité de l'action, sans l'ajout d'un complément (*un gros N* → *un gros bosseur*).

— Condition 4 : le nom est dérivé d'un verbe transitif.

Soulignons que l'absence de complément est requise pour les conditions 1 et 3 pour éviter des effets de coercion tels qu'illustrés en (36), où les noms d'agent occasionnels (*agresseur*, *envoyeur*, *acheteur*) dérivés de verbes transitifs peuvent être interprétés comme des noms d'agent dispositionnels quand ils sont utilisés avec des compléments indéfinis.

(36) Cet homme est un grand agresseur de personnes âgées / un gros envoyeur d'emails / un gros acheteur de voitures anciennes

La condition 4 permet de garantir la monosémie des noms validant les conditions 1, 2 ou 3. Dans le cas de verbes intransitifs, seuls les noms ne validant pas les conditions 1 et 3 peuvent être identifiés, par défaut comme des noms occasionnels monosémiques, qu'ils valident ou non la condition 2.

Le tableau 9.1 présente les diagnostics qui peuvent être inférés sur la base de ces quatre conditions.

À partir de cette grille d'identification, nous sélectionnons 50 noms d'agent déverbaux monosémiques en *-eur* de chaque type, parmi ceux utilisés dans le chapitre 7. Nous nous limitons à 50 car de nombreux noms d'agent sont polysémiques au regard de cette typologie (*animateur*, *racketteur* et *inventeur* étant respectivement statutaire et occasionnel, occasionnel et dispositionnel, et statutaire, occasionnel et dispositionnel), et il n'est donc pas évident d'avoir un échantillon équilibré plus important. Nous basons dans la suite notre analyse sur l'étude distributionnelle des 150 noms ainsi sélectionnés.

Sous-type	Cond. 1	Cond. 2	Cond. 3	Cond. 4
Statutaire	1	0	0	1
Dispositionnel	0	0	1	1
Occasionnel	0	1	0	1
	0	0/1	0	0

TABLE 9.1 – Grille d'identification des sous-types monosémiques des noms d'agent déverbaux

9.3 Pertinence de la typologie

Nous faisons l'hypothèse que la pertinence distributionnelle d'une classe peut être testée à l'aide de deux critères : l'homogénéité sémantique de la classe, et sa singularité vis-à-vis d'autres classes. Ainsi, pour évaluer la pertinence des trois classes formées à l'aide des noms d'agent statutaires, occasionnels et dispositionnels, nous cherchons à montrer d'une part que chacune d'elles forme un ensemble relativement homogène sur le plan distributionnel, et d'autre part qu'il existe des différences significatives entre ces trois classes sur le plan distributionnel.

Pour voir si les noms d'agent sélectionnés forment trois ensembles pertinents sur le plan distributionnel, nous adoptons dans un premier temps une approche *bottom-up* de clustering (section 9.3.1), afin de tester la capacité des modèles distributionnels à discriminer les noms d'agent sur la base de ce trait sémantique. Nous évaluons ensuite dans quelle mesure les classes formées par ces noms d'agent diffèrent sur le plan distributionnel en comparant les propriétés morphosémantiques de leurs barycentres (section 9.3.2).

9.3.1 Clustering des noms d'agent statutaires, occasionnels et dispositionnels

Sur la base des 150 noms d'agent monosémiques précédemment sélectionnés, nous souhaitons évaluer dans quelle mesure leurs propriétés distributionnelles permettent de les discriminer. Nous faisons l'hypothèse que si les traits statutaire, occasionnel et dispositionnel sont significatifs dans la représentation distributionnelle de ces noms, ils devraient permettre de discriminer les noms. En d'autres termes, nous nous attendons à ce que les noms soient regroupés sur la base de ces traits, et que chacune des catégories fasse l'objet d'un cluster distinct.

Nous opérons un clustering des vecteurs des 150 noms d'agent à l'aide du *spherical kmeans*. Nous fixons à 3 le nombre c de clusters sur la base de

l'hypothèse que ces noms forment trois classes. Comme précédemment, nous répétons l'opération sur la base des vecteurs calculés dans les cinq modèles distributionnels afin de tester la stabilité des résultats. La variation liée aux modèles est évaluée à l'aide de l'indice de Rand dans le tableau 9.2.

	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5
Modèle 1	1	0.715	0.649	0.686	0.781
Modèle 2	0.715	1	0.652	0.576	0.721
Modèle 3	0.649	0.652	1	0.525	0.613
Modèle 4	0.686	0.576	0.525	1	0.655
Modèle 5	0.781	0.721	0.613	0.655	1

TABLE 9.2 – Indice de Rand du clustering des noms d'agent statutaires, occasionnels et dispositionnels sur 5 modèles

Le tableau 9.2 montre que le partitionnement réalisé sur les différents modèles varie de façon plus ou moins importante. Ainsi, on observe un accord bien moins important entre les modèles 3 et 4 (0.525) qu'entre les modèles 1 et 5 (0.781). Globalement, le clustering semble moins stable que celui obtenu dans la section 8.4.1. Nous faisons donc ici le choix de ne pas nous baser sur les résultats d'un unique modèle, qui ne serait donc pas nécessairement représentatif, mais de d'examiner le clustering obtenu dans chacun des modèles.

La répartition des noms d'agent dans les clusters en fonction de leur type, et ce dans chaque modèle, est donnée dans les tableaux présentés en 9.3. Précisons qu'il n'y a pas d'équivalence, d'un modèle à l'autre, entre les étiquettes C1, C2 et C3 puisque celles-ci sont arbitraires et permettent simplement d'identifier les regroupements de noms. Le nombre total (T) d'items est indiqué pour chaque cluster.

L'analyse du partionnement des 150 noms d'agent dans les cinq modèles sur la base des tableaux 9.3 montre une certaine variation. Le calcul du χ^2 de Pearson montre cependant une relation significative entre le clustering et le type sémantique des noms d'agent (p-value < 0.05) dans les cinq modèles. Il y a donc une corrélation entre le regroupement des noms d'agent et leur type référentiel. On peut donc faire l'hypothèse que la nature statutaire, occasionnelle ou dispositionnelle des noms d'agent se traduit sur le plan distributionnel et permet de discriminer les noms. Notons que les effectifs sont inégaux, avec souvent un cluster bien plus peuplé que les autres. C'est notamment le cas du cluster 2 dans le modèle 3 qui contient plus de 56% des noms, alors que la répartition est bien plus équilibrée dans le modèle 4. Des tendances émergent cependant.

	Modèle 1		
	C1	**C2**	**C3**
S	4	4	42
O	28	8	14
D	16	23	11
T	48	35	67

	Modèle 2		
	C1	**C2**	**C3**
S	6	42	2
O	8	11	31
D	20	4	26
T	34	57	59

	Modèle 3		
	C1	**C2**	**C3**
S	19	30	1
O	9	33	8
D	5	22	23
T	33	85	32

	Modèle 4		
	C1	**C2**	**C3**
S	10	36	4
O	18	18	14
D	13	23	14
T	41	77	32

	Modèle 5		
	C1	**C2**	**C3**
S	36	13	1
O	10	26	14
D	6	19	25
T	52	58	40

TABLE 9.3 – Distribution des noms d'agent statutaires (S), occasionnels (O) et dispositionnels (D) dans les 3 clusters C1, C2 et C3 sur les 5 modèles

Les clusters obtenus dans le modèle 1 suggèrent une certaine discrimination des noms d'agent statutaires, occasionnels et dispositionnels. En effet, 84% des noms statutaires se trouvent dans le cluster 3, 56% des noms occasionnels dans le cluster 1, et 46% des noms dispositionnels dans le cluster 2. En retour, le cluster 1 est majoritairement constitué de noms occasionnels (58%), le cluster 2 de noms dispositionnels (66%), et le cluster 3 de noms statutaires (63%). Chaque classe semble ainsi occuper un cluster distinct, et chaque cluster semble être majoritairement dédié à un type spécifique, bien que cela soit moins net pour les noms dispositionnels. Notons que nous faisons un constat similaire en ce qui concerne la répartition des noms d'agent dans les clusters du modèle 5.

A contrario, la spécificité de certains clusters est moins claire dans le modèle 2. Ainsi, si 84% des noms statutaires sont regroupés dans un même cluster (C2), respectivement 62% et 52% des noms d'agent occasionnels et dispositionnels sont regroupés dans un même et unique cluster (C3). Malgré cela, on peut remarquer que le cluster 1 est composé à 59% de noms dispositionnels, le cluster 2 à 74% de noms statutaires, et le cluster 3 à 53% de noms occasionnels. La distinction entre noms occasionnels et dispositionnels est ici moins nette.

Les clusters du modèle 3 se caractérisent eux aussi par une incapacité à capter distinctement les différences entres les trois catégories, l'intersection concernant ici les noms statutaires et occasionnels. En effet, on constate que 60% des noms statutaires se trouvent dans le cluster 2, au même titre que

66% des noms occasionnels. Bien que les noms dispositionnels soient en plus grande partie dans le cluster 3, à raison de 46%, on en trouve tout de même 44% dans le cluster 2. Le cluster 2 semble donc ici particulièrement hétérogène, regroupant de façon presque équitable les trois types sémantiques. Les clusters 1 et 3 sont cependant de fait bien plus spécialisés, le premier intégrant à hauteur de 58% des noms statutaires, et le second à hauteur de 72% des noms dispositionnels.

Le modèle 4 présente lui un certain recouvrement entre les noms statutaires et dispositionnels. Les noms occasionnels sont distribués de façon similaire dans les 3 clusters (36%, 36% et 28%), et la plus grande proportion de noms statutaires et dispositionnels – respectivement 72% et 46% – sont regroupés dans le même cluster (C2). Si le cluster 3 contient autant de noms occasionnels que dispositionnels (à hauteur de 44% chacun), les clusters 1 et 2 sont tous les deux davantage spécialisés, le premier autour des noms occasionnels (44%), et le second autour des noms statutaires (47%).

Il apparaît à ce stade que les trois sous-classes se distinguent sur le plan distributionnel, mais que la clarté de la discrimination deux à deux des catégories varie en fonction des modèles. Afin de préciser ces résultats, nous conduisons une analyse par cluster des sous-classes deux par deux. En d'autres termes, nous demandons à l'algorithme de partitionner les 100 noms issus de deux des trois sous-classes en 2 clusters. Le calcul du χ^2 de Pearson (avec la correction de continuité de Yates) montre que la distinction entre les noms d'agent statutaires et dispositionnels est significative dans les 5 modèles, alors que la distinction entre les noms d'agent statutaires et occasionnels n'est significative que dans 4 modèles, et celle entre noms d'agent occasionnels et dispositionnels ne l'est que dans 3 modèles (avec p-value$<$0.05).

Globalement, les résultats du clustering soutiennent la distinction entre noms d'agent statutaires, occasionnels et dispositionnels. Ils montrent que les classes ne sont pas toutes aussi distinctes les unes que les autres, les noms d'agent statutaires étant plus clairement séparés des deux autres classes que ne le sont les deux autres sous classes. Néanmoins, ces spécificités n'invalident pas la pertinence distributionnelle de la tripartition, dont on peut assumer (du moins, dans une certaine mesure) qu'elle est déterminée par une distinction sémantique entre les trois classes.

9.3.2 Barycentres des noms statutaires, occasionnels et dispositionnels

Pour approfondir l'hypothèse d'une tripartition, nous reprenons la méthode d'analyse de classes de mots par le biais de leurs réprésentations distri-

butionnelles présentée dans les chapitres précédents. Nous construisons donc les barycentres des noms d'agent statutaires, occasionnels et dispositionnels, et en calculons les 100 plus proches voisins moyens sur les cinq modèles. Les 10 plus proches voisins moyens des trois barycentres sont donnés dans le tableau 9.4.

Statutaire	Occasionnel	Dispositionnel
coiffeur	assassin	farceur
plombier	meurtrier	filou
garagiste	geôlier	hâbleur
soigneur	bourreau	râleur
contremaître	tortionnaire	fanfaron
cuisinier	agresseur	vantard
dentiste	imposteur	séducteur
masseur	complice	poltron
apprenti	traître	menteur
magasinier	héros	ivrogne

TABLE 9.4 – Dix plus proches voisins des barycentres des noms d'agent déverbaux monosémiques en *-eur* statutaires, occasionnels et dispositionnels

Nous nous posons la question de savoir si l'espace sémantique que l'on observe par le biais du barycentre est bien caractérisé par les traits sémantiques visés, à savoir le caractère statutaire, occasionnel ou dispositionnel de l'agentivité. Une première façon de procéder consiste à étudier le recouvrement entre les amorces (dont on connaît le caractère statutaire, occasionnel ou dispositionnel) et les voisins des barycentres. Ce taux de recouvrement est donné dans le tableau 9.5.

Voisins / Amorces	Statutaires	Occasionnels	Dispositionnels
Statutaires	11	2	-
Occasionnels	-	7	-
Dispositionnels	1	1	3

TABLE 9.5 – Recouvrement entre les amorces et les voisins des barycentres des noms d'agent statutaires, occasionnels et dispositionnels

Le tableau 9.5 montre que 11 des amorces ayant servi à construire le barycentre des noms d'agent statutaires se retrouvent dans son voisinage,

contre sept pour les noms occasionnels, et trois pour les noms dispositionnels. Cela signifie qu'au moins 11, 7 et 3 des voisins des barycentres des noms statutaires, occasionnels et dispositionnels sont eux-mêmes respectivement des noms statutaires, occasionnels et dispositionnels. Mais on peut aussi noter la bonne séparation des catégories. Plus précisément, on ne retrouve qu'une seule amorce dispositionnelle dans le voisinage du barycentre des noms d'agent statutaires, une amorce dispositionnelle dans celui des noms occasionnels, et deux amorces statutaires dans celui des noms occasionnels. Les barycentres semblent donc dessiner entre eux des zones de voisinage relativement imperméables.

L'analyse du recouvrement entre amorces et voisins ne suffit cependant pas à tirer des conclusions sur les propriétés statutaires, occasionnelles ou dispositionnels de ces barycentres puisque d'une part les amorces sont relativement peu nombreuses, et d'autre part leur sélection reposait sur des critères assez restrictifs de monosémie. Ces amorces ne sont donc pas pleinement représentatives de ce qui se passe dans le lexique, les noms d'agent monosémiques n'étant pas majoritaires, d'autant plus au regard de cette typologie, et ne peuvent suffire à l'évaluation du voisinage. Pour évaluer la sensibilité des barycentres à ces trois traits sémantiques, nous procédons à une analyse plus globale des voisins au regard de ces traits, par rapport aux conditions 1 à 3 décrites précédemment.

Cette analyse requiert cependant une première phase de filtrage du voisinage. En effet, la caractérisation selon cette typologie et ses conditions d'identification ébauchées dans la section 9.2 ne s'applique qu'aux noms d'agent. Or, les 100 plus proches voisins de ces barycentres ne sont pas tous des noms d'agent. Nous nous limitons donc pour notre analyse aux voisins pertinents. Nous appelons « voisins pertinents »les voisins qui sont des noms et qui sont candidats à l'agentivité. En d'autres termes, nous excluons des analyses suivantes les mots qui ne sont pas automatiquement étiquetés comme noms (par exemple les ajectifs *vaniteux*, *cupide*, *fourbe*, dans le voisinage du barycentre des noms d'agent dispositionnels), ainsi que les noms qui ne dénotent pas des agents – en l'occurrence des noms qui dénotent des éventualités (*ingratitude*, *manigance*, *infortune* dans le voisinage des noms d'agent occasionnels) et des noms strictement relationnels (*proche*, *ami*, *amoureux* et *protégé* dans le voisinage des noms occasionnels).

De fait, l'ensemble des 100 plus proches voisins du barycentre des noms d'agent statutaires se révèlent des voisins pertinents, contre 90 pour le barycentre des noms d'agent occasionnels, et 79 pour celui des noms d'agent dispositionnels. À ce titre, l'ensemble des voisins présentés dans le tableau 9.4 sont considérés comme des voisins pertinents. La suite des analyses se fait donc sur la base de ces voisins pertinents.

L'annotation des voisins pertinents au regard des conditions 1 à 3 (cf. p.184) est donnée dans le tableau 9.6. Notons qu'un nom peut ne valider aucune condition (*sbire*, *complice*) ou au contraire en validant plusieurs du fait de la polysémie entre les trois sous-classes (*rêveur* valide les conditions 2 et 3, et *cambrioleur* les trois conditions). Ces noms font donc l'objet de plusieurs annotations. Nous n'annotons pas la condition 4 car nous ne cherchons pas à diagnostiquer la monosémie.

	Cond. 1	Cond. 2	Cond. 3
Statutaires	98	6	9
Occasionnels	24	29	43
Dispositionnels	12	3	71

TABLE 9.6 – Nombre de voisins pertinents des barycentres des noms d'agent statutaires, occasionnels et dispositionnels validant les conditions 1 à 3

La distribution des voisins pertinents en fonction des conditions qu'ils valident se révèle statistiquement significative (p-value$<2.2^{-16}$ au χ^2 de Pearson). En d'autres termes, davantage de voisins d'une sous-classe donnée valident la condition nécessaire à la classification dans cette sous-classe donnée (voir tableau 9.1) que les autres voisins, et ce de façon significative.

Le tableau 9.6 montre que les voisins du barycentre des noms d'agent occasionnels sont plus uniformément distribués que les autres voisins au regard des Conditions 1 à 3. Cela est cohérent avec notre précédente remarque concernant le fait que les noms d'agent occasionnels sont moins homogènes en termes de distribution linguistique que les autres noms d'agent. Le fait que 29 voisins valident la condition 2 peut sembler assez faible dans le cas des voisins occasionnels. Il faut cependant rappeler que la condition 2, contrairement aux conditions 1 et 3 pour les noms statutaires et dispositionnels, n'est pas nécessaire pour la classification en tant que noms occasionnels – c'est-à-dire que les noms d'agent occasionnels dérivés de verbe intransitif ont peu de chance de valider cette condition. Néanmoins, le fait demeure qu'un nombre significativement plus important de voisins du barycentre des noms d'agent occasionnels valident la condition 2, ce qui suggère une plus grande sensibilité à ce trait. Les résultats sont somme toute congruents avec l'idée que les différences distributionnelles entre les noms d'agent statutaires, occasionnels et dispositionnels peuvent en grande partie expliquer la partition de la classe des noms d'agent.

Deux observations corroborent cette affirmation. Tout d'abord, on remarque que le voisinage du barycentre des noms d'agent dispositionnels

contient 18 adjectifs (contre 0 pour les autres barycentres), ainsi que de nombreux noms qui sont convertis à partir d'adjectifs ou souvent utilisés comme adjectifs (*maniaque*, *sadique*, *teigneux*, *débrouillard*, *rêveur*, *ronchon*) [1]. Cela peut s'expliquer par le fait que les noms d'agent dispositionnels décrivent des attitudes agentives qui sont associées à des habitudes et des façons de se comporter. Ils sont proches de mots dénotant des propriétés, étant utilisés pour caractériser des référents d'une façon comparable aux adjectifs. Deuxièmement, les 10 voisins non pertinents présents dans le voisinage du barycentre des noms d'agent occasionnels sont des noms, parmi lesquels au moins 5 ont au moins un sens événementiel (contre 0 pour les autres barycentres), c'est-à-dire qu'ils sont compatibles avec *avoir lieu* ou *se produire*, et peuvent dénoter l'occurrence d'un événement. Ces noms (*agissement*, *infortune*, *malheur*, *manigance*, *méfait*), qu'ils impliquent ou non un agent, partagent avec les noms d'agent occasionnels le fait de faire référence à un événement particulier, ce qui pourrait expliquer leur présence dans le voisinage du barycentre des noms d'agent occasionnels – si ce voisinage, comme nous le soutenons, est lié à la dénotation d'occurrences particulières d'actions.

Des remarques additionnelles peuvent être faites concernant la sous-classe des noms d'agent statutaires. Comme confirmation de leur homogénéité, le voisinage de ce barycentre semble être à la fois plus cohérent et plus distant des autres que ne le sont les deux autres barycentres. Il n'y a aucun voisin non pertinent dans son voisinage, contrairement à ce qui est observé pour les deux autres sous-classes. Par ailleurs, les scores de proximité moyens sur cinq modèles des barycentres des noms d'agent statutaires et occasionnels (0.637), et des barycentres des noms d'agent statutaires et dispositionnels (0.629) sont plus faibles que celui entre les barycentres des noms d'agent occasionnels et dispositionnels (0.702). Cette distance est confirmée par le taux de recouvrement variable entres les voisinages des différentes sous-classes, présenté dans le tableau 9.7.

Les observations convergent pour montrer que les noms d'agent statutaires forment la sous-classe de noms d'agent la plus distinctive, mais aussi la plus représentée. En effet, la comparaison des 100 plus proches voisins des barycentres des noms d'agent statutaires avec ceux obtenus en section 7.3 à partir de l'ensemble des noms d'agent déverbaux monosémiques en *-eur* montre de plus fortes similitudes que dans le cas des deux autres sous-types. En effet, le barycentre des noms d'agent déverbaux monosémiques en *-eur* partage 42 voisins avec le barycentre des noms d'agent statutaires, contre 30

1. L'identité formelle entre noms et adjectifs peut être à l'origine d'erreurs de tagging. Nous n'avons pas exploré ce point, et nous nous fions entièrement à l'annotation fournie par le parseur, ce qui est cohérent avec la décision prise dans ce chapitre d'analyser des vecteurs de mots étiquetés automatiquement comme noms.

	Statutaires	Occasionnels	Dispositionnels
Statutaires	-	3	2
Occasionnels	3	-	16
Dispositionnels	2	16	-

TABLE 9.7 – Nombre de voisins partagés par les barycentres des noms d'agent statutaires, occasionnels et dispositionnels

avec celui des noms occasionnels et 24 avec celui des noms dispositionnels. Puisque nous sommes partis du même nombre de noms statutaires, occasionnels et distributionnels, on peut faire l'hypothèse que cette plus grande similarité s'explique par une différence de proportion des trois sous-types dans l'échantillon de noms déverbaux monosémiques, le type statutaire étant potentiellement plus représenté que les types occasionnel et dispositionnel. Corollairement, on peut faire l'hypothèse que ce sous-type est davantage représenté dans le lexique.

Une remarque additionnelle peut être faite concernant la connotation morale des voisins des barycentres des noms d'agent statutaires, occasionnels et dispositionnels. Alors que les voisins du barycentre des noms d'agent statutaires ne sont globalement pas connotés (*plombier*, *couturier*, *radiologue*), les voisins des noms d'agents occasionnels incluent de nombreux noms qui font référence à des actions nuisibles (*assassin*, *fraudeur*, *oppresseur*), et les voisins des noms dispositionnels des noms qui décrivent des tendances à agir d'une mauvaise façon ou d'une façon dévalorisée par la société (*arriviste*, *goinfre*, *fainéant*). La même propriété peut être trouvée dans certaines des amorces occasionnelles (*agresseur*, *kidnappeur*, *saboteur*) et dispositionnelles (*baratineur*, *emmerdeur*, *magouilleur*), mais c'est davantage proéminent chez les voisins. Une association lexicale récurrente semble apparaître entre la dénotation des noms d'agent occasionnels et dispositionnels et la référence à des actions nuisibles ou des fautes morales sur autrui. Bien que les dénominations connotées positivement, ou neutres, existent chez les noms d'agent occassionnels et dispositionnels du français (*bienfaiteur*, *mangeur*), il se pourrait qu'ils ne constituent qu'une minorité, ou qu'il s'agisse d'un groupe moins saillant parmi les noms d'agent occassionnels et dispositionnels.

De fait, si on croise la capacité de *-eur* à former des noms dispositionnels, et la sur-représentation des noms en *-eur* dans le lexique, cela explique le constat fait dans la section 7.3.2 d'une connotation négative des voisins du barycentre des noms d'agent en *-eur* monosémiques. Ces observations confortent par ailleurs les observations faites sur la différence de connotation des noms en *-euse* et *-rice*, les premiers tendant à désigner des agents

statutaires (*directrice*), les seconds des agents occasionnels (*fraudeuse*) ou dispositionnels (*allumeuse*).

9.4 Extension morphologique de la typologie

Nous avons établi la pertinence sur le plan distributionnel de cette tripartition sémantique des noms d'agent. Mais cette analyse est établie sur la base des noms d'agent prototypiques du français, à savoir les noms d'agent déverbaux en *-eur*, et l'on peut se demander quelle est l'extension de cette typologie. Puisque nous avons montré que les noms d'agent n'étaient pas nécessairement construits, et pas nécessairement des déverbaux en *-eur*, on peut se demander si ces noms sont aussi sensibles à cette tripartition, et si oui, dans quelle mesure cette tripartition est corrélée à des propriétés morphologiques spécifiques.

Il est cependant nécessaire de poser dans un premier temps le fait que la typologie peut s'étendre à d'autres noms que les noms d'agent déverbaux en *-eur*. En effet, nous avons mis en avant dans la section 9.1 le fait que cette distinction est très majoritairement envisagée à l'aune des noms d'agent prototypiques (en français les noms déverbaux en *-eur*). Rien n'est dit sur son application à d'autres types de nom. La question se pose tout particulièrement pour les noms d'agent non déverbaux pour lesquels les conditions évoquées dans la section 9.2 ne s'appliquent pas.

Un premier indice est le fait que les voisins pertinents des barycentres de ces trois classes, dont nous avons vu en section 9.3.2 qu'ils avaient tendance à respecter les conditions de validation des trois classes, ont des profils morphologiques variés, comme des noms dénominaux ou déverbaux suffixés en *-iste* (*trapéziste*), *-ier* (*braconnier*), *-aire* (*manutentionnaire*), *-ard* (*débrouillard*), ou *-ien* (*comédien*), etc. Or, une partie de ces noms sont déjà observés dans le voisinage des barycentres des noms d'agent étudiés dans les Chapitres 7 et 8, se qualifiant de fait pour l'agentivité. Des exemples de ces voisins communs sont donnés en 37.

(37) a. Noms affixés : plombier (G1, S2) [2], arriviste (G62, D15), faussaire (G34, O43)

b. Converts : drogué (G32, S100), criminel (G31, O18)

c. Composés : photographe (G59, S30), ventriloque (G35, S51)

2. Les nombres entre parenthèses indiquent le rang occupé par le nom dans un voisinage donné. Les lettres S,O et D identifient les voisinages des barycentres des noms respectivement statutaires, occasionnels, dispositionnels. La lettre G identifie le barycentre des noms d'agent en *-eur* prototypiques présenté dans le chapitre 7.

d. Noms simples : voyou (G18, D35, O91), bandit (G46, O40), héros (G88, O10)

e. Noms complexes non construits : ivrogne (G39, O58, D10), apprenti (G73, S9)

Dans la mesure où nous sommes en présence de noms d'agent proches de barycentres sensibles à cette distinction sémantique entre statutaire, occasionnel et dispositionnel, et que certains de ces noms valident eux-mêmes les conditions nécessaires à l'identification de ces trois classes, on peut dès lors envisager que certains de ces voisins sont des noms d'agent statutaires, occasionnels et dispositionnels. Puisque ces noms d'agent ne sont pas des noms déverbaux en *-eur*, on peut conclure que cette distinction s'applique aux noms d'agent indépendamment de leur profil morphologique. Des exemples de noms non suffixés en *-eur* candidats à la classification en tant que noms d'agent statutaires, occasionnels et dispositionnels, et tirés des voisinages des barycentres correspondants, sont donnés respectivement en (38a), (38b) et (38c).

(38) a. plombier, garagiste, vétérinaire, électricien, couturier, comptable, laborantin, maréchal-ferrant, chirurgien, assistant, proxénète, taxidermiste, radiologue, surveillant

b. assassin, meurtrier, tortionnaire, traître, criminel, conspirateur, coupable, fraudeur, commanditaire, violeur, oppresseur, gêneur, fugitif, incendiaire

c. filou, fanfaron, vantard, arriviste, goinfre, fainéant, bavard, débrouillard, mythomane, poivrot, débauché, ronchon, goujat, fayot

Si l'on présume que la distinction entre noms d'agent statutaires, occasionnels et dispositionnels s'applique aux noms d'agent non suffixés en *-eur*, on peut se demander si elle est corrélée à des propriétés morphologiques des noms d'agent. L'analyse morphologique des voisins des barycentres des noms d'agent statutaires, occasionnels et dispositionnels au regard de leur construction morphologique, de leur suffixe et de leur type de base met en évidence des tendances significatives, et permet d'ébaucher plusieurs hypothèses concernant la corrélation entre les propriétés morphologiques des noms d'agent et leur sous-classification sémantique.

La construction morphologique des voisins pertinents des barycentres des noms d'agent statutaires, occasionnels et dispositionnels est donnée dans le tableau voir tableau 9.8.

Le tableau 9.8 montre que l'affixation est plus proéminente dans le voisinage du barycentre des noms d'agent statutaires que dans les deux autres

	Affixé	Convert	Composé	Simple	Complexe	Extragram.	Indet.
S	76	6	8	3	6	0	1
O	39	15	1	19	6	0	10
D	22	9	4	13	4	1	26

TABLE 9.8 – Type morphologique des voisins pertinents des barycentres des noms d'agent statutaires (S), occasionnels (O) et dispositionnels (D)

voisinages. En excluant la formation affixale, les noms converts, les noms simples et les noms à la construction indéterminée (c'est-à-dire en relation de conversion mais dont l'orientation est indéterminée, voir section 7.3.1) sont en revanche les plus représentés parmi les autres voisins pertinents. Les noms simples se trouvent principalement dans les voisinages des barycentres des noms d'agent occasionnels et dispositionnels, alors que les constructions morphologiquement indéterminées dominent dans le voisinage des noms d'agent dispositionnels.

Ce point nous semble remarquable dans la mesure où, suivant la ligne d'analyse présentée dans la section 7.3.1, nous annotons comme indéterminés les noms qui sont en relation de conversion avec un autre mot lorsque la direction de la conversion ne peut être identifiée sur des bases morphologiques. Or, des 26 voisins indéterminés dans le voisinage du vecteur moyen des noms d'agent dispositionnels, 24 sont en relation de conversion avec un adjectif (*pervers*) et deux avec un verbe (*escroc*). Des 10 voisins indéterminés parmi le voisinage du barycentre des noms d'agent occasionnels, sept sont en relation de conversion avec un ajectif (*scélérat*) et trois avec un verbe (*assassin*). De fait, si l'on ajoute à cela le fait que les neuf voisins converts dans le voisinage du barycentre des noms d'agent dispositionnels le sont d'adjectifs, il apparaît que le profil morphologique le plus représenté parmi les voisins pertinents du barycentre des noms d'agent dispositionnels est la relation de conversion avec un adjectif, ce qui est une caractéristique distinctive de ce voisinage. Inversement, seuls 10 des 15 voisins converts du voisinage des barycentres des noms d'agent occasionnels sont convertis d'adjectifs, et les cinq autres sont convertis de verbes. Les voisins en relation de conversion avec un adjectif se trouvent donc principalement dans la sous-classe des noms d'agent dispositionnels.

Cela est confirmé par l'analyse de la catégorie grammaticale des bases des voisins dérivés parmi les voisins pertinents des trois barycentres (voir tableau 9.9).

Il y a en effet une dépendance significative entre les différents voisinages et la catégorie grammaticale de la base, comme on peut le voir dans le tableau

	Nom	Verbe	Adjectif
Statutaires	52	29	1
Occasionnels	12	31	11
Dispositionnels	3	19	9

TABLE 9.9 – Catégories grammaticales des bases des voisins dérivés pertinents des barycentres des noms d'agent statutaires, occasionnels et dispositionnels

9.9 (p-value$<$0.00001 au χ^2). Les bases nominales se trouvent majoritairement dans le voisinage du barycentre des noms d'agent statutaires. Les voisins du barycentre des noms d'agent occasionnels sélectionnent quant à eux préférentiellement des bases verbales. Concernant les voisins du barycentre des noms d'agent dispositionnels, si on considère qu'en plus des 31 items dans le tableau 9.9, 24 sont possiblement dérivés d'adjectifs, contre sept dans le cas des voisins du barycentre des noms occasionnels, et 0 dans le cas du barycentre des noms statutaires, alors une corrélation possible apparaît entre les bases adjectivales et le voisinage du barycentre des noms d'agent dispositionnels.

	Action	Objet	Propriété	Domaine	Institution
Statutaires	31	36	2	12	1
Occasionnels	37	5	11	2	0
Dispositionnels	20	1	10	1	0

TABLE 9.10 – Types sémantiques des bases des voisins dérivés pertinents des barycentres des noms d'agent statutaires, occasionnels et dispositionnels

Le tableau 9.10 présente le type sémantique des bases des voisins dérivés parmi les voisins pertinents des trois barycentres. On constate que les voisins des barycentres des noms d'agent occasionnels et dispositionnels sont préférentiellement dérivés de mots qui dénotent des actions ou des propriétés, alors que les bases dénotant des objets ou des domaines se trouvent plutôt parmi les voisins des noms d'agent statutaires. La prédominance des bases actionnelles au sein du voisinage du barycentre des noms d'agent occasionnels peut s'expliquer par la référence à des événements particuliers qui est souvent dépendante de l'existence d'une base morphologique qui dénote des occurrences d'action (*meurtrier*, *sauveur*). La prévalence des actions sur les propriétés est moins évidente dans le cas des noms d'agent dispositionnels, mais ne vaut que dans la mesure où les voisins indéterminés ne sont pas pris en compte. Dans le cas où l'on considère les voisins indéterminés comme des noms désajectivaux, les voisins des noms d'agent dispositionnels seraient

principalement dérivés de bases dénotant des propriétés (*crétin*, *mythomane*). Cette indétermination conforte d'une certaine façon la vision selon laquelle les noms d'agent dispositionnels sont sémantiquement hybrides entre action et propriété. Concernant les voisins des noms d'agent statutaires, ils dénotent généralement des professions ou des spécialistes, lesquels peuvent être pour certains définis par la production ou la manipulation d'objets (*machiniste*, *perruquier*), ou par des activités associées à la connaissance ou la pratique de domaines (*chirurgien*, *cuisinier*) – d'où la sélection de bases dénotant des objets ou des domaines. Dans ces cas-là, la composante sémantique actionnelle n'est pas héritée, mais directement engendrée dans la structure sémantique des noms d'agent dérivés, en relation avec le référent de la base.

Certaines corrélations entre la sélection de l'affixe et le voisinage des noms d'agent statutaires, occasionnels et dispositionnels peuvent être identifiées. Le tableau 9.11 présente la distribution des suffixes utilisés dans les plus proches voisins dérivés des trois barycentres. Il apparaît que les noms en *-iste*, *-ier*, *-ien*, et *-logue* tendent à être des voisins du barycentre des noms d'agent statutaires, alors que les noms en *-aire* forment plutôt des voisins du barycentre des noms d'agent occasionnels, et *-ard* des voisins du barycentre dispositionnel. En considérant que les voisins des barycentres des noms d'agent statutaires, occasionnels et dispositionnels ont de fortes chances d'être respectivement des noms d'agent statutaires, occasionnels et dispositionnels, on peut faire l'hypothèse que les suffixes agentifs s'associent préférentiellement à des sous-types agentifs.

	-eur	*-ier*	*-iste*	*-ien*	*-aire*	*-ard*	*-logue*	*-on*
Statutaires	24	31	11	6	1	0	2	1
Occasionnels	26	5	0	1	4	0	0	0
Dispositionnels	15	0	1	1	0	5	0	0

TABLE 9.11 – Suffixes des voisins dérivés pertinents des barycentres des noms d'agent statutaires, occasionnels et dispositionnels

Nous avons pu montrer dans ce chapitre que la sous-typologie est pertinente sur le plan distributionnel, sous la forme d'une tripartition. Nous avons pu mettre en évidence que cette typologie ne s'appliquait pas qu'aux noms d'agent déverbaux en *-eur*, mais plus largement à l'ensemble des noms d'agent. Nous avons ébauché une première caractérisation morphosémantique des trois sous-types. Ainsi, il ressort de nos analyses que les noms statutaires tendent à être des noms dénominaux, suffixés en *-eur* ou *-ier*, et dont la base dénote un objet ou un domaine, la relation prédicative se construisant sur la

manipulation ou la pratique du référent de la base. La dénotation de propriétés comportementales des noms dispositionnels est liée à leur construction par conversion à partir d'adjectif. Les noms occasionnels tirent quant à leur le potentiel de référence à un événement spécifique de leurs bases actionnelles, principalement verbales.

Plus largement, le travail présenté dans cette partie montre que la sémantique joue un rôle primordial en morphologie. Cela est notamment illustré par l'influence du type sémantique de la base sur le type de noms d'agent formés, avec une association toute particulière des noms de spécialistes et des noms de domaines, des noms d'artisans et des noms d'objet, ou encore des noms ou verbes d'action et les noms occasionnels. Nos résultats soutiennent par ailleurs la relative indépendance de la sémantique vis-à-vis du niveau formel. Nous avons en effet déterminé que le sens – et notamment sur un plan référentiel et ontologique – ne détermine pas totalement la sélection des affixes ou des bases, et réciproquement, que les affixes ne déterminent pas les propriétés sémantiques des dérivés. La dérivation agentive se caractérise par la relation *many-to-many* entre les sous-types lexicaux d'agent et les affixes, quand bien même des relations privilégiées émergent.

Sur un plan méthodologique, nous avons montré qu'il était possible d'explorer des propriétés sémantiques fines à l'aide d'un dispositif distributionnel et des données lexicales contrôlées. Si un contrôle complet sur le dispositif est inenvisageable, notamment lorsqu'il repose sur de grandes quantités de données et des traitements automatiques, cela permet d'aller plus loin dans l'analyse et de creuser des questions toujours plus complexes. Cette étude illustre le rôle que la sémantique distributionnelle peut jouer dans l'analyse linguistique lorsque les méthodes traditionnelles (notamment les tests linguistiques et l'introspection) ne sont plus suffisantes. S'ils peuvent apporter des éléments de réponse, les espaces distributionnels ne sont cependant parfois pas eux-mêmes suffisants, et nécessitent l'utilisation conjointe d'autres outils linguistiques. C'est ce que nous nous attachons à illustrer dans la partie V.

Cinquième partie

Degrés de technicité des noms d'action

À l'image des différences que l'on observait pour les suffixes *-eur*, *-euse* et *-rice* dans la partie III, l'expérience présentée dans la partie II montrait une variation du score proximité entre le verbe et le nom d'action en fonction du suffixe du nom d'action, notamment lorsque l'on considérait les suffixes *-age*, *-ion* et *-ment*. Si ces suffixes sont souvent considérés comme concurrents, des différences ont néanmoins été établies sur la base de critères syntaxiques, formels ou encore sémantiques. Un des critères évoqués, mais à notre connaissance peu exploré, et pas de façon systématique, concerne le domaine ontologique auquel les noms d'action se rattachent et le degré de concrétude de l'action dénotée. Il est ainsi suggéré que le suffixe *-age* serait davantage utilisé dans le domaine industriel (Dubois 1962, Fleischman 1980, Uth 2010), et le suffixe *-ion* dans le domaine scientifique (Dubois 1962). Si l'hypothèse d'une spécialisation ontologique est partagée, elle n'a cependant pas été évaluée de façon empirique. Or, la mise à l'épreuve de cette hypothèse est assez simplement opérationnalisable dans les modèles distributionnels.

Nous nous proposons dans cette partie d'examiner la différenciation sémantique des noms d'action déverbaux en *-age*, *-ion* et *-ment* dans les espaces vectoriels. Plus précisément, nous nous demandons s'il y a effectivement une différence sémantique entre les noms construits par ces trois procédés – est-ce qu'ils forment des classes sémantiques distinctes ? – et si cette différence relève bien de domaines ontologiques liés à l'industrie et aux sciencs. À ce titre, nous reprenons l'approche expérimentale qui consiste à caractériser sémantiquement des classes définies formellement, et appliquons l'utilisation d'une représentation unifiée dans les espaces vectoriels aux noms d'action en *-age*, *-ion* et *-ment*. L'analyse comparative des voisinages permet ainsi de faire ressortir à la fois une grande cohérence sémantique et formelle au sein des classes, et une spécialisation des trois classes que nous décrivons en termes de technicité.

Cette approche nous conduit ensuite à caractériser plus précisément cette spécialisation en termes de technicité. Afin d'appliquer la méthode utilisée pour l'agentivité dans la partie IV, nous avons tout d'abord besoin de délimiter clairement la catégorie formée par les noms techniques. Nous mobilisons d'autres outils d'analyse afin de pallier ce manque. La formulation d'une définition reposant sur des critères quantitatifs opérationnalisables nous permet de discriminer plus précisément les degrés de technicitié exhibés par les noms d'action en *-age*, *-ion* et *-ment*, et que nous étendons plus largement à l'ensemble des noms d'action. Nous montrons sur la base de ces critères que les noms en *-age* se caractérisent bien par un plus fort degré de technicité, contrairement aux noms en *-ion* ou aux noms converts qui se définissent par un degré de technicité plus faible voire nul. La comparaison des noms en fonction de leur degré de technicité dans les espaces vectoriels montre que ces

derniers captent bien, à l'image de la concrétude ou de la sous-spécification, la technicité des noms.

Sur le plan méthodologique, cette partie développe un dispositif expérimental plus complexe que les parties précédentes, intégrant sémantique distributionnelle et approche statistique. Elle illustre la capacité des espaces vectoriels à offrir de nouvelles pistes d'analyse et à les valider en combinaison avec d'autres méthodes.

La partie V est construite comme suit. Nous explorons dans le chapitre 10 les propriétés sémantiques des noms d'action en *-age*, *-ion* et *-ment* sur le plan distributionnel, ce qui nous permet de corroborer l'existence de différences, et que nous caractérisons en termes de technicité. Nous développons dans le chapitre 11 la définition de la technicité des noms d'action, et nous opérationnalisons cette définition à l'aide d'une tâche d'annotation par des locuteurs ainsi que de critères empiriques calculés à partir de ressources lexicales et de corpus, qui nous permettent de valider l'hypothèse d'une plus grande technicité des noms d'action en *-age*. Enfin, nous évaluons dans le chapitre 12 la capacité des espaces vectoriels à capter la technicité des noms d'action en comparant la représentation unifiée de groupes de noms définis par leur degré de technicité.

Le travail présenté dans cette partie a fait l'objet de plusieurs publications (Wauquier *et al.* 2020b, Wauquier *et al.* à paraître). Les données présentées sont accessibles à l'adresse `https://github.com/mwauquier/PhdData/tree/main/Part5`.

Chapitre 10

Différenciation distributionnelle

Dans ce chapitre, nous explorons la différenciation sémantique des noms d'action issus de procédés dérivationnels concurrents, les suffixes *-age*, *-ion* et *-ment*. Nous partons du constat qu'il existe des différences tendancielles, qui peuvent être étudiées à grande échelle et de façon systématique. Pour répondre à cela, nous nous proposons d'appliquer aux noms d'action en *-age*, *-ion* et *-ment* la méthodologie mise en place dans le chapitre 6 pour la comparaison sémantique des noms d'agent en *-euse* et *-rice*, et dans le chapitre 8 pour la comparaison des noms d'agent en *-aire*, *-ant*, *-eur*, *-ien*, *-ier*, et *-iste*, et ce afin de faire émerger de possibles divergences sémantiques dans les représentations vectorielles des trois sous-groupes de noms d'action.

Nous présentons dans un premier temps les arguments en faveur d'une spécialisation sémantique des suffixations en *-age*, *-ion* et *-ment* tels qu'ils ont été identifiés dans la littérature (section 10.1). Dans un second temps, nous présentons l'adaptation de la méthodologie des barycentres développée pour les noms d'agent en *-eur*, *-euse* et *-rice* appliquée aux noms d'action en *-age*, *-ion* et *-ment*, et montrons qu'elle permet effectivement de mettre en évidence une différence entre ces trois suffixations relative à la technicité des noms d'action construits (section 10.2).

10.1 Concurrence des noms d'action en *-age*, *-ion* et *-ment*

Comme évoqué en section 3.1.3, la construction morphologique des noms d'action peut reposer sur une grande variété de procédés dérivationnels. L'existence de ces différents procédés, et la possibilité de construire plusieurs lexèmes distincts à partir d'un même lexème de base, posent néanmoins la question de la spécificité de chaque procédé. La question est particulièrement

étudiée dans le cas des suffixes *-age*, *-ion* et *-ment*, souvent considérés comme concurrents. Nous discutons dans un premier temps la notion de concurrence vis à vis de ces trois suffixes (section 10.1.1), puis nous passons en revue les critères mis en avant pour soutenir l'idée d'une différenciation (section 10.1.1).

10.1.1 *Affrontage*, *affrontation* ou *affrontement* des suffixes

Parmi l'ensemble des procédés affixaux disponibles pour construire des nominalisations déverbales, les trois principaux suffixes utilisés pour former des noms d'action déverbaux sont les suffixes *-age*, *-ion* et *-ment* (Fradin 2014, Missud et Villoing 2020). Ces trois suffixes se caractérisent par une forte similarité de leurs règles morphologiques, puisqu'ils reposent sur le même type d'*input* – traditionnellement un verbe, malgré l'existence de constructions, notamment néologiques, sur base nominale, comme *pipelettage* (Lombard et Huyghe à paraître) – et produisent le même type d'*output*, à savoir un nom. La forte polysémie qui caractérise ces suffixes est globalement partagée par les trois suffixes. Les suffixes *-age*, *-ion* et *-ment* permettent ainsi de construire des noms d'action ou d'événement (*atterissage*, *localisation*), des noms de résultat ou de produit (*pliage*, *ornement*), mais aussi des noms de moyen ou d'instrument (*chauffage*, *ventilation*), de propriété ou d'état (*assemblage*, *abattement*), de trajet (*passage*, *croisement*), de manière (*tissage*, *ciselle-ment*), de lieu (*garage*, *campement*), des noms collectifs (*pavage*, *rédaction*) ou encore des noms de période (*hivernage*, *glaciation*) (Namer 2009, Fradin 2012 2016).

Les suffixes *-age*, *-ion* et *-ment* sont à ce titre souvent étudiés conjointement dans le cadre de leur concurrence. On trouve en effet de nombreux cas de doublets (*ajustage* et *ajustement*, *finissage* et *finition*, *habillage* et *habille-ment*) voire de triplets (*affrontation* et *affrontement* étant attestés dans le TLFi, la forme *affrontage*[1] étant marginalement attestée sur le web) impliquant ces suffixes.

Ces cas de doublets ou de triplets représentent deux réalités. Dans un premier cas, il s'agit de noms construits sur deux lexèmes formellement identiques, mais fondamentalement distincts, à l'image de *élevage* et *élèvement*, construits sur deux lexèmes *élever* distincts (voir Fradin 2014). On peut alors se demander ce qui justifie l'usage de l'un ou l'autre suffixe pour construire

1. « Dans le troisième numéro du Marvel Crossover sorti en juillet 97, on peut donc déplorer l'**affrontage** Silver Surfer/Superman et admirer l'**affrontement** Daredevil/Batman »à l'adresse `https://www.du9.org/chronique/daredevil-batman/`.

ces deux lexèmes distincts.

Dans un second cas, les noms dérivés sons construits à partir d'un même lexème unique, à l'image de *pavage* et *pavement*, construit à partir de *paver*. Il arrive que des différences puissent être observées entre ces noms dérivés, en termes par exemple de contextes d'usage, ou de propriétés aspectuelles ou sémantiques (Dal *et al.* 2018), qui permettent alors d'expliquer la coexistence des deux dérivés, par l'occupation de niches spécifiques (Aronoff et Lindsay 2016, Missud 2019). L'examen de ce type de doublets ou triplets permet ainsi de mettre en avant les différences entre des procédés concurrents. Les traits distinctifs ne sont cependant pas systématiquement exclusifs à l'un ou l'autre procédé, et ne peuvent pas nécessairement être généralisés. Mais il y a aussi des cas où les noms dérivés sont strictement identiques sur le plan sémantique, à l'image des paires *engluage* et *engluement*, *ravalage* et *ravalement*, *triplage* et *triplement*, *désensablage* et *désensablement*, *relogeage* et *relogement*, ou *épinglage* et *épinglement* (Dubois 1962, Dal *et al.* 2018, Fradin 2016). Il ressort de la présence de ces dérivés interchangeables, et, dans le cas précédent, de l'absence de régularité apparente pour les dérivés non interchangeables, l'hypothèse que ces suffixes sont les exposants d'une même règle, au même titre que les suffixes *-ité* et *-té* pour les noms de qualité.

Pourtant, les spécificités de ces trois suffixations ont été longuement étudiées dans la littérature, notamment au regard d'autres procédés comme les suffixes *-age*, *-ion* et *-ment* eux-mêmes, ainsi que la conversion en *-éé* (Ferret *et al.* 2010, Ferret et Villoing 2012), ou la conversion de verbe à nom de façon générale (Missud 2019, Missud et Villoing 2019). Des tendances ont ainsi pu être mises au jour. Nous reprenons dans la section 10.1.2 les principaux critères discutés dans la littérature.

10.1.2 Tendances à la spécialisation

Divers critères ont été proposés pour expliquer la sélection des suffixes *-age*, *-ion* et *-ment*. Une première série de différences concerne les bases sélectionnées par les suffixes.

On retrouve parmi eux la transitivité du verbe. Il est communément admis que le suffixe *-age* préfère les bases transitives, alors que le suffixe *-ment* sélectionnerait plutôt des bases intransitives (Dubois 1962, Dubois et Dubois-Charlier 1999, Fradin 2014). La préférence du suffixe *-ment* est aussi étendue aux verbes réflexifs ou à la voix passive (Kelling 2001, Fradin 2014). Martin (2010) montre qu'il n'y a pas de préférence significative en termes de télicité pour *-age*, *-ion*, *-ment*, et *-erie*, mais soutient néanmoins que *-ion* serait prototypiquement plus télique que *-ment*. Missud (2019) montre quant à elle la préférence du suffixe *-age* (par rapport à la conversion) pour les verbes

d'accomplissement.

Les spécificités des arguments du verbe sont aussi mises en avant pour distinguer les trois suffixes. S'inspirant des protorôles thématiques de Dowty (1991), Kelling (2001) argue que les sujets des verbes sélectionnés par le suffixe *-age* sont prototypiquement plus agentifs que les sujet des verbes sélectionnés par le suffixe *-ment*. S'intéressant à l'objet et non au sujet du verbe, Fradin (2014) note quant à lui que le suffixe *-age* est préféré lorsque l'objet du verbe dénote un référent spécifique, un référent non humain, un objet concret ou un objet inerte. Le suffixe *-ment* est quant à lui préféré pour un référent général, un référent humain, un objet abstrait ou un objet actif.

Des critères formels et morphologiques [2] ont aussi été proposés pour expliquer la concurrence entre les suffixes *-age*, *-ion* et *-ment*. Lapraye (2017) et Missud et Villoing (2020) indiquent que le suffixe *-age* sélectionne préférentiellement les bases courtes et relevant du premier groupe, notamment des verbes converts ou préfixés (en *dé-*, plus marginalement en *é-* ou *en-*). Le suffixe *-ment* sélectionnerait quant à lui plutôt des bases verbales préfixés en *a-* ou *en-*. Le suffixe *-ion* privilégie quant à lui les bases verbales suffixées en *-iser* et *-ifier* [3] (Missud et Villoing 2020) et les bases savantes (Fradin 2014). Fradin (2014) soutient à ce titre qu'il n'y a pas de concurrence entre le suffixe *-ion* d'une part et les suffixes *-age* et *-ment* d'autre part. Seuls les suffixes *-age* et *-ment* seraient réellement rivaux. L'existence de dérivés interchangeables impliquant le suffixe *-ion* et un autre suffixe (*accommodage*, *accommodement* et *accommodation*, ou *cadastration* et *cadastrage*) invite néanmoins à s'interroger sur les ressemblances ou différences sémantiques entre les noms d'action en *-age*, *-ion* et *-ment*. Nous les considèrerons à ce titre dans le cadre de cette étude comme des concurrents.

Des différences relatives aux dérivés construits par les suffixes ont aussi été proposées. Martin (2010) met notamment en avant comme facteurs de différenciation des noms en *-age* et *-ment* la longueur de la chaîne événementielle décrite par les noms, c'est-à-dire la présence d'un plus grand nombre de sous-événements, ainsi que l'incrémentalité de l'action décrite, c'est-à-dire la possiblité de thématiser – au regard des sous-événements de l'action décrite – toutes les parties de l'objet auquel s'applique le procès décrit par le nom. Ainsi, pour une base donnée (*gonfler*, *plisser*), l'auteur soutient que l'événement décrit par le nom en *-ment* (*gonflement*, *plissement*) sera nécessairement inclus dans l'événement décrit par le nom en *-age* (*gonflage*,

2. Nous n'explorons pas ici les critères phonotactiques à l'œuvre.

3. La nature suffixale des chaînes *-isation* et *-ification* est discutée dans la littérature (Lignon *et al.* 2014, Dal et Namer 2015). Nous ne nous positionnons pas sur cette question, et considérons pour le moment les noms en *-isation* et *-ification* comme des noms suffixés en *-ion* à partir de bases verbales en *-iser* et *-ifier*.

plissage), et que le nom en *-ment* statisfera davantage la relation d'incrémentalité que le nom en *-age*. Ferret *et al.* (2010) soulignent par ailleurs la valeur imperfective des noms en *-age*, c'est-à-dire leur capacité à dénoter le procès dans son déroulement et non dans sa globalité.

S'intéressant aux compléments des noms construits, Dubois et Dubois-Charlier (1999) suggèrent que le complément des noms en *-age* est inanimé et qu'il correspond à l'objet du verbe de base (repris par Fradin 2014, alors qu'il sera au contraire animé – voire humain – et correspond au sujet du verbe pour les noms en *-ment*. Les auteurs ajoutent qu'en cas de concurrence pour une base donnée (*apponter*), le nom en *-ment* (*appontemenent*) tend à traduire un résultat, et le nom en *-age* (*appontage*) l'action. Enfin, dans la continuité des considérations de Kelling (2001) sur les sujets des verbes de base, Martin (2010) suggère que les noms construits en *-age* sont plus protoagentifs que les autres, car le suffixe *-age* signalerait la présence d'une action intentionnelle.

Des distinctions ontologiques ont enfin été proposées. Martin (2010) soutient que la suffixation en *-age* invite à une interprétation physique du nom construit (lorsque la base verbale n'implique elle-même pas une interprétation physique), contrairement aux suffixations en *-ion* et *-ment*. Cette idée est corroborée par Fradin (2016) qui indique que le suffixe *-age* privilégie la construction de noms à partir des thèmes concrets. Dubois (1962) va plus loin en indiquant que le suffixe *-age* tend à priviligier la dénotation de processus de domaines spécialisés, notamment d'opérations industrielles, contrairement au suffixe *-ment*. Le suffixe *-ion* serait quant à lui plus fortement associé au lexique scientifique et technique (Dubois 1962). La thèse d'une association particulière du suffixe *-age* au domaine industriel est développée par Fleischman (1980, cité par Uth 2010) dans l'étude diachronique du suffixe *-age*. Fleischman (1980) fait l'hypothèse que l'augmentation du nombre de nominalisations en *-age* au 19e siècle (démontré par Uth 2010) est directement liée à la révolution industrielle et au besoin croissant de désigner de nouvelles technologies et de nouveaux procédés technologiques. Le français aurait emprunté la terminologie technique à l'anglais, qui utilisait massivement le suffixe *-age* lui-même emprunté précédemment au français. Bien que cette hypothèse d'emprunts successifs reste à prouver, il ressort de ces études une association forte entre le suffixe *-age* et le domaine industriel. Dal *et al.* (2018) suggèrent cependant que les suffixes *-age* et *-ment* ne se distinguent pas relativement à l'appartenance à un domaine de spécialité.

En tout état de cause, les différents critères précédemment présentés reposent principalement sur l'étude individuelle des noms, qui ne permet pas d'avoir une vision extensive du comportement des noms construits par ces procédés. Ces études se construisent sur des données parfois limitées, et se

basent assez peu sur l'étude systématique des similarités ou des différences entre les procédés. La stabilité de certaines observations est ainsi remise en cause par l'analyse de nouvelles données. L'étude des noms d'action en *-age* à partir de formes néologiques (c'est-à-dire absentes des ressources dictionnairiques) contredit par exemple l'hypothèse d'une prédilection marquée pour les verbes transitifs, et une préférence aspectuelle pour les activités (Lombard et Huyghe à paraître). L'examen en contexte de doublets en *-age* et *-ment* relevés sur la Toile contredit par ailleurs l'existence de tendances nettes spécifiques aux suffixations (Dal *et al.* 2018).

Si des tendances à la spécialisation des suffixes *-age*, *-ion* et *-ment* apparaissent aux travers de ces études, il émerge surtout la nécessité d'employer une approche plus systématique, permettant l'analyse plus extensive des différences entre ces trois suffixes. C'est ce que nous proposons de faire, en explorant les différences sémantiques des noms construits en *-age*, *-ion* et *-ment* à l'aide de la sémantique distributionnelle. Plus précisément, nous souhaitons explorer l'hypothèse de spécificités sémantiques et ontologiques des noms construits liées au caractère concret, physiques et spécialisés des actions dénotées.

10.2 Comparaison distributionnelle des noms d'action en *-age*, *-ion* et *-ment*

Sur la base des observations précédemment présentées, nous faisons l'hypothèse que les noms d'action en *-age*, *-ion* et *-ment* diffèrent relativement à leur domaine ontologique et à leur degré de concrétude et de spécialité. Plus précisément, nous cherchons à vérifier que les noms en *-age* relèvent de domaines techniques ou concrets comme l'industrie, les noms en *-ion* des sciences, et que les noms en *-ment* ne sont pas spécialement marqués. Pour tester cela, nous adoptons une approche contrastive similaire à celle développée pour les suffixes agentifs *-eur*, *-euse* et *-rice*. En partant du principe que les noms construits par chaque suffixe forment un sous-groupe sémantique distinct, nous construisons la représentation vectorielle de chaque groupe, que nous comparons ensuite.

10.2.1 Construction des barycentres des noms d'action

Dans la lignée de ce qui a été proposé pour les suffixes *-eur*, *-euse* et *-rice* dans la partie III, et pour les suffixes *-ant*, *-aire*, *-eur*, *-ien*, *-ier* et *-iste* dans le chapitre 8, nous construisons un barycentre par suffixe, suivant la formule

5.1. La comparaison des barycentres des noms d'action en *-age*, *-ion* et *-ment* passe par l'analyse de leurs 100 plus proches voisins.

Pour les raisons évoquées en section 3.2.2, nous utilisons le corpus *Wikipedia2018* pour construire le modèle distributionnel sur lequel nous basons nos analyse. Nous faisons ici le choix d'utiliser le corpus dans sa forme lemmatisée, mais ne jugeons pas nécessaire de l'étiqueter morphosyntaxiquement, comme cela avait été fait pour la partie IV. En effet, l'ambiguïté liée aux suffixes *-age*, *-ion* et *-ment* relève de la polysémie sémantique mais pas syntaxique. La seule exception concerne le suffixe *-ment*, qui forme aussi des adverbes à partir d'adjectifs, mais cela ne se traduit pas par des ambiguïtés formelles entre noms d'action et adverbes comme cela était le cas entre nom d'agent et adjectif avec *enchanteur* par exemple. L'information catégorielle n'est donc pas ici utile, et l'utilisation d'un corpus uniquement lemmatisé limite donc le bruit et allège le modèle final.

Afin de limiter les effets de l'instabilité intrinsèque des modèles distributionnels, nous utilisons un modèle concaténé, tel que présenté en section 5.3.2. Nous avons montré que l'approche par concaténation et l'approche par voisins communs produisaient les mêmes résultats. Si nous avons choisi l'approche par voisins communs dans les parties III et IV, nous favorisons ici l'approche par concaténation CONCAT présentée dans la section 5.3.2, qui permet un traitement allégé puisque cela nous permet de travailler à partir d'un unique modèle – bien que concaténé à partir de cinq modèles distincts – sans étape intermédiaire. Les cinq modèles sont entraînés selon les mêmes paramètres que ceux utilisés dans le reste de ce travail, à savoir l'architecture CBOW, l'algorithme *Negative Sampling*, une fréquence minimum de 5, une fenêtre de 5, et 100 dimensions.

Pour construire les barycentres des noms d'action en *-age*, *-ion* et *-ment*, nous devons sélectionner les membres de chaque classe. Dans la partie II, nous avions analysé les noms d'action en extrayant l'ensemble des noms relevant du champ 'Nominalisation' de Lexeur. Or, nous avons vu que les noms présents dans ce champs ne sont pas tous des noms d'action (*fumerie*, *machinerie*). Nous procédons à une première étape de nettoyage des données afin de contrôler au mieux l'expérience et de garantir la délimitation de la catégorie des noms d'action. Nous opérons pour cela un filtrage manuel dont l'objectif est d'éliminer les couples pour lesquels le lien entre le nom et l'action dénotée par le verbe est nul ou très douteux. L'annotation est réalisée par trois annotateurs avec adjudication. Nous cherchons à exclure les noms (i) dont le lien avec la base verbale renseignée dans Lexeur est faible (*reportage* par rapport à *reporter*, *chantage* par rapport à *chanter*) ou inexistant (*pleurage* par rapport à *pleurer*) sur des bases morphologiques et sémantiques ; (ii) qui n'ont pas de sens processif (*diction*, de *dire*).

Nous extrayons pour cela les noms contenus dans le champ 'Nominalisation' des familles déverbales de Lexeur construits en *-age*, *-ion* et *-ment*, pour une total de 2 695 noms dont 1 687 en *-age*, 1 357 en *-ion* et 1 222 en *-ment*. Du fait des conditions expérimentales liées aux modèles distributionnels, nous ne conservons que les paires Nom-Verbe [4] pour lesquelles les deux lexèmes ont une fréquence supérieure ou égale à 5 dans le corpus *Wikipedia2018*, et ce afin de garantir l'homogénéité des résultats à travers ce travail.

Sur la base des paires conservées, nous excluons les paires dont l'une des formes au moins n'est plus en usage en synchronie (à l'image de *aberrer* par rapport à *aberration*) ou uniquement dans des expressions figées (à l'image de *empoigne* par rapport à *empoigner* dans *foire d'empoigne*, ou *entremise* par rapport à *entremettre* dans *par l'entremise de*). Nous décidons aussi de limiter dans une certaine mesure l'impact de la polysémie, à l'image de *garage* qui a un sens processif et un sens locatif. Pour cela, nous excluons les noms pour lesquels nous considérons que le sens processif est moins fréquent que le sens non processif – comme c'est le cas dans le cas de *garage*. Plus largement, lorsque le sens jugé prédominant de la base et celui du dérivé ne coincident pas, nous excluons la paire (à l'image de *testament* et *tester*, ou *pelage* et *peler*). Lorsqu'il y a un désaccord net entre les annotateurs, nous faisons le choix d'exclure la forme. Nous conservons les formes qui ne sont pas des noms d'action en tant que tels, mais qui entrent dans des constructions à verbe support comme *faire du N* jugées sémantiquement équivalentes à l'action dénotée par la base verbale, à l'image de *patin* vis à vis de *patiner*. Enfin, nous ne conservons que les noms suffixés en *-age*, *-ion* et *-ment* [5].

4. Nous effectuons ce filtrage sur la base des paires Nom-Verbe et non pas sur le lexème nominal uniquement afin de garantir le lien sémantique et morphologique avec la base.

5. Sont considérés comme des noms déverbaux en *-age*, *-ion* et *-ment* les noms issus de familles déverbales dans Lexeur et n'étant pas en relation de conversion supposée avec ce verbe. Sont considérés comme noms converts tout nom d'action présent dans les listes de converts de Tribout (2010), indépendamment de sa forme, que l'orientation Verbe-Nom soit clairement établie ou non (*ascension*). Ne sont pas considérés comme converts les noms en relation de conversion avec un verbe pour lesquels l'orientation Nom-Verbe est établie par Tribout (2010) (*promotion*). Des noms comme *naufrage*, *confection*, *tourment* – pour lesquels ils existent dans Lexeur les verbes *naufrager*, *confectionner* et *tourmenter* – sont ainsi analysés comme des converts malgré la présence de la chaîne de caractères finale *-age*, *-ion* ou *-ment*, du fait de la relation de conversion identifiée par Tribout (2010). *A contrario* des noms comme *vision* ne sont pas considérés comme des converts, puisqu'ils ne sont pas considérés comme des converts de verbe ni par (Tribout 2010) (mais en relation de conversion Nom-Verbe avec le verbe *visionner*) ni dans la ressource Lexeur (base verbale *voir*). Les noms annotés comme *-age*, *-ion* ou *-ment* sont de fait absents des listes de converts. Sont annotés comme 'autre' les noms absents des listes de converts, et non porteurs des chaînes *-age*, *-ion* et *-ment*.

Nous nous retrouvons à l'issue de ce filtrage avec 1 828 noms, dont 629 noms en *-age*, 750 en *-ion* et 449 en *-ment*. Nous nous servons de ces 1 828 noms pour construire les barycentres des noms d'action en -age, -ion et -ment, dont nous analysons les 100 plus proches voisins en section 10.2.2.

10.2.2 Comparaison des voisinages distributionnels

Nous analysons les 100 premiers voisins de chaque suffixe. Un aperçu des 10 premiers voisins de chaque barycentre est donné dans le tableau 10.1.

-age	*-ion*	*-ment*
usinage	généralisation	déplacement
polissage	manipulation	étirement
meulage	dégradation	durcissement
piquage	simplification	ajustement
perçage	stimulation	relâchement
sablage	contamination	traitement
pliage	dispersion	adoucissement
remplissage	dénaturation	utilisation
salage	transformation	échauffement
soufflage	récupération	enfoncement

TABLE 10.1 – Dix plus proches voisins des barycentres des noms d'action en *-age*, *-ion* et *-ment*

Le tableau 10.1 illustre un premier fait remarquable, à savoir la grande cohérence formelle des voisins, qui se confirme à l'échelle des 100 plus proches voisins.

En effet, on constate une forte homogénéité morphologique des voisins : 82% des voisins du barycentre des noms d'action en *-age* sont eux-même des noms d'action en *-age*. De même, 80% des voisins du barycentre des noms en *-ion* sont eux-mêmes suffixés en *-ion*, et 73% pour *-ment*. À titre de comparaison, le taux était de 46% pour *-eur*, 27% pour *-euse*, et 17% pour *-rice*. Ces dérivés ne se mélangent pas. Les dérivés en *-age*, *-ion* et *-ment* occupent des zones distinctes de l'espace vectoriel, ce qui suggère donc l'existence d'une distinction claire entre les trois groupes formés par ces dérivés. Cette homogénéité n'est pas le fruit d'un recouvrement amorce/voisins, puisqu'on ne retrouve respectivement que 53%, 45% et 37% d'amorces dans les voisinages des barycentres correspondants. Cela confirme que les barycentres se trouvent donc localisés dans une zone occupée par les noms construits par ces suffixes, au-delà des noms utilisés pour construire les barycentres.

Le détail de la distribution des voisins des barycentres en fonction de leur construction morphologique est donné dans le tableau 10.2.

	-age	*-ion*	*-ment*	convert	autre
Barycentre des noms en *-age*	82	4	8	3	4
Barycentre des noms en *-ion*	0	80	1	6	13
Barycentre des noms en *-ment*	11	7	73	6	3

TABLE 10.2 – Constructions morphologiques des 100 plus proches voisins des barycentres des noms d'action en *-age*, *-ion* et *-ment*

Outre la forte connivence morphologique entre les barycentres et leurs voisins respectifs, le tableau 10.2 montre certaines tendances relatives aux différents suffixes. Ainsi, on constate que l'autre construction morphologique la plus représentée dans le voisinage du barycentre des noms d'action en *-age* est la suffixation en *-ment*, et réciproquement que la suffixation en *-age* est aussi la seconde construction morphologique la plus représentée dans le voisinage du barycentre des noms d'action en *-ment*. *A contrario*, ces suffixes sont virtuellement absents du voisinage du barycentre en *-ion*, qui présente une plus forte concentration de noms convertis et de noms construits avec d'autres procédés. Cela suggère un recouvrement des zones occupées par les suffixes *-age*, *-ion* et *-ment* légèrement plus important entre *-age* et *-ment* qu'avec *-ion*. Cela semble corroboré par le nombre de voisins partagés par les trois barycentres. En effet, on note que les barycentres des noms en *-age* et *-ion* partagent deux voisins, et les barycentres des noms en *-ion* et *-ment* six voisins, quand les barycentres des noms en *-age* et *-ment* en partagent huit. Ce recouvrement reste néanmoins modeste, surtout au regard du recouvrement que nous avons observé en section 8.3.1 pour les noms d'agent concurrents en *-ant*, *-aire*, *-eur*, *-ien*, *-ier* et *-iste*.

Cela va donc dans le sens d'une identité assez forte pour chaque suffixe dans l'espace distributionnel, traduisant des différences sémantiques entre les noms d'action suffixés en *-age*, *-ion* et *-ment*. Ces différences restent cependant à caractériser.

L'analyse sur le plan sémantique et référentiel des 100 plus proches voisins des barycentres des noms d'action en *-age*, *-ion* et *-ment* montre tout d'abord que les trois listes de voisins comportent quasi exclusivement des noms d'action (ou possédant au moins une acception processive). Les seules exceptions sont les noms *trépan* dans le voisinage du barycentre des noms en *-age* (nom d'instrument), *stabilité* et *complexité* dans le voisinage du barycentre des noms en *-ion* (noms de qualité) et *inconfort* dans le voisinage du barycentre des noms en *-ment* (nom de qualité).

Sur un plan ontologique, on constate que les voisins du barycentre des noms en *-ion* correspondent largement à des noms dénotant des procédés ou phénomènes relatifs aux sciences, qu'il s'agisse de médecine ou biologie (*coagulation*, *succion*, *cicatrisation*, *mastectomie*) ou de chimie (*dilution*, *dénaturation*, *cristallisation*, *polymérisation*) – entre autres – ce qui va dans le sens de la description proposée par Dubois (1962). On retrouve aussi des noms relatifs à des processus psychologiques, comme *compréhension*, *détermination*, ou *perception*, et, en plus grande quantité, des noms relativement généraux voire sous-spécifiés comme *modification*, *action* ou *utilisation*, caractérisés par une importante polysémie. Certains noms sous-spécifiés se retrouvent ainsi dans le lexique scientifique transdisciplinaire (Hatier 2016), comme *application*, *pratique* ou *synthèse*, ce qui accentue le penchant scientifique du barycentre des noms d'action en *-ion*.

Le voisinage du barycentre des noms en *-age* diffère fortement de celui du barycentre des noms en *-ion*. On y trouve un grand nombre de noms relatifs à des techniques ou procédés industriels, tels que *soudage*, *usinage*, ou *brasage*, et peu voire aucun nom général ou sous-spécifié. Les plus généraux comme *stockage*, *traitement*, *nettoyage* ou *lavage* dénotent des actions qui semblent intrinsèquement plus techniques que celles des noms sous-spécifiés trouvés dans le voisinage du barycentre des noms en *-ion*. Différents types de noms se distinguent parmi les noms dénotant des techniques ou actions industrielles : *séchage*, *rinçage*, *lavage* diffèrent de *meulage*, *extrusion*, *chromage*. Les unes semblent renvoyer à des actions plus ordinaires que d'autres, qui semblent plus techniques. On note l'absence de noms relatifs aux sciences, ainsi que celle des processus cognitifs que l'on observait précédemment. Ces observations vont là encore dans le sens des observations précédemment faites dans la littérature et évoquées en section 10.1.2 (Dubois 1962, Fradin 2014).

Le voisinage du barycentre des noms en *-ment* semble quant à lui plus hétéroclite. On y trouve à la fois des noms généraux voire sous-spécifiés comme *déplacement* ou *traitement*, mais aussi des noms plus spécifiques, pour certains relevant de techniques particulières, comme *relèvement* ou *ensablement*. On note l'absence à la fois de noms relatifs aux sciences et de noms relatifs à l'industrie. Ce voisinage se distingue à ce titre des deux voisinages précédemment étudiés, et corrobore l'hypothèse d'une non spécialisation – du moins ontologique – du suffixe *-ment*.

Ces premières observations permettent de dessiner deux profils sémantiques très distincts pour les noms d'action suffixés en *-age* et *-ion* qui sont en accord avec les observations précédemment obtenues à l'aide d'autres approches. Nous faisons notamment émerger de la comparaison des barycentres construits une différence relative au domaine ontologique – industrie pour les noms en *-age*, sciences pour les noms en *-ion*. Plus précisément, les voisins

des barycentres semblent se distinguer en termes de technicité. Le voisinage du barycentre des noms en *-age* semble dénoter des actions plus techniques que les autres voisinages.

Notre approche permet de confirmer sur un plan expérimental ces observations, en les faisant émerger des usages, et à grande échelle. La caractérisation des profils sémantiques ainsi ébauchés reste cependant à ce stade intuitive. Nous souhaitons dans la suite de ce travail caractériser de façon plus objective les spécificités sémantiques de chaque suffixe. L'exploration des espaces vectoriels ne nous permet cependant pas d'approfondir cette spécialisation sémantique des noms d'action en *-age* et *-ion*, au risque d'une circularité de la caractérisation. Pour approfondir la description que nous ébauchons, nous devons recourir à d'autres outils linguistiques. Nous étayons dans le chapitre 11 le constat d'une technicité variable des noms d'action en nous dotant d'une définition et de critères permettant de quantifier la technicité des noms d'action.

Chapitre 11

Technicité des noms d'action

Nous avons fait dans le chapitre 10 le constat, qui restait impressionniste, d'une différenciation distributionnelle des noms d'action en *-age*, *-ion* et *-ment* en termes de technicité. Les voisins du barycentre des noms d'action en *-age* semblaient plus techniques que ceux du barycentre des noms d'action en *-ion*. Par ailleurs, nous avons observé une grande cohérence entre forme et sens puisque les voisinages étaient principalement occupés par des noms du même procédé. Cela tendait à confirmer l'hypothèse que la suffixation en *-age* tend à former des noms d'action plus techniques que la suffixation en *-ion* ou en *-ment*.

La vérification de cette hypothèse se heurte cependant à l'absence d'une définition lexicale de la technicité, hors de toute considération terminologique. L'objectif de cette section est donc de proposer une définition et une quantification de la technicité des noms d'action. Nous nous attachons dans un premier temps à définir la technicité sur le plan lexical (section 11.1). Nous explorons ensuite plus précisément l'hypothèse de la technicité des noms d'action telle qu'elle peut être perçue par les locuteurs en développant une tâche d'annotation, qui nous permet de dresser un premier bilan de la technicité des noms d'action en *-age*, *-ion* et *-ment* (section 11.2). Enfin, nous opérationnalisons à plus grande échelle la définition de la technicité des noms d'action, et modélisons statistiquement la différence de technicité des noms d'action en fonction de leur procédé dérivationnel.

Si le constat d'une différence de technicité a émergé sur la base des noms d'action en *-age*, *-ion* et *-ment*, nous nous posons la question subsidiaire de l'extension de cette distinction aux autres noms d'action. Nous faisons donc le choix dans la suite de cette étude d'étendre notre analyse à l'ensemble des noms d'action.

11.1 Définition de la technicité

L'étude de la technicité des noms d'action nécessite une définition claire de la notion de technicité. Or, cette notion reste à notre connaissance très peu étudiée en tant que telle, tant en terminologie qu'en sémantique. Nous nous proposons dans cette section de montrer les limites actuelles de la définition linguistique et lexicale de la technicité (section 11.1.1). Puis nous explorons la notion de technicité par le biais des travaux de Simondon en philosophie (section 11.1.2), afin d'aboutir à une proposition de définition de la technicité (section 11.1.3).

11.1.1 Flou linguistique

La notion de technicité est relativement peu définie en linguistique (Mudraya 2006). On parle de corpus technique, de vocabulaire technique, mais la notion elle-même de technicité est assez peu étayée. Divers travaux évoquent la notion de corpus ou de langue technique, mais ces notions ne sont jamais précisément décrites en tant que telles.

Cette absence de consensus explicite se traduit ainsi par une très grande diversité de corpus dits techniques. Par exemple, certains travaux basent leur analyse du domaine technique sur des corpus relevant d'un domaine spécifique, comme le domaine nautique (Baroni et Bisi 2004), celui du pétrole et des sciences et technologies (Mudraya 2006), des technologies nucléaires et de la médecine (Habert *et al.* 1996), des télécommunications (Drouin 2003), ou de l'informatique (Wang *et al.* 2009). Dans ces études, la notion de technicité suppose de circonscrire préalablement un domaine de spécialité, dans une perspective terminologique. *A contrario*, d'autres études font le choix de travailler sur des corpus intégrant un nombre varié de domaines et de genre, à l'image de l'encyclopédie en ligne *Wikipedia* (Nazar *et al.* 2012). Des auteurs comme Siddiqui *et al.* (2016) considèrent de la même façon comme techniques des documents aussi divers que des brevets, des documents juridiques, des accords immobiliers, des archives historiques et des articles scientifiques (entre autres).

Un ersatz de définition émerge par le biais de l'opposition entre technique et général. On retrouve cette opposition au niveau des corpus, mais aussi au niveau de la langue. Cette opposition se fait principalement au niveau du lexique. Mudraya signale ainsi que « the division between technical and non technical vocabulary is far from distinct »(Mudraya 2006 : 238). Wang *et al.* (2009) évoquent des « technical terms », par opposition à des termes qui ne seraient pas techniques. Chez Drouin (2003) et Fuentes (2001), cette opposition se traduit notamment par la caractérisation des corpus utilisés, où l'on

retrouve les expressions « non technical corpora »et « academic and technical corpora ». L'adjectif *technique* y est à chaque fois utilisé sans qu'une réelle définition ne soit proposée. Notons qu'une seconde opposition, tout aussi peu définie, émerge entre technique et scientifique. Ces deux notions sont ainsi mises en regard dans (Nazar *et al.* 2012) avec l'expression « scientific and technical corpora »ou « technical or specialized meanings »dans (Fuentes 2001 : 111). Cette distinction est par contre absente chez d'autres auteurs, tels que Siddiqui *et al.* (2016), qui intègrent des données issues à la fois des domaines technique (au travers des brevets) et scientifique (au travers des articles scientifiques).

Une autre difficulté quant à la définition de la technicité émerge de son apparente proximité avec les notions de terme technique et de spécialisation (définies relativement à l'émetteur et au destinataire de l'énoncé analysé, voir Josselin-Leray 2005), issues de la terminologie. Dans ce contexte, ces notions sont très dépendantes du domaine et du corpus. Or, l'approche que nous avons choisie est lexicale, pas terminologique. Notre étude ne se limite pas à un domaine de spécialité, comme le montre notre choix de corpus (une encyclopédie) qui n'est donc pas un corpus spécialisé en tant que tel puisqu'il ne s'adresse pas à un public spécifique du domaine, mais au contraire au lecteur tout venant. De fait, nous distinguons ici l'idée d'appartenance et d'usage telle que mise en avant par Mudraya (2006) ou Fuentes (2001) : on parlera de vocabulaire technique lorsqu'un mot appartient à un sous-langage (ou « technolecte »chez Boulanger 1996) et d'usage technique pour parler d'un emploi terminologique dans un contexte spécialisé. Cela distinguera à ce titre des noms comme *thermoformage*, qui dénote une action intrinsèquement technique, indépendamment du domaine de spécialité dont elle relève, et *rasage*, dont la technicité de l'action dépend entièrement du domaine dans lequel elle s'inscrit. Le nom *rasage* désigne tantôt une technique d'épilation relativement peu technique, tantôt une opération d'usinage de métaux très technique.

11.1.2 Définition philosophique

La technicité est étudiée par Simondon (1958) dans une approche plus philosophique, sur laquelle nous construisons par la suite notre définition de la technicité.

Simondon (1958) décrit la technicité au travers de plusieurs aspects. Le premier d'entre eux est l'agentivité. Il envisage en effet la technicité comme étroitement liée à l'homme : « [l'homme] est parmi les machines qui opèrent avec lui »(Simondon 1958 : 12). La technicité implique donc des agents, mais aussi des instruments. C'est ainsi à l'aune des objets – qu'il qualifie de tech-

niques – que le philosophe définit la technicité : « la technicité se manifest[e] par l'emploi d'objets techniques »(Simondon 1958 : 156). La technicité y est envisagée comme un phénomène à degrés variables, à l'image de la technicité des objets, dépendante de leur niveau de perfectionnement et de complexité. La complexité est elle-même estimée selon le caractère inné ou acquis de la connaissance nécessaire à son utilisation (Simondon 1958). Un savoir inné, non réfléchi, lié à un objet d'usage ou de la vie quotidienne, traduira une action moins technique qu'une action qui est le fruit d'une "opération réfléchie, d'une connaissance rationnelle élaborée par les sciences" (Simondon 1958 : 85). Plus l'opération dénotée par le nom d'action nécessite une connaissance construite et acquise, plus elle présentera un haut degré de technicité. D'un point de vue ontologique, Simondon indique que « la technique touche au commerce, à l'agriculture, à l'industrie »(Simondon 1958 : 97).

11.1.3 Proposition de définition

Dans la lignée de Simondon (1958), et pour la suite de ce travail, nous considérons qu'un nom d'action est technique lorsqu'il dénote une action relevant d'un domaine technique. Nous entendons par domaine technique l'industrie et l'agriculture, tels que proposés par Simondon (1958) et Dubois (1962), domaines auxquels nous ajoutons l'artisanat : si Simondon l'excluait du fait d'une moindre technicité, il nous semble qu'il s'y rattache néanmoins du fait notamment des connaissances et des outils nécessaires pour la réalisation des actions liées au domaine. Plus que l'appartenance à ces domaines précis, nous retiendrons qu'un nom technique est spécifique à un domaine particulier, et qu'il ne sera donc pas utilisé dans d'autres domaines ou dans un contexte plus général (Kocourek 1982). Du fait de sa spécificité, le nom technique est peu transparent pour un public non initié, et nécessite des connaissances particulières pour comprendre l'action dénotée par le nom (Kocourek 1982). Ces différents éléments permettent de définir la technicité des noms d'action de la façon suivante.

Nom d'action technique Nom peu transparent pour un public non initié, dénotant une action précise complexe, dont la réalisation et la connaissance nécessitent un savoir acquis et qui est spécifique à un domaine particulier. Les noms d'action techniques appartiennent aux domaines de l'industrie, de l'agriculture et de l'artisanat.

De fait, nous considérons comme technique tout nom qui dénote une action réalisée intentionnellement par un agent et dont la complexité nécessite de la part de l'agent un savoir ou une connaissance acquise qui peut (mais pas nécessairement) être au cœur d'une tâche ou d'un métier bien défini.

Les actions techniques ne peuvent être réalisées avec succès par des agents tout-venants dans la vie de tous les jours, dans la mesure où elles impliquent en plus de connaissances spécifiques un dispositif ou des conditions de réalisation spécifiques (un environnement, des outils, un contexte). Cela oppose des noms d'action comme *danse* et *ébreuillage*. N'importe qui peut danser n'importe quand, n'importe où, sans être un danseur professionnel, alors que l'ébreuillage requiert des compétences particulières et un outillage spécifique afin d'être réalisé correctement. La complexité et la spécificité de l'action se traduisent par le caractère non familier du nom pour des noms experts. Selon notre approche, des domaines autres que l'industrie et l'agriculture, comme les disciplines scientifiques, peuvent aussi fournir des noms techniques dénotant des actions complexes impliquant des outils ou machines, ainsi que des connaissances particulières (à l'image de *inoculation* ou *cassation*). Cependant, l'appartenance à un domaine spécialisé n'est pas une condition nécessaire ou suffisante pour qu'un nom soit considéré comme technique. Ainsi, un nom comme *aimance* relève bien d'un domaine spécialisé – la psychologie – mais il ne répond cependant pas à notre définition d'un nom technique dans la mesure où il n'implique pas la réalisation d'une action complexe par un agent possédant des compétences acquises spécifiques à l'aide d'un outillage dédié.

Soulignons que si notre définition suggère qu'il existe une classe des noms d'action qui serait clairement délimitée (à l'image de celle des noms d'agent), nous considérons la technicité comme une propriété gradable, qui s'instancie à des degrés variables. Ainsi, on peut considérer la technicité sous la forme d'un continuum, avec d'une part des noms très techniques, et de l'autre des noms non techniques. Parmi les noms très techniques, on retrouve des noms comme *puddlage*, qui est obscur pour les non experts, et dont l'action dénotée relève du domaine de la métallurgie, et uniquement de ce domaine, nécessite un savoir-faire ainsi qu'un dispositif spécifique. À l'opposé de ce continuum, on retrouve des noms sous-spécifiés (ou noms capsules, selon la terminologie de Huyghe (2018)) – c'est-à-dire des noms généraux et abstraits, au sémantisme vague (Halliday et Hasan 1976), et dont la spécification sémantique est relative au contenu propositionnel qu'ils encapsulent. Ces noms, à l'image de *analyse* et *explication* sont donc par nature l'exact opposé des noms d'action techniques. Entre les deux, on placera des noms décrivant des actions peu techniques (*clouage*), moyennement techniques (*toilettage*), et fortement techniques (*alunissage*), dont le niveau de technicité varie relativement aux connaissances nécessaires à la réalisation de l'action dénotée, à la complexité du dispositif nécessaire, ou encore à la familiarité de l'action.

La définition ci-dessus permet de déduire trois propriétés linguistiques (ci-après désignées T1 à T3) de la technicité des noms d'action. La première

propriété que nous dégageons est la **spécialisation** (T1) : un nom d'action technique est plus souvent utilisé dans des contextes spécialisés que dans des contextes généraux puisqu'il dépend d'un domaine particulier et que l'action qu'il dénote est spécifique. Dans la lignée de Simondon (1958) qui considère que la technicité est une notion plus précise que la spécialisation, nous considérons la spécialisation comme propriété mais pas essence de la technicité. La spécificité dénotationnelle donne lieu à une seconde propriété, à savoir l'**obscurité** (T2) : un nom d'action technique est plus souvent décrit et expliqué aux non experts qu'un nom non technique puisqu'il dénote une réalité qui n'est pas familière aux non experts. La description peut être une définition dictionnairique, un article dans une encyclopédie, etc. À l'aune de la définition de la technicité, un nom d'action non technique est réciproquement défini comme un nom transparent pour un public de non spécialistes, et dont l'action dénotée n'est ni spécifique, ni particulière à un domaine de spécialité. La troisième propriété que nous dérivons de notre définition est l'**univocité** (T3) : un nom d'action technique tend à être plus univoque qu'un nom non technique (avec comme exemple le plus extrême les noms génériques), du fait de la spécificité de l'action dénotée. La monosémie peut être utilisée comme approximation de l'univocité. *A contrario*, l'équivocité peut être approchée par le biais de la polysémie et de la sous-spécification.

11.2 Technicité perçue des noms d'action

Sur la base de la définition que nous proposons en 11.1.3, nous testons l'hypothèse d'une plus grande technicité des noms d'action en *-age* par rapport aux noms d'action en *-ion* et *-ment*. Plus précisément, nous vérifions que cette distinction telle que nous la décrivons est bien perçue par les locuteurs. Pour cela nous mettons en place une tâche d'annotation que nous présentons en section 11.2.1), avant de présenter les résultats de cette annotation (section 11.2.2). Cette tâche constitue une première étape préliminaire, consistant à ce stade à tester le guide préalablement à une annotation à plus large échelle. L'annotation de la technicité perçue s'avère une tâche complexe, notamment du fait de la difficulté à circonscrire cette notion, avec une forte variation inter-individuelle puisque dépendante de l'expérience des locuteurs. Nous présentons dans ce qui suit la tâche (section 11.2.1) ainsi que les résultats de cette première phase (section 11.2.2).

11.2.1 Annotation de la technicité perçue

Pour évaluer la technicité des noms d'action sur la base de notre définition, nous avons mis en place une tâche d'annotation. Comme nous souhaitons considérer les noms d'action plus largement que *-age*, *-ion* et *-ment*, nous commençons dans un premier temps par réintégrer à nos données les noms d'action de Lexeur précédemment exclus du fait de leur procédé dérivationnel, soit entre autres les noms en *-ure* (*soudure*), *-erie* (*tromperie*), *-ance* (*jouissance*), *-aison* (*flottaison*), etc. Comme pour les noms en *-age*, *-ion* et *-ment*, nous procédons au filtrage des autres noms d'action selon les mêmes critères que ceux explicités en section 10.2.1. Nous récupérons à l'issue de ce filtrage 325 noms converts [1] et 290 noms construits par d'autres procédés, que nous réintégrons aux 1 828 noms initiaux, pour un total de 2 443 noms.

Du fait de la lourdeur de la tâche, l'annotation a porté sur 287 noms d'action sélectionnés aléatoirement parmi l'ensemble des 2 443 noms filtrés, comprenant 47 noms en *-age*, 141 noms en *-ion*, 47 noms en *-ment*, 21 noms converts et 31 noms construits par d'autres procédés dérivationnels. Sur la base d'un guide d'annotation (voir Annexe A), il était demandé aux annotateurs d'évaluer la technicité des noms d'action sur une échelle de 0 (pas technique) à 5 (très technique). Le guide donne une définition de la technicité, puis explicite certains aspects qui facilitent le diagnostic de technicité, comme l'existence d'un agent ou d'un instrument en lien avec l'action, le degré de familiarité du terme, le niveau de complexité de l'action associée. Des exemples illustrent la tâche au regard des critères explicités par le guide.

Cette tâche a été réalisée par trois annotateurs. À l'issue de l'annotation individuelle, des sessions d'adjudication et de normalisation des annotations se sont tenues avec les trois annotateurs pour arriver à une annotation unique des 287 noms d'action sur une échelle de 0 (pas technique) à 2 (très technique). Nous avons pris la décision de passer d'une échelle à six niveaux à une échelle à trois niveaux lors de l'adjudication pour offrir une plus grande souplesse lors de l'annotation par les locuteurs, mais limiter ensuite le désaccord et rendre plus opérationnalisable sur un plan statistique la mesure.

La normalisation des annotations a impliqué l'attribution par défaut d'un score normalisé de 0 (correspondant à une technicité nulle) lorsque les annotateurs donnent tous un score de 0 ou 1, un score normalisé de 1 (technicité moyenne) lorsque les scores des annotateurs oscillent entre 2 et 3, et un score normalisé de 2 (technicité élevée) lorsque les annotateurs s'accordent sur un score de 4 ou 5. Il y avait adjudication lorsque les scores attribués par les annotateurs empiétaient sur deux niveaux de technicité (scores de 3 et 4 par exemple). L'adjudication a mené à différentes décisions que nous présentons

1. Les principes d'identification des noms converts sont détaillés en section 10.2.2.

ci-après.

L'adjudication permet de traiter des cas relativement simples comme *touraillage* ou *intensification*. Le nom *touraillage*, dénotant une étape du traitement du malt, n'a pas fait l'objet d'une attribution automatique d'un score normalisé car il a été annoté respectivement 3, 4 et 4 par les 3 annotateurs, ce qui le place donc à cheval entre moyennement technique (par annotateur 1) et très technique (par annotateurs 2 et 3). Après discussion, il a été décidé de normaliser son score de technicité à 2 (très technique, sur l'échelle de 0 à 2) car ce nom s'avère obscur pour les non experts, l'action qu'il dénote implique un savoir-faire particulier et n'est pas faite au quotidien par le tout-venant, et implique l'utilisation d'un instrument (la touraille), même si l'instrument n'est pas caractérisé par un haut niveau de complexité, et l'action elle-même non plus (séchage par air chaud). Dans le cas de *intensification*, l'adjudication a permis de trancher entre action non technique (annotateurs 2 et 3, avec un score respectif de 1 et 0) et action moyennement technique (annotateur 1, avec un score de 2). Un score normalisé de 0 (non technique) a été attribué après adjudication à *intensification* car le nom n'implique pas nécessairement d'agent, ni d'action volontaire.

L'adjudication a été plus complexe pour des noms comme *arabisation* et *conscientisation*, qui relèvent de procédures spécifiques, mais qui n'appartiennent pas à des domaines techniques. Ces noms, se trouvant en périphérie, relèvent d'actions cognitives plus ou moins volontaires et ne nécessitant pas forcément un expert. Les annotateurs étaient en désaccord sur le caractère plus ou moins technique (respectivement 3 et 4 pour l'annotateur 1) et non technique (0 pour les annotateurs 2 et 3). L'adjudication a abouti à l'attribution d'un score normalisé de 0 (non technique) pour les deux noms, car même s'ils impliquent des phénomènes particuliers ou des connaissances particulières, la réalisation de l'action qu'ils dénotent n'impliquent pas l'apprentissage d'un savoir-faire particulier et ne nécessitent pas l'action volontaire d'un agent identifié dans un contexte bien spécifique. Le nom *arabisation* désigne par exemple un processus qui s'étend dans le temps et qui implique des populations et pas juste un individu. Plus largement, les actions relevant de processus conceptuels (comme *stylisation*) sont annotés comme non techniques et se voient attribués un score normalisé de 0 pour ces mêmes raisons.

Un autre cas traité lors de l'adjudication concerne les noms pour lesquels une acception technique et une acception non technique coexistent. La perception de la technicité dépend alors de l'annotateur. C'est notamment le cas de noms comme *polarisation*, qui peut correspondre à une réalité abstraite, comme dans le syntagme *la polarisation du débat*, et à une réalité physique et intrinsèquement technique, comme dans le syntagme *la polarisation des élec-*

trodes. La décision a été prise dans ces cas là d'attribuer le score normalisé de 1 (moyennement technique).

L'annotation de la technicité perçue se caractérise de fait par une grande variation inter-locuteur (voire intra-locuteur) puisqu'elle est fortement dépendante de l'expérience linguistique individuelle. La technicité est par essence une appréciation personnelle, qu'il est délicat de capter au travers d'un guide d'annotation ou d'un score unique.

Notons par ailleurs que les trois annotateurs n'étaient pas naïfs, puisqu'ils connaissaient l'objectif de la tâche et l'hypothèse qui la sous-tendait d'une plus grande technicité des noms en *-age*. Il y a donc un risque de biais dans l'annotation qui aurait pu amener les annotateurs à surévaluer le caractère technique des noms en *-age*, et à sous-évaluer les autres noms d'action, pour prouver l'hypothèse testée. À ce titre, il serait nécessaire de reprendre la tâche d'annotation pour impliquer un plus grand nombre d'annotateurs, qui ne seraient pas familiers de l'hypothèse testée, et couvrant un plus grand nombre de noms.

Cette première tâche permet néanmoins de travailler sur la base de 287 noms d'action pour lesquels un score normalisé est fourni. Nous détaillons les résultats de cette première phase d'annotation dans la section 11.2.2.

11.2.2 Résultats de l'annotation

À l'issue de l'annotation, 198 noms ont été annotés comme non techniques (0), 68 comme moyennement techniques (1) et 21 comme très techniques (2). Le détail de la répartition des noms en fonction de leur construction morphologique et de leur score de technicité perçue est donné dans le tableau 11.1.

Score	*-age*	*-ion*	*-ment*	*conv*	*autre*
0	19	94	40	19	26
1	19	38	6	2	3
2	9	9	1	0	2

TABLE 11.1 – Nombre et type morphologique des noms d'action en fonction de leur score de technicité perçue

Le tableau 11.1 montre que la proportion de *-age* au sein de chaque groupe augmente à mesure que le score de technicité perçue augmente. Ainsi, les noms en *-age* représentent 43% des noms jugés comme très techniques (2), alors qu'ils ne représentent que 10% des noms jugés non techniques (0). *A contrario*, les noms en *-ion* représente toujours à peu près la moitié de l'effectif

(48% pour 0, 56% pour 1 et 43% des 2) de chaque groupe. La forte présence des noms en *-ion* s'explique par la sur-représentation de ce suffixe dans les données initialement soumises à l'annotation.

Cela suggère que les noms jugés non techniques tendent à être moins souvent en *-age* que les noms jugés techniques. Par extension, on peut voir ces résultats comme un premier indice de la technicité des noms d'action en *-age*, même si cela ne préjuge en rien du degré de technicité des noms d'action en *-ion*.

Les scores moyens de technicité perçue pour chaque procédé dérivationnel considéré sont donnés dans le tableau 11.2.

	min	Q1	moyen	mediane	Q3	max
-age	0	0	0.7872	1	1	2
-ion	0	0	0.3972	0	1	2
-ment	0	0	0.1702	0	0	2
conversion	0	0	0.09524	0	0	1
autre	0	0	0.2258	0	0	2

TABLE 11.2 – Jugement de technicité des noms d'action

Le tableau 11.2 confirme de façon plus nette la tendance des noms d'action en *-age* à être perçus comme plus techniques que les noms d'action en *-ion*. En effet, on constate que le score moyen de technicité perçue pour les noms d'action en *-age* est supérieur à celui des noms d'action en *-ion* (0.78 contre 0.40). Cela est par ailleurs conforté par la médiane, puisque plus de la moitié des noms d'action en *-age* ont un score de technicité supérieur ou égal à 1, alors que la médiane est à 0 pour les noms en *-ion*. Concernant le suffixe *-ment*, l'annotation de ce premier échantillon suggère que les noms en *-ment* sont encore moins techniques que les noms en *-ion*, avec un score moyen de technicité perçue de 0.17, 75% de ces noms ayant un score de technicité perçue inférieur ou égal à 0 (contre 1 pour *-ion*). Le tableau suggère par ailleurs que la conversion est le procédé formant les noms les moins techniques.

Notons cependant que ces premiers résultats ont été obtenus sur la base d'un échantillon relativement limité de noms. Afin de tirer de réelles conclusions sur la différence de technicité des noms d'action en *-age*, *-ion* et *-ment*, il nous faut réaliser une annotation à plus grande échelle. Nous proposons dans la suite une approche automatique de cette annotation.

11.3 Modélisation statistique de la technicité

L'annotation par des locuteurs est une tâche coûteuse parce que chronophage. Nous cherchons dans la suite à opérationnaliser cette notion de technicité de façon à pouvoir annoter de façon automatique la technicité d'un nom selon des critères empiriques qui puissent être quantifiés. Il s'agit de proposer une annotation reproductible, indépendante de la subjectivité des locuteurs, sur la base d'une définition quantitative de la technicité, de façon à pouvoir annoter un nombre plus important de noms de façon plus rapide. Nous présentons dans ce qui suit la mise en place de critères statistiques sur la base de notre définition, ainsi que son opérationnalisation (section 11.3.1). Nous analysons ensuite la technicité des noms d'action au regard de cette annotation (section 11.3.2) ainsi qu'au regard de la technicité perçue par les locuteurs (section 11.3.3). Enfin, nous évaluons dans quelle mesure ces critères permettent de différencier les noms en *-age*, *-ion* et *-ment* (section 11.3.4).

11.3.1 Sélection de critères

Dans la section 11.1.3, nous avons proposé une définition de la technicité ainsi que des corrollaires linguistiques que nous rappelons ici. La première est la spécialisation (*T1*). Nous considérons ainsi qu'un nom d'action technique, du fait de la spécificité de l'action qu'il dénote, revêt un caractère spécialisé se traduisant par un usage différencié en fonction de la spécialisation du contexte d'énonciation. Sa spécificité se traduit par ailleurs par son caractère obscur (*T2*) pour des non-spécialistes, favorisant son apparition dans des ressources lexicographiques ou encyclopédiques. Enfin, cette spécificité a pour conséquence l'univocité du nom (*T3*), qui aura donc tendance à être moins polysémique que des noms non techniques.

Ces propriétés restent à opérationnaliser. Nous allons les traduire en mesures permettant d'évaluer le degré de spécialisation, d'obscurité et d'univocité du nom d'action. Pour cela, nous sommes guidée par deux exigences : la possibilité d'automatiser ces calculs, et la nécessité de s'appuyer sur des ressources linguistiques disponibles en français. Le tableau 11.3 présente les critères que nous dérivons des propriétés linguistiques précédemment présentées.

Le degré de spécialisation d'un nom (propriété *T1*) est estimé en comparant son emploi dans un corpus technique et un corpus de référence (Lemay *et al.* 2005) et au moyen de lexiques qui fournissent des informations sur l'appartenance à un domaine et sur la transdisciplinarité (Hatier 2016), conçue comme un critère de non spécialisation. Concernant la propriété d'obscurité,

Propriété	Critère
T1 - Spécialisation	- Ratio des fréquences relatives dans un corpus technique et un corpus de référence - Nombre de marqueurs lexicographiques de domaines dans des dictionnaires ou encyclopédies - Présence ou absence dans des lexiques scientifiques transdisciplinaires
T2 - Obscurité	- Présence ou absence d'une entrée dans une encyclopédie
T3 - Univocité	- Nombre de synonymes - Nombre de définitions dans des dictionnaires - Présence ou absence dans des lexiques de noms génériques

TABLE 11.3 – Critères de technicité des noms d'action

nous utilisons un unique critère : la présence ou absence de ce nom comme entrée dans une encyclopédie. Nous faisons ainsi l'hypothèse qu'un nom technique comme *thermoformage* aura une plus grande probabilité de faire l'objet d'une entrée encyclopédique qu'un nom sous-spécifié comme *démarche*. L'univocité est approchée par son opposé, l'équivocité, que nous approximons à l'aide de plusieurs critères, à savoir la polysémie et la sous-spécification.

Ces critères sont calculés à partir de différentes ressources que nous présentons dans le tableau 11.4.

Dans cette étude, nous choisissons le corpus *LM10* comme corpus de référence puisqu'il est considéré comme un bon exemple, parmi d'autres, de discours non technique. Le choix du corpus *Wikipedia2018* comme corpus technique est soutenu par sa nature encyclopédique, par le fait qu'il intègre un large choix de domaines techniques, mais aussi par l'absence de larges corpus techniques diversifiés pour le français. D'autres corpus ont ainsi été considérés, tels que Scientext (trop petit), frWaC (plus bruité), ou encore un corpus de brevets (constitution trop chronophage dans le cadre de ce travail, et périmètre thématique trop limité), ce qui pourrait être une piste de poursuite du travail.

Le corpus *Wikipedia2018* est utilisé pour le ratio des fréquences mais aussi pour tester la présence ou absence d'un article décrivant l'action dénotée par le nom. Seuls les articles dont le titre est strictement identique au nom sont pris en compte. Par exemple, nous considérons que *serrage* ne fait pas l'objet d'un article dans *Wikipedia2018*, même si on trouve des articles intitulés *collier de serrage* ou *noix de serrage*, car on ne trouve aucun article simplement

Nom	Taille	Description
Wikipedia2018	600 millions de mots	Corpus encyclopédique construit à partir de la version française de Wikipedia (dump de 2018)
LM10	200 millions de mots	Corpus journalistique français constitué d'articles du journal Le Monde publiés entre 1991 et 2000
DES	83 395 entrées	Dictionnaire électronique des synonymes (Manguin *et al.* 2004)
TLFi	54 280 entrées	Version électronique du dictionnaire Trésor de la Langue Français (Dendien et Pierrel 2003)
GLAWI	1 481 346 entrées	Dictionnaire électronique construit à partir de la version française du *Wiktionary* (Hathout et Sajous 2016)
LexiTrans	1 611 entrées	Lexique scientifique transdisciplinaire (LST) (Drouin 2010)
LexNSS	305 entrées	Liste de noms sous-spécifiés (NSS) extraits de (Legallois et Grea 2006)

TABLE 11.4 – Ressources et lexiques utilisés pour le calcul des critères de technicité

intitulé *serrage*. Concernant les lexiques, nous utilisons à la fois de larges dictionnaires généraux du français (*Trésor de la Langue Française* et *Glawi*) et des lexiques plus modestes et spécifiques qui donnent accès à la synonymie, au vocabulaire transdisciplinaire (comme indice de non spécialisation) et aux noms capsules (comme indice d'équivocité). Nous choisissons d'utiliser plusieurs dictionnaires afin de réduire l'impact de choix lexicographiques spécifiques.

Nos critères ainsi que les ressources sur lesquelles on les évalue sont donnés dans le tableau 11.5.

Comme nous pouvons le voir dans le tableau 11.5, nous utilisons des mesures simples qui vérifient la présence d'un nom dans des lexiques ou qui compte le nombre d'items lexicaux (définitions, synonymes) que l'on trouve dans une entrée. Le ratio des fréquences est calculé en divisant pour un nom donné sa fréquence relative dans le corpus techniques par sa fréquence relative dans le corpus de référence (Hatier 2016). Pour des raisons statistiques, nous ajoutons systématiquement un millionnième aux fréquences relatives. Ce choix est motivé par l'absence de certains noms dans le corpus *LM10*, qui

Propriété	Nom	Description
T1	RATIO_FREQR	Ratio des fréquences relatives (par million de mots) dans *Wikipedia2018* et *LM10*
	NB_CAT_W18	Nombre de marqueurs de catégorie dans *Wikipedia2018*
	NB_DOM_T	Nombre de marqueurs lexicographiques de domaine dans TLFi
	NB_DOM_G	Nombre de marqueurs lexicographiques de domaine dans GLAWI
	LST	Présence ou absence dans LexiTrans
T2	PAGE_W18	Présence ou absence d'un article dans *Wikipedia2018*
T3	NB_SYN	Nombre de synonymes dans le *DES*
	NB_DEF_T	Nombre de définitions dans le TLFi
	NB_DEF_G	Nombre de définitions dans GLAWI
	NSS	Présence ou absence du LexNSS

TABLE 11.5 – Implémentation des critères de technicité

se voyaient dès lors attribuer une fréquence de 0, et qui empêche le calcul du ratio des fréquences.

Ces critères sont exploratoires et nous sommes consciente que ce premier essai d'approximation de la technicité doit être raffiné à l'avenir au regard des résultats préliminaires que nous présentons en section 11.3.2 et suivant.

Notons que la technicité est estimée par la combinaison de ces critères. Ils visent à mettre en évidence des tendances concernant le degré de technicité des noms d'action. À ce titre, il n'y a pas de seuil de valeur à utiliser comme indice de caractérisation binaire d'un nom comme technique ou non technique.

11.3.2 Annotation automatique de la technicité

Les critères que nous venons de présenter nous permettent de tester empiriquement l'hypothèse d'une plus grande technicité des noms en *-age* et d'une plus faible technicité des noms en *-ion* et *-ment*. Suivant la définition de la technicité que nous donnons en section 11.1.3, nous nous attendons donc à ce que les noms d'action en *-age* aient des valeurs plus élevées que les noms en *-ion* et *-ment* pour deux critères liés à *T1* et *T2*, et des valeurs plus faibles pour les autres critères : ils auront un ratio de fréquence plus élevé (RATIO_FREQR) et auront une plus grande probabilité d'avoir un article

dans le corpus *Wikipedia2018* (PAGE_W18), mais ils seront moins représentés dans le lexique transdisciplinaire (LST) et parmi les noms sous-spécifiés (NSS), auront moins de synonymes (NB_SYN) de définitions (NB_DEF) et de marqueurs de domaines (NB_DOM).

Pour tester nos prédictions quant aux critères de technicité, nous annotons automatiquement les 2 443 noms d'action sélectionnés en section 10.2.1. Le tableau 11.6 présente l'annotation de quatre noms sélectionnés pour illustrer l'opposition entre noms techniques (*alunissage* et *cimentage*) et non techniques (*correction* et *revendication*), conformément à notre définition de la technicité. Nos prédictions sont indiquées dans le tableau par les symboles (+) et (-).

Critère	Technicité	*alunissage*	*cimentage*	*correction*	*revendication*
RATIO_FREQR	+	3.06	2.18	1.14	0.23
NB_CAT_W18	-	3	0	0	0
NB_DOM_T	-	1	2	7	3
NB_DOM_G	-	1	0	4	1
LST	-	Non	Non	Non	Non
PAGE_W18	+	Oui	Non	Oui	Non
NB_SYN	-	1	0	87	45
NB_DEF_T	-	1	3	33	6
NB_DEF_G	-	1	1	8	3
NSS	-	Non	Non	Non	Oui

TABLE 11.6 – Scores de technicité des noms *alunissage*, *cimentage*, *correction* et *revendication*

Tout d'abord, les résultats présentés dans le tableau 11.6 montrent que les 4 exemples sont plutôt bien décrits en termes de technicité par les critères que nous avons implémentés, bien que certains critères, considérés invididuellement, ne se conforment pas systématiquement à nos attentes. Par exemple, l'absence d'un nom dans le lexique transdisciplinaire (LST) ne garantit pas sa technicité, puisqu'il peut relever du vocabulaire non scientifique (*revendication*). Par ailleurs, la présence de plusieurs définitions (NB_DEF_G et NB_DEF_T) n'est pas nécessairement un bon indice de la non technicité d'un nom (*cimentage*). Pourtant, les noms techniques *alunissage* et *cimentage*, dénotant respectivement une manœuvre spécifique d'un engin spatial et un procédé de l'industrie de la construction, ont plusieurs valeurs qui s'approchent de nos attentes (valeur élevée pour le ratio de fréquences RATIO_FREQR, moins de synonymes, de définitions et de marqueurs de domaines, respectivement NB_SYN, NB_DEF et NB_DOM). De la même façon, la non technicité semble être globalement bien capturée, comme le montrent les noms

correction et *revendication*. Tous deux ont un nombre élevé de synonymes et de définitions, et pour *correction*, un nombre important de marqueurs de domaines.

Le tableau 11.6 souligne le rôle de nos critères comme des indicateurs de tendances et non comme des délimiteurs de classes. Nous n'avons pas de noms techniques et non techniques en tant que tel, mais des noms qui ont un degré de technicité plus important que d'autres. Parmi les noms qui ont des degrés de technicité élevés au regard de nos critères, c'est-à-dire des noms qui se conforment globalement à la plupart de nos prédictions, on retrouve *hydroformage*, *zingage*, *cardage* et *oxycoupage*, et parmi ceux au plus faible degré de technicité, les noms *association*, *division*, *commencement* et *approbation*.

Nous fournissons dans le tableau 11.7 les valeurs moyennes des critères de technicité pour les noms d'action en *-age*, *-ion* et *-ment*. Nous fournissons aussi à titre indicatif les valeurs obtenues pour les noms convertis (catégorie *conv*) et pour l'ensemble des autres noms d'action (catégorie *autre*) extraits de Lexeur et conservés suite à notre filtrage. La présence dans les lexiques de noms transdisciplinaires et sous-spécifiés est présentée sous la forme d'un pourcentage de noms appartenant à ces lexiques. L'existence d'un article correspondant au nom dans le corpus *Wikipedia2018* est aussi donné sous la forme d'un pourcentage.

Critères	*-age*	*-ion*	*-ment*	conv	autre
Ratio_freqr (10^5)	4.81	2.66	3.05	1.03	1.97
Nb_cat_w18	0.78	0.93	0.42	0.92	0.9
Nb_dom_t	1.2	2.65	1.27	3.6	1.99
Nb_dom_g	0.65	0.93	0.46	1.37	0.96
Lst (%)	0.2	8.13	2.23	11.08	6.21
Page_w18 (%)	48.97	69.87	34.08	77.85	58.28
Nb_syn	3.03	13.83	11.01	27.78	18.23
Nb_def_t	2.52	5.8	4.34	10.61	6.12
Nb_def_g	2.01	2.54	1.98	7.62	4.39
Nss (%)	0	2	1.11	6.15	2.07

Table 11.7 – Annotation des noms d'action

Les résultats présentés dans le tableau 11.7 corroborent l'hypothèse d'un plus grand degré de technicité des noms d'action en *-age*, et d'un plus faible degré de technicité des noms d'action en *-ion*. Nous pouvons voir que les noms en *-age* ont en moyenne un nombre significativementplus faible de sy-

nonymes[2] (NB_SYN), de définitions[3] (NB_DEF) et de marqueurs de domaines[4] (NB_DOM) que les noms en *-ion*. Par ailleurs, ils sont proportionnellement moins présents dans le lexique transdisciplinaire (significativement sauf pour 'autre') et parmi les noms sous-spécifiés que les noms en *-ion*. Notons que les différences de valeurs affichées par les suffixes sont plus élevées pour les mesures extraites du TLFi que pour celles extraites de *GLAWI*.

Cependant, un critère ne va pas dans le sens de nos prédictions, à savoir PAGE_W18 (présence ou absence d'une page dans *Wikipedia2018*). Le pourcentage de noms ayant une page dans *Wikipedia2018* est significativement plus faible pour le suffixe *-age* que pour le suffixe *-ion* (49% pour le premier, contre 70% pour le second). Bien que la différence entre chaque paire de suffixes soit statistiquement significative (à l'exception des noms étiquetés comme *autre*), le critère échoue dans l'approximation de la technicité des noms d'action telle que nous l'avions envisagée. Il semble au contraire diagnostiquer leur non-technicité.

Une remarque doit être faite concernant le critère du ratio des fréquences relatives (RATIO_FREQR). En effet, on constate que si le ratio était compris entre 0.23 et 3.06 dans le cas des noms *alunissage*, *cimentage*, *correction* et *revendication* (tableau 11.6), il s'avère ici bien plus élevé (entre 103 000 et 481 000). Cela s'explique par l'ajout d'un millionième à chaque fréquence relative. Avec cet encodage des fréquences relatives, on se retrouve de fait avec des valeurs aberrantes. C'est notamment le cas de *zingage*, qui a une fréquence relative de 4.43243 dans le corpus *Wikipedia2018*, mais qui est absent du corpus LM10. Suite au recodage des fréquences, *zingage* se voit donc avec un ratio des fréquences relatives de 4 432 433. Un total de 210 noms reçoivent ainsi des valeurs extrêmes qui viennent gonfler la moyenne des ratios. Notons malgré cela que la différence entre les valeurs moyennes affichées par les différents procédés dans le tableau 11.7 n'est significative que pour la distinction entre les conversions et les noms en *-age*. L'incapacité de ce critère à discriminer les noms d'action techniques des noms non techniques peut probablement s'expliquer par le choix du corpus *Wikipedia2018*. Il n'est sans doute pas suffisamment technique, et n'est donc par conséquent pas approprié pour évaluer les propriétés concernant la spécialisation et l'obscurité. L'inadéquation du corpus semble corroborée par le critère NB_CAT_W18

2. Le test de Tukey post-hoc montre que la différence n'est pas significative pour *-ment* et *-ion*.

3. La différence n'est pas significative pour *-ment* et *-age* dans le cas de GLAWI, et pour *-ion* et *autre* dans le cas de TLFi.

4. La différence n'est pas significative pour *-ment* et *-age* d'une part et *-ion* et *autre* d'autre part pour GLAWI, et pour les paires *-ment*/*-age*, *-ion*/*autre* et *-ment*/*autre* dans le cas du TLFi.

(nombre de catégories dans *Wikipedia2018*), qui ne s'avère pas significatif (sauf dans l'opposition de *-ment* aux autres procédés).

Le tableau 11.7 montre aussi que les noms en *-ment* sont dans une position intermédiaire relativement à nos critères. On constate que les valeurs moyennes pour *-ment* se situent entre celles de *-age* et de *-ion* pour les critères tels que le nombre de synonymes, le nombre de définitions dans le TLFi ou la présence dans les lexiques de noms transdisciplinaires et sous-spécifiés. Les valeurs moyennes pour les autres critères comme le nombre de catégories dans *Wikipedia2018* ou le nombre de marqueurs de domaine dans les deux dictionnaires est plus faible que pour *-age*. *A contrario*, les valeurs moyennes du suffixe *-ment* sont rarement supérieures à celles de *-ion*. Ces observations suggèrent que les noms d'action en *-ment* sont plus proches de ceux en *-age* que de ceux en *-ion* relativement à la technicité des actions qu'ils dénotent.

Nous pouvons nous demander à ce stade dans quelle mesure les critères que nous proposons sont bien complémentaires ou au contraire se recouvrent, et notamment dans le cas des nombres de définitions et de marqueurs de domaines, puisqu'il s'agit des mêmes mesures réalisées sur des ressources différentes. Nous donnons les scores de corrélation des différents critères dans la figure 11.1.

La figure 11.1 montre une forte corrélation positive entre les critères NB_DEF_T et NB_DOM_T : plus on a de définitions dans *TLFi*, plus on a de marqueurs de domaines. Le même constat, dans une certaine mesure, peut être fait pour ces mêmes critères calculés à partir de *GLAWI*. Cela va dans le sens de la polysémie des noms non techniques. On observe ainsi une corrélation positive entre le nombre de synonymes d'une part, et le nombre de définitions et de marqueurs de domaines (dans *TLFi* comme dans *GLAWI*). Il existe par ailleurs une corrélation positive non négligeable entre la présence dans le lexique scientifique transdisciplinaire d'une part et le nombre de synonymes, de définitions et de marqueurs de domaines d'autre part. Les noms dans le LST tendent à avoir de nombreuses définitions et de nombreux marqueurs de domaines. Enfin, on observe une corrélation positive notable, bien que prévisible, entre la présence d'une page dans le corpus *Wikipedia2018* et le nombre de catégories dans *Wikipedia2018* – qui s'explique par le fait que cette valeur n'est pertinente que lorsqu'il y a une page –, mais aussi avec le nombre de synonymes, de définitions et de marqueurs de domaines dans *GLAWI* et *TLFi*. Ces corrélations semblent confirmer que le critère PAGE_W18 ne pointe pas dans le sens de la technicité, ne validant pas notre hypothèse.

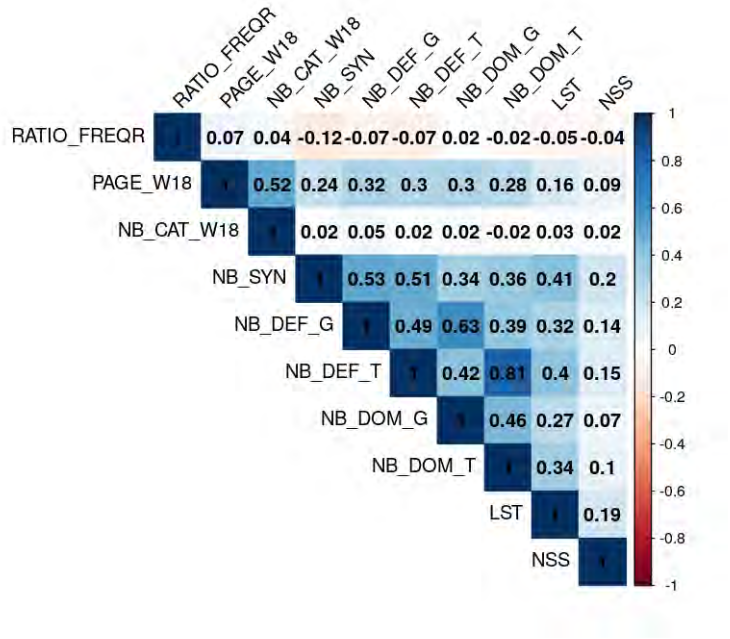

FIGURE 11.1 – Matrice de corrélation des critères de technicité

11.3.3 Corrélation avec la technicité perçue

Les résultats obtenus avec les critères empiriques semblent converger avec ceux obtenus sur la base du jugement des locuteurs. Pour quantifier cela, nous calculons la corrélation entre les critères calculés automatiquement et les scores de technicité perçue. La matrice de corrélation est donnée dans la figure 11.2. Le jugement des locuteurs correspond à NORM.

La figure 11.2 montre une corrélation négative entre le jugement et le nombre de synonymes (-0.24) ainsi qu'avec le nombre de définitions dans le *TLFi* (-0.23), et dans une moindre mesure avec le nombre de définitions dans *GLAWI* (-0.16) et avec la présence dans le lexique scientifique transdisciplinaire (-0.13). Plus le nombre de définitions et de synonymes augmente, moins le nom est jugé technique par les locuteurs. Par ailleurs, on observe une faible corrélation positive entre la technicité perçue et le ratio des fréquences (0.19). Les autres valeurs semblent indiquer une corrélation quasi nulle entre les autres critères et la technicité perçue.

La comparaison de cette matrice avec celle présentée en figure 11.1 fait

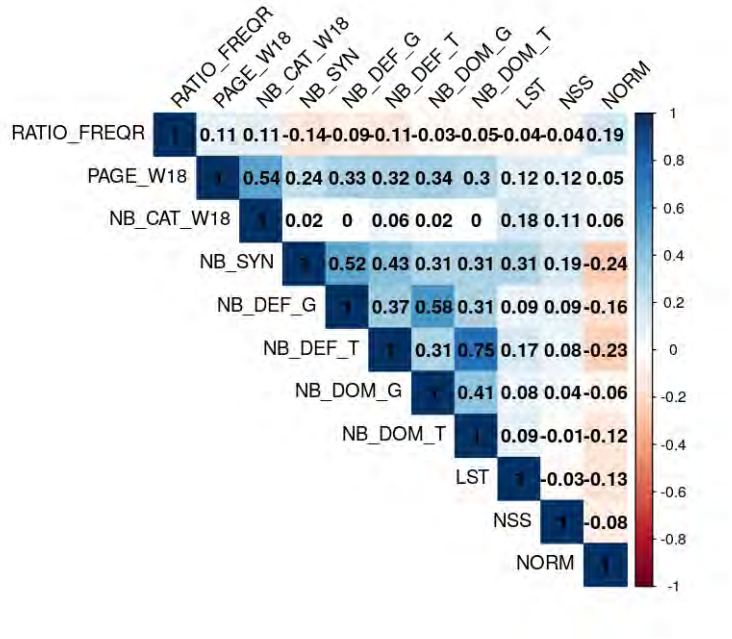

FIGURE 11.2 – Matrice de corrélation des critères de technicité et de la technicité perçue

émerger des variations concernant les corrélations entre critères. Cela s'explique par le fait que la seconde matrice a été calculée à partir d'un échantillon limité d'items, ce qui ne permet pas d'aboutir aux mêmes généralisations. Cela est notamment illustré par la corrélation entre le nombre de définitions et le nombre de marqueurs de domaines dans TLFI qui diminue passant de 0.8 à 0.7. Les tendances restent cependant similaires.

Nous évaluons dans quelle mesure le score de technicité perçue par les locuteurs est expliquée par nos critères de technicité à l'aide d'une régression linéaire [5] pour voir dans quelle mesure nos critères permettent d'expliquer la valeur du jugement de locuteurs.

Les prédicteurs conservés suite à la régression pas à pas (c'est-à-dire les prédicteurs les plus significatifs) sont le ratio des fréquences, la présence d'une page dans le corpus *wikipedia2018*, le nombre de synonymes et le nombre de

5. Le modèle de régression linéaire est estimé à l'aide de la fonction *train* de la librairie *caret*, méthode *lmStepAIC*, avec l'option *train control* par validation croisée à 10 *folds*.

	Estimate	Std error	t value	Pr (>|t|)
(Intercept)	0.392474	0.057011	6.884	3.77e-11
RATIO_FREQR	0.017852	0.008345	2.139	0.03326
PAGE_W18	0.174830	0.075553	2.314	0.02139
NB_SYN	-0.005001	0.001725	-2.899	0.00404
NB_DEF_T	-0.018720	0.006659	-2.811	0.00528

TABLE 11.8 – Coefficients du modèle de régression linéaire

définitions dans le *TLFi*. Le modèle confirme une nouvelle fois l'influence négative des synonymes et des définitions (plus leur nombre augmente, plus la technicité diminue), et l'influence positive de la présence d'une page et du ratio des fréquences. Il est intéressant de noter que la présence d'une page dans le corpus contribue à expliquer le jugement alors qu'il n'était pas spécialement corrélé au jugement (0.05).

Une première conclusion que l'on peut tirer de cette modélisation concerne la validation des propriétés linguistiques de la technicité. En effet, à l'exception de PAGE_W18 dont nous avons observé le caractère paradoxal vis-à-vis de la propriété d'obscurité (*T2*), les critères conservés instancient les propriétés de spécialisation (*T1*) et d'univocité (*T3*). Cela semble valider ces deux propriétés comme indices de technicité des noms d'action. Ce constat est cependant à nuancer à l'échelle de chaque critère, tous ne permettant pas d'expliquer le score de technicité perçue. Plusieurs hypothèses peuvent être faites à ce stade pour expliquer ce résultat. Tout d'abord, la taille de l'échantillon sur lequel la modélisation se base peut être insuffisant, puisque ce calcul de corrélation repose sur les 287 noms pour lesquels nous avions un score de technicité perçue. Le nombre d'items jugés techniques était ainsi sans doute trop faible (21) pour permettre une généralisation et donc une modélisation représentative. L'intégration d'un plus grand nombre de noms, suite à une nouvelle phase d'annotation et de normalisation, devrait permettre de confirmer ou non ces résultats. Deuxièmement, comme certains des critères automatiques sont corrélés, il est normal qu'ils ne soient pas tous intégrés dans le modèle. Enfin, il se peut que les critères choisis ne permettent effectivement pas d'estimer de façon pertinente la technicité des noms d'action, auquel cas il nous faut nous pencher sur de nouveaux critères.

11.3.4 Pouvoir discriminant des critères

Nous évaluons le potentiel prédictif de l'ensemble de nos critères empiriques de technicité pour voir dans quelle mesure la technicité, au travers de ces critères, permet de discriminer les noms d'action sur la base de leur

procédé dérivationnel. Leur pouvoir discriminant est estimé à l'aide d'un arbre de décision qui prédit le suffixe d'un nom à partir des valeurs des différents critères. Nous classons [6] l'ensemble des 2 443 noms d'action sur la base de l'ensemble des critères annotés automatiquement. L'arbre de décision est donné en figure 11.3. Les valeurs figurant sous les feuilles correspondent, dans le cluster, au nombre de noms relevant des cas suivants : noms en *-age*, noms relevant d'autres procédés, converts, noms en *-ion*, noms en *-ment*. Les valeurs catégorielles sont encodées par des indicateurs fictifs, 1 correspondant à Oui et 0 à Non.

La figure 11.3 montre que les critères utilisés par le modèle pour discriminer les suffixes sont le nombre de définitions dans GLAWI (NB_DEF_G), le nombre de synonymes (NB_SYN) et la présence d'une page dans *Wikipedia2018* (PAGE_W18). Les conversions sont principalement identifiées sur la base de leur nombre de définitions (supérieur ou égal à 6), les noms en *-age* sur leur nombre de définitions (inférieur à 6) et leur nombre de synonymes (inférieur à 3). Les noms en *-ion* et *-ment* sont identifiés par leur nombre de définitions (inférieur à 6), de synonymes (supérieur ou égal à 3), ainsi que la présence (pour les *-ion*) ou non (pour les *-ment*) d'une page dans *Wikipedia2018*.

Le modèle ainsi construit obtient une précision globale de 47%. Nous présentons la matrice de confusion associée dans le tableau 11.9. La matrice de confusion présente le type morphologique prédit par le modèle (en colonne) en fonction du type morphologique réel des items (en ligne). Le nombre d'items correctement prédits est indiqué en gras.

Le tableau 11.9 montre que la précision du modèle varie en fonction du suffixe ciblé. Ainsi, 77% des noms en *-age* sont bien classifiés, c'est-à-dire sont bien prédits comme des noms en *-age* par le modèle, contre 40% des noms en *-ion*, et 31% des noms en *-ment*. Les noms converts sont correctement discriminés à hauteur de 69%. Les noms construits par d'autres procédés ne sont quant à eux jamais correctement étiquetés, et le modèle ne prédit jamais l'étiquette *autre*.

Dans le cadre de leur concurrence, nous cherchons à évaluer plus précisément le potentiel prédictif de ces critères pour la discrimination des noms d'action en *-age*, *-ion* et *-ment*. Pour cela, nous classons exclusivement les 1 828 noms d'action en *-age*, *-ion* et *-ment* sur la base de l'ensemble des critères annotés automatiquement. L'arbre de décision est donné en figure 11.4. Les valeurs données sous les feuilles correspondent respectivement au nombre de noms en *-age*, *-ion* et *-ment* dans le cluster. La présence ou l'absence d'une page dans *Wikipedia2018* sont indiquées respectivement par les valeurs '=1'

6. Nous utilisons la fonction `rpart` du package `rpart` de R, méthode `class`.

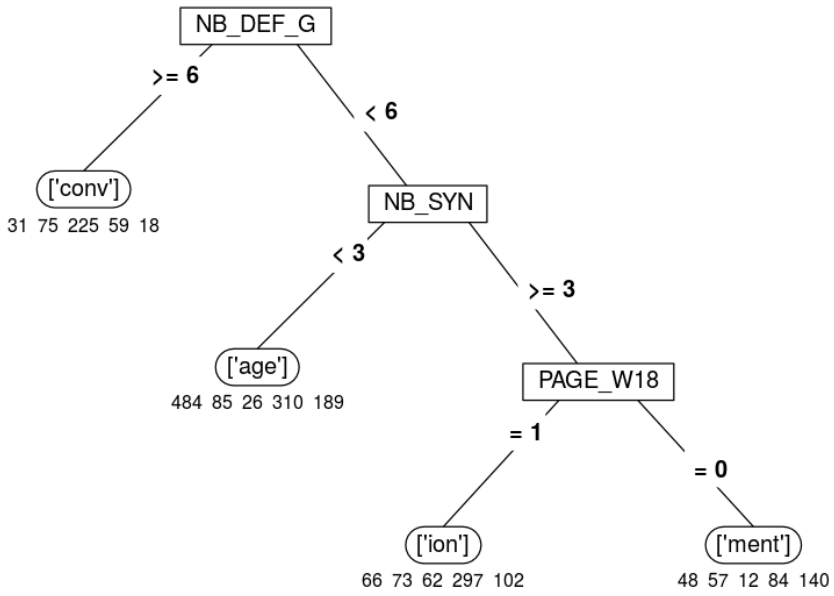

FIGURE 11.3 – Classification des noms d'action à partir des critères de technicité

et '=0' (les valeurs O et N ayant été remplacées par des indicateurs fictifs).

La figure 11.4 montre à la fois les règles de décision inférées à partir des données ainsi que la contribution des critères à la classification. On observe à nouveau que seuls 3 des 10 critères initialement fournis sont utilisés pour la classification des noms en *-age*, *-ion* et *-ment*. Si l'on retrouve les critères (NB_SYN) et PAGE_W18 précédemment utilisés pour classifier l'ensemble des noms d'action, on constate que le critère NB_DEF_G est remplacé par NB_DEF_T, son équivalent dans l'autre dictionnaire.

La règle associée au critère PAGE_W18, présente dans les deux modèles,

		Prédit				
		-age	*-ion*	*-ment*	conv	autre
Observés	*-age*	**484**	66	48	31	0
	-ion	310	**297**	84	59	0
	-ment	189	102	**140**	18	0
	conv	26	62	12	**225**	0
	autre	85	73	57	75	**0**

TABLE 11.9 – Matrice de confusion de la prédiction du suffixe des noms d'action

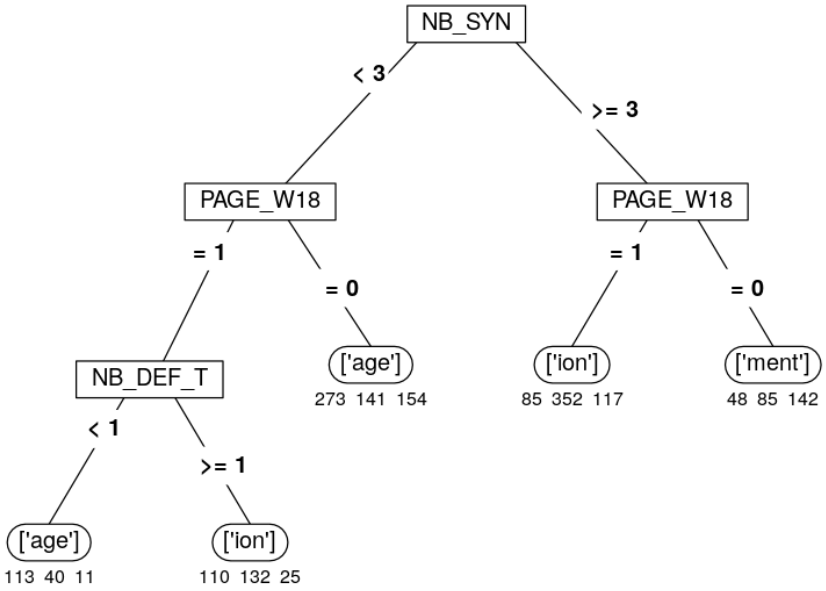

FIGURE 11.4 – Classification des noms d'action en *-age*, *-ion* et *-ment* à partir des critères de technicité

confirme les observations tirées du tableau 11.7, à savoir que le critère de la présence ou l'absence d'une page pour le nom n'opérationnalise pas de façon satisfaisante la propriété *T2*. L'exclusion de ce critère dégrade cependant la précision globale du modèle, qui passe à 53%, contre la précision de 55% obtenu par le modèle correspondant à l'arbre en 11.4.

La précision globale de 55% signifie qu'à peine plus de la moitié des noms d'action sont correctement classifiés par le modèle (c'est-à-dire que le suffixe est bien identifié). Tous les suffixes ne sont cependant pas classifiés de la même façon. Les performances du modèle sont données dans le tableau 11.10.

		Prédit		
		-age	*-ion*	*-ment*
Observés	*-age*	**386**	195	48
	-ion	181	**484**	85
	-ment	165	142	**142**

TABLE 11.10 – Matrice de confusion de la prédiction du suffixe des noms d'agent en *-age*, *-ion* et *-ment*

Le tableau 11.10 montre que 386 noms d'action en *-age* sont identifiés comme tels (soit 61% des *-age*), alors que 195 des noms en *-age* sont identi-

fiés comme des noms en *-ion* (31%) et 48 comme des noms en *-ment* (8%). Le modèle atteint des scores similaires pour les suffixes *-age* et *-ion* (respectivement 61% et 65%), mais une performance plus faible pour les noms en -ment (32%). On en conclut que globalement, sur la base des critères, les suffixes *-age* et *-ion* sont relativement bien identifiés, mais pas les noms en *-ment*. Les critères ne permettent pas de discriminer *-ment*. Autrement dit, les noms en *-ment* ne se distinguent pas aussi nettement que les noms en *-age* et *-ion* au regard de ces critères.

L'examen des noms mal classifiés montre que les trois critères NB_SYN, NB_DEF_T et PAGE_W18 ne discriminent peut-être pas parfaitement les noms en fonction de leur suffixe, mais plutôt en fonction de leur degré de technicité. En effet, parmi les noms en *-age* mal classifiés (c'est-à-dire qui ont été associés aux noms en *-ion* ou en *-ment*), nous trouvons des noms comme *papotage*, *batifolage* ou *babillage*, qui sont étiquetés comme nominalisations en *-ment* par le modèle. Cela s'explique par leur faible degré de technicité, qui se traduit par un nombre important de synonymes. Bien qu'ils soient mal classifiés sur le plan formel, ils sont néanmoins bien classifiés sur le plan sémantique, relativement à leur degré de technicité. D'autres noms mal classifiés sur le plan formel, à l'image de *damage* étiqueté comme nominalisation en *-ion* du fait de sa définition dans TLFi, montre que certaines règles sont un peu trop restrictives, et jouent un rôle un peu trop important.

L'analyse des erreurs met en lumière une certaine hétérogénéité au sein des noms en *-ion*. Ainsi, 11% des noms en *-ion* mal classifiés (c'est-à-dire prédits comme *-age* ou *-ment*) sont suffixés en *-ification* et *-isation* (*panification*, *dessalinisation*), et étiquetés comme *-age*. Ce sont ainsi 23% des noms en *-ification* et 58% des noms en *-isation* qui sont étiquetés comme des noms en *-age* par le modèle. Les noms en *-ification* et *-isation* ne sont étiquetés comme *-ion* par le modèle qu'à hauteur respectivement de 68% et 39%, et comme *-ment* à hauteur de 9% et 3%. À titre de comparaison, seuls 18% des noms en *-ion* sont étiquetés comme des noms en *-age* par le modèle, les autres noms en *-ion* étant étiquetés comme noms en *-ion* et *-ment* à hauteur de 69% et 13% respectivement. Les noms en *-ification* et, plus encore, en *-isation* tendent donc à être davantage étiquetés comme des noms en *-age* que ne le sont les autres noms en *-ion*. Cela suggère donc une plus grande technicité des noms en *-ification* et *-isation*, qui peut s'expliquer par leur construction à partir de verbes en *-iser* et *-ifier*. Ces verbes, construits à partir de bases nominales ou adjectives, dénotent des prédicats de changement d'état (Lignon 2013), impliqués dans de nombreux processus chimiques (*acidifier*), physiques (*électrifier*) ou biologiques (*momifier*). Ils sont ainsi fortement mobilisés dans le domaine scientifique pour exprimer la causalité (El Khamissy 2016). Les noms dérivés de ces verbes héritent ainsi de la tech-

nicité intrinsèque de leur base.

Le regroupement des noms en fonction de la technicité de leur suffixe tend à améliorer les résultats. La reprise de la classification, mais sur la base d'une nouvelle répartition des 1 828 noms en deux groupes basés sur la technicité supposée des suffixes, montre ainsi une nette progression de la précision. Nous constituons un premier groupe dont nous faisons l'hypothèse qu'il est plus technique, et qui regroupe les noms en *-age*, *-ification*, *-isation* et *-ment* d'une part, noté *-Age*), et un groupe dont nous faisons l'hypothèse qu'il n'est pas technique, constitué des noms en *-ion*, et noté *-Ion*. L'arbre ainsi construit, n'exploitant que les critères NB_SYN et PAGE_W18, obtient une précision globale de 74%. La matrice de confusion correspondante est présentée dans le tableau 11.11

		Prédit	
		-Age	*-Ion*
Observés	*-Age*	1085	152
	-Ion	330	261

TABLE 11.11 – Matrice de confusion de la prédiction de la technicité des noms d'action

L'analyse plus en détail des résultats montre que 92% des noms en *-age*, 77% des noms en *-ification*, 91% des noms en *-isation* et 82% des noms en *-ment* sont considérés comme des noms techniques par le modèle. Plus surprenant, on constate que 56% des noms en *-ion* sont considérés comme des noms techniques.

La même classification mais réalisée sur l'ensemble des 2 443 noms d'action, avec l'intégration des noms converts et des noms construits par d'autres procédés dans le groupe des noms non techniques *-Ion*, montre des résultats similaires, bien que légèrement moins bons, avec une précision globale de 71%. Les tendances précédemment mises en avant pour les suffixes *-age*, *-ification*, *-isation* et *-ment* d'une part, et *-ion* d'autre part, se maintiennent, bien que dans des proportions plus faibles. De nouvelles tendances émergent néanmoins de cette classification, à savoir la non technicité des noms converts (91% d'entre eux prédits comme appartenant au groupe *-Ion*) et l'étonnante hétérogénéité des noms construits par d'autres procédés, dont un peu moins de la moitié (44%) sont prédits comme appartenant au groupe des noms techniques *-Age*.

Ces résultats confirment d'une part que les noms en *-age*, *-ification*, *-isation* et *-ment* tendent à être plus techniques que les noms en *-ion*, les converts, et les noms construits par d'autres procédés. Mais cela montre

aussi d'autre part que qu'il existe une forte hétérogénéité, en termes de technicité, au sein des noms en *-ion* et des noms construits par d'autres procédés, puisque un peu moins de la moitié d'entre eux sont considéré comme techniques. Une analyse qualitative des noms mal classés reste à mener pour comprendre les raisons de cette hétérogénéité, mais ces résultats confirment néanmoins le pouvoir discriminant de nos critères relativement aux procédés dérivationnels des noms d'action.

Le constat d'une distinction entre les noms en *-ion*, *-isation* et *-ification* se confirme lorsque l'on regarde dans le détail les scores moyens de technicité perçue, en distinguant les noms en *-ification* et *-isation* des noms en *-ion* (tableau 11.12).

	min	Q1	moyen	mediane	Q3	max
-age	0	0	0.7872	1	1	2
-ion	0	0	0.2128	0	0	2
-isation	0	0	0.4894	0	1	2
-ification	0	0	0.4894	0	1	2
-ment	0	0	0.1702	0	0	2
conversion	0	0	0.09524	0	0	1
autre	0	0	0.2258	0	0	2

TABLE 11.12 – Scores moyens de technicité perçue des noms d'action (détails)

Cela est aussi confirmé par la reprise des scores obtenus par chaque suffixe à partir de chaque critère calculé automatiquement en corpus (tableau 11.13).

Critères	*-age*	*-ion*	*-ification*	*-isation*	*-ment*	conv	autre
RATIO_FREQR (10^5)	4.81	3.02	1.66	1.14	3.05	1.03	1.97
NB_CAT_W18	0.78	0.90	1.04	1.02	0.42	0.92	0.9
NB_DOM_T	1.2	3.04	1.06	1.51	1.27	3.6	1.99
NB_DOM_G	0.65	1.04	0.51	0.60	0.46	1.37	0.96
LST (%)	0.16	9.14	2.68	8.51	2.23	11.08	6.21
PAGE_W18 (%)	48.97	72.42	57.14	68.09	34.08	77.85	58.28
NB_SYN	3.03	16.43	2.38	8.40	11.01	27.78	18.23
NB_DEF_T	2.52	6.66	2.22	3.51	4.34	10.61	6.12
NB_DEF_G	2.01	2.79	1.54	1.83	1.98	7.62	4.39
NSS (%)	0	2.54	0	0	1.11	6.15	2.07

TABLE 11.13 – Annotation des noms d'action (détails)

Le tableau 11.13 montre que les valeurs moyennes des critères pour les noms en *-ification* et *-isation* sont globalement situées entre celles obtenues pour les noms en *-age* et *-ion*. Plus précisément, les valeurs des noms

en *-ification* sont soit inférieures à celles des noms en *-age* (notamment pour les critères NB_DOM_T et NB_DOM_G, NB_SYN, NB_DEF_T et NB_DEF_G) soit comprises entre celles des noms en *-age* et en *-ion* (notamment pour les critères PAGE_W18, LST et NB_CAT_W18). Les valeurs obtenues pour *-isation* sont quant à elles généralement comprises entre celles de *-ification* et de *-ion*, à l'image des critères NB_DOM_T, LST, PAGE_W18, NB_SYN ou encore NB_DEF_T. Cela suggère que les noms en *-isation*, bien que plus techniques que les noms en *-ion*, sont un peu moins techniques que les noms en *-ification*.

Les résultats combinés des arbres de décision et des différents scores de technicité – perçue ou calculée – permettent de dresser une première image des noms d'action en fonction de leur technicité sous la forme d'un continuum (figure 11.5). Aux extrémités se trouveraient les noms d'action en *-age* (très techniques) et les conversions (non techniques). Une étude plus approfondie devrait être faite pour identifier plus précisément les technicités relatives des noms en *-ification*, *-isation*, qui fluctuent en fonction des critères et scores considérés, mais ces noms semblent néanmoins globalement plus techniques que les noms en *-ion* et en *-ment*.

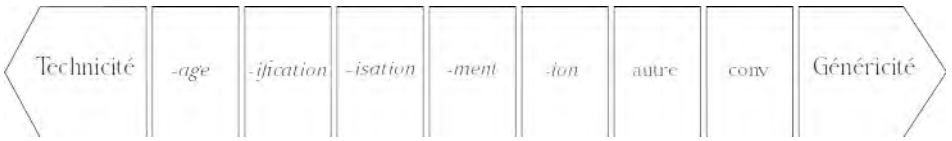

FIGURE 11.5 – Continuum de technicité des noms d'action

Cette expérience questionne la pertinence de certains de nos critères, en particulier ceux liés aux corpus. Bien qu'il soit utilisé de façon opposée à ce que l'on attendait, le critère PAGE_W18 améliore la précision globale du premier modèle, et particulièrement la classification des noms en *-ion* et *-ment*. Concernant le ratio des fréquences relatives, il se peut que le corpus *Wikipedia2018* ne soit pas suffisamment technique, et trop diversifié pour favoriser les noms en *-age*. Concernant la présence d'un article dans le corpus *Wikipedia2018*, il faudrait revoir son implémentation et l'hypothèse derrière ce critère.

Dans cette étude, nous avons exploré le facteur de technicité dans la concurrence entre les noms d'action français en *-age*, *-ion* et *-ment*. Nous avons dans un premier temps proposé une définition des noms d'action techniques comme étant des noms non familiers pour les non experts, qui dénotent une action complexe, spécifique, nécessitant des compétences acquises, et ancrées notamment dans des domaines comme l'industrie, l'agriculture ou l'artisanat. Nous avons montré que des propriétés linguistiques pouvaient être

dérivées de cette définition, et calculées à partir de corpus et de ressources lexicales, afin de caractériser la technicité des nominalisations. Certains de ces critères, comme le nombre de synonymes et de définitions, ont montré leur efficacité dans la discrimination des noms en *-age*, plus techniques, des noms d'action en *-ion*, moins techniques. D'autres critères se révèlent moins pertinents, notamment ceux calculés à partir des corpus, comme le ratio des fréquences. Nous avons aussi vu que les noms d'action en *-ion* sont plus hétérogènes qu'attendus relativement à leur technicité, principalement du fait de la présence des noms d'action en *-ification* et *-isation*, à la technicité inhérente.

Nous avons considéré dans ce travail la technicité comme une notion composite à laquelle les multiples dimensions (spécialisation, obscurité, univocité) contribuaient de façon similaire. Le fait que certains critères soient plus systématiquement explicatifs, à l'image du nombre de domaines, de définitions ou de synonymes, indique que le caractère monolithique de la notion de technicité n'est peut-être pas pertinent. Les différentes analyses montre que l'importance des dimensions dans la caractérisation de la technicité est variable, l'univocité jouant un rôle primordial. À l'inverse, l'obscurité ne semble pas contribuer à cette caractérisation, ou du moins pas comme attendu.

Ce travail est une première tentative de caractérisation et d'approximation de la technicité au travers de critères empiriques. Il offre un aperçu nouveau sur la distinction entre les noms d'action en *-age*, *-ion* et *-ment* du français, et met en lumière la limite de certains critères exploratoires. En particulier, il montre la nécessite d'améliorer les critères qui approximent l'obscurité. Nous souhaitons aussi explorer d'autres critères, à l'image de la concrétude (Köper et Im Walde 2016, Pierrejean 2020), ou l'instrumentalité (Missud 2019). Il y a en effet un certain recouvrement entre la notion de technicité et la notion plus basique de concrétude, puisque les noms d'action techniques dénotent principalement des actions concrètes (*usinage*, *extrusion*), par opposition à des actions plus abstraites, notamment cognitives (*compréhension*, *perception*). Ce lien entre concrétude et technicité est par ailleurs renforcé par la préférence du suffixe *-age* pour les thèmes concrets (Fradin 2016). Nous souhaitons aussi explorer de nouveaux critères exploitant les corpus, et l'utilisation du TF-IDF sur le corpus *Wikipedia2018*. On peut ainsi faire l'hypothèse qu'un nom technique aura plus de chance d'avoir un tfidf élevé qu'un nom non technique, car il apparaîtra dans un nombre plus limité d'articles qu'un nom sous-spécifié. Enfin, nous souhaiterions affiner la quantification de la technicité au travers des scores. D'une part, nous projetons de poursuivre et développer l'annotation de la technicité perçue, en déployant une annotation à plus large échelle – tant sur le nombre d'items annotés que d'annotateurs mobilisés. Nous souhaitons par ailleurs produire un

score de technicité unique, qui aggrège les critères pertinents et la technicité perçue.

Chapitre 12

La technicité dans les modèles distributionnels

Nous avons montré jusque-là que les noms d'action en *-age*, *-ion* et *-ment* se distinguent sur la base de leur degré de technicité. Cela a été confirmé dans le chapitre 11 au moyen de critères empiriques et par le jugement de plusieurs locuteurs, mais ce constat se basait initialement sur l'examen introspectif et contrastif dans le chapitre 10 des voisins des barycentres des noms d'action en *-age*, *-ion* et *-ment*.

On peut à ce stade se demander si cette différence de technicité caractérise réellement les barycentres constitués sur des bases formelles. En d'autres termes, on peut se demander si la distinction qui sous-tend les trois sous-classes formées par les noms d'action en *-age*, *-ion* et *-ment* dans les modèles distributionnels repose bien sur leur degré variable de technicité, et non sur l'homogénéité distributionnelle des procédés.

Cela amène à s'interroger sur la sensibilité des modèles distributionnels au trait sémantique de technicité. Nous avons vu que les modèles distributionnels permettaient d'analyser voire de discriminer sur le plan distributionnel des noms sur la base de leur agentivité (partie IV), de leur concrétude (Frassinelli *et al.* 2017, Pierrejean 2020) et de leur sous-spécification (Ho-Dac *et al.* 2020), entre autres (Gupta *et al.* 2015). On peut faire l'hypothèse que la technicité est présente au même titre dans les représentations distributionnelles.

Ce chapitre consacre deux approches complémentaires à cette question. Nous évaluons dans un premier temps la technicité des barycentres construits à partir des noms d'action en *-age*, *-ion* et *-ment* en analysant quantitativement leurs voisins (section 12.1). Dans un second temps, nous abordons la question par le biais des groupes de noms techniques et non techniques, afin d'évaluer la correspondance distributionnelle entre le trait sémantique de la technicité et la suffixation en *-age*, *-ion* et *-ment* (section 12.2).

12.1 Technicité des barycentres des noms d'action en *-age*, *-ion* et *-ment*

Nous souhaitons quantifier plus précisément la différence de technicité des barycentres constitués dans la section 10.2.1 afin d'évaluer dans quelle mesure les classes des noms d'action en *-age*, *-ion* et *-ment* se distinguent par leur degré variable de technicité dans les espaces distributionnels. Nous évaluons ce degré en analysant les voisins. Il s'agit ici de vérifier que les voisins du barycentre des noms d'action en *-age* ont un degré de technicité plus élevé que ceux des barycentres des noms d'action en *-ion* et *-ment*.

Nous pouvons d'abord confronter ces voisins au score de technicité perçue par les locuteurs. Ce score donne un indice clair de la technicité, mais il souffre d'une couverture limitée par rapport aux voisinages étudiés. Ainsi, respectivement 7, 9 et 7 des 100 plus proches voisins des barycentres des noms d'action en *-age*, *-ion* et *-ment* ont été annotés. Ces voisins sont donnés en (39a), (39b) et (39c). Leur score de technicité perçue est indiqué entre parenthèses.

(39) a. assemblage (0), gonflage (0), clouage (0), calibrage (1), aiguisage (1), utilisation (0), filage (2)

b. simplification (0), dispersion (0), utilisation (0), modification (0), cristallisation (1), acceptation 0, quantification (1), purification (1), identification (0)

c. utilisation (0), tassement (0), usure (0), glissement (0), arrachage (0), positionnement (0), clouage (0)

Bien que modeste par sa couverture, le score de technicité perçue des voisins présentés en (39) fait émerger une première tendance allant dans le sens d'une plus grande technicité des voisins du barycentre des noms d'action en *-age*. Ces derniers affichent ainsi un score moyen de technicité de 0.57 (sur une échelle de 0 à 2), 43% d'entre eux ayant un score de technicité perçue supérieur ou égale à 1. *A contrario*, seuls 33% des voisins du barycentre des noms d'action en *-ion* ont un score supérieur ou égal à 1, pour un score moyen de 0.33, contre un score moyen de 0 pour les voisins du barycentre des noms d'action en *-ment*. Ces résultats sont cependant issus d'échantillons très limités, et il est difficile d'en tirer une quelconque généralisation.

S'ils ont l'inconvénient de ne pas aggréger l'information en une seule valeur, les critères de technicité calculés automatiquement offrent cependant une meilleure couverture des voisins, puisque respectivement 62, 50 et 52 des 100 plus proches voisins des barycentres des noms d'action en *-age*, *-ion* et *-ment* font l'objet d'une annotation automatique dans le cadre de l'étude

présentée dans le chapitre 11. L'avantage de ces critères est que l'on peut annoter les 38, 50 et 48 de voisins restants. Les valeurs moyennes des critères pour les 100 voisins des barycentres des noms d'action en *-age*, *-ion* et *-ment* sont données dans le tableau 12.1.

Critères	*-age*	*-ion*	*-ment*
RATIO_FREQR	464000	1.60	1.06
NB_CAT_W18	1.19	0.78	0.37
NB_DOM_T	2.46	5.51	3.12
NB_DOM_G	0.9	1.66	0.92
LST (%)	2	25	13
PAGE_W18 (%)	72	92	54
NB_SYN	4.11	19.02	19.49
NB_DEF_T	4.54	9.5	6.75
NB_DEF_G	2.41	3.82	3.03
NSS (%)	0	5	2

TABLE 12.1 – Valeurs moyennes des critères des voisins des barycentres des noms d'action en *-age*, *-ion* et *-ment*

Le tableau 12.1 va globalement dans le sens de nos hypothèses. On observe des valeurs plus faibles pour le barycentre des noms d'action en *-age* pour le nombre de marqueurs de domaines NB_DOM_T et NB_DOM_G (la différence n'est pas significative [1] entre *-ment* et *-age* dans les deux cas), le nombre de définitions NB_DEF_T et NB_DEF_G (la différence est significative pour *-age* et *-ion* pour les deux ressources, et est aussi significative pour *-ion* et *-ment* pour le TLFi), le nombre de synonymes NB_SYN, la présence dans le LST. On observe par ailleurs que le ratio des fréquences relatives RATIO_FREQR est plus élevé [2] pour le barycentre des noms d'action en *-age* que pour les deux autres barycentres. Contrairement à nos hypothèses, les voisins du barycentres des noms d'action en *-age* ont significativement plus de marqueurs de catégories dans *Wikipedia2018* (NB_CAT_W18), et le pourcentage de voisins faisant l'objet d'une page dans *Wikipedia2018* (PAGE_W18) est plus faible pour le barycentre des noms d'action en *-age* que pour celui des

1. Nous calculons la significativité à l'aide d'une ANOVA et d'un test de Tukey post-hoc, avec un niveau de confiance fixé à 0.95 pour les critères continus (RATIO_FREQR, NB_CAT_W18, NB_DOM_T, NB_DOM_G, NB_SYN, NB_DEF_T et NB_DEF_G). La significavité des critères catégoriels (LST, PAGE_W18 et NSS) est évaluée à l'aide du χ^2 de Pearson et d'une analyse des résidus normalisés.

2. Notons la présence de valeurs aberrantes pour le barycentre des noms d'action en *-age*, absentes des autres voisinages, dues à notre choix de recoder la fréquence relative des noms d'action. Le résultat est néanmoins significatif.

noms d'action en *-ion*. Ce dernier constat va cependant dans le sens des observations faites dans le chapitre 11 concernant le caractère inapproprié de ce critère, les résultats n'étant par ailleurs pas significatifs pour le barycentre des noms d'action en *-age*. Notons que le critère de la présence parmi les noms sous-spécifié (NSS) n'est pas significatif.

La plus grande technicité du barycentre des noms d'action en *-age* émerge donc sur la base de l'examen quantitatif des voisins, fondé sur les critères empiriques et sur l'intuition des locuteurs. Les critères permettent ainsi de discriminer les voisinages que nous venons d'analyser. Intéressons-nous maintenant à la sensibilité des modèles distributionnels, avec comme point de départ non pas les suffixes, mais la technicité des noms d'action.

12.2 Vers une classe de noms techniques

Dans cette section, nous nous libérons de la contrainte formelle et nous nous basons sur le seul trait sémantique de la technicité. Nous nous inspirons en cela de l'approche utilisée dans la partie IV sur l'agentivité : le critère de technicité nous permet d'identifier des classes sémantiques de noms au degré de technicité variable, que nous pouvons analyser par le biais d'une représentation vectorielle globale.

Nous présentons dans ce qui suit la constitution de barycentres de noms techniques et non techniques (section 12.2.1), puis nous en présentons les principales caractéristiques (section 12.2.2). Enfin, nous comparons les barycentres construits sur des bases formelles et sémantiques afin d'évaluer leur intersection (section 12.2.3).

12.2.1 Représentation vectorielle de la technicité

Nous cherchons à analyser les noms d'action en fonction de leur technicité, en nous libérant de la contrainte formelle du suffixe. Nous souhaitons à ce titre envisager les noms d'action au travers de sous-groupes sémantiques différenciés par leur degré variable de technicité, que l'on puisse représenter par des barycentres. L'analyse comparative des voisinages de ces barycentres nous permet alors d'étudier dans quelle mesure les sous-groupes définis par les degrés de technicité se distinguent sur le plan distributionnel.

À l'image des barycentres précédemment construits, nous devons dans un premier temps sélection les amorces qui représenteront les classes sémantiques de noms techniques. En l'absence d'un marqueur formel, il nous faut choisir un autre critère afin de circonscrire ces groupes, ici le degré de technicité. Plusieurs critères sont ainsi à notre disposition. S'ils offrent une bonne cou-

verture, les critères empiriques sont multiples, et ne proposent pas de score qui aggrège l'ensemble des informations annotées. De ce fait, nous choisissons plutôt de nous baser sur le score de technicité perçue. Ce score se caractérise par une couverture assez limitée, puisque seuls 287 noms ont fait l'objet d'une annotation, mais il offre une évaluation directe de la technicité des noms d'action.

Puisque le score de technicité perçue se présente sur une échelle de 0 (non technique) à 2 (très technique), la question du nombre de groupes et de leur définition se pose. Nous pourrions partir d'une opposition binaire entre les noms non techniques (score de 0) et les noms techniques indépendamment de leur degré de technicité (score de 1 ou 2), ou choisir de distinguer les trois niveaux de technicité, en considérant que le choix du score 1 ou 2 est significatif pour les locuteurs. Pour avoir un aperçu le plus complet possible, nous faisons le choix de constituer 4 classes présentées dans le tableau 12.2.

Classe	Description	Technicité perçue	Effectif
Tech0	Noms non techniques	0	198
Tech1	Noms moyennement techniques	1	68
Tech2	Noms très techniques	2	21
Tech12	Noms techniques	1 ou 2	89

TABLE 12.2 – Description des classes de noms techniques

Nous faisons le choix de constituer un quatrième groupe, en plus des groupes définis par les trois niveaux de l'annotation, afin de pouvoir analyser les noms techniques par opposition aux noms non techniques, indépendamment de leur degré de technicité.

Le détail de la répartition des procédés dérivationnels à l'origine des amorces pour chaque classe est donné dans le tableau 12.3.

	-age	*-ion*	*-ment*	conv	autre
Tech0	19	94	40	19	26
Tech1	19	38	6	2	3
Tech12	28	47	7	2	5
Tech2	9	9	1	0	2

TABLE 12.3 – Constructions morphologiques des amorces utilisées pour construire les barycentres des classes de noms non techniques Tech0, et techniques Tech1, Tech12 et Tech2

Pour constituer les barycentres, nous utilisons le même modèle distributionnel que celui utilisé dans le chapitre 10, à savoir la concaténation de

cinq modèles entraînés sur le corpus *Wikipedia2018* préalablement lemmatisé, avec les mêmes paramètres par défaut. Nous calculons les barycentres des quatre groupes de noms Tech0, Tech1, Tech2 et Tech12 suivant la formule 5.1, et nous récupérons les 100 plus proches voisins de chaque barycentre. Nous donnons dans le tableau 12.4 les dix plus proches voisins des quatre barycentres.

Tech0	Tech1	Tech12	Tech2
manipulation	désinfection	désinfection	soufflage
simplification	récupération	récupération	soudure
généralisation	utilisation	filtration	usinage
utilisation	filtration	utilisation	distillation
dégradation	traitement	soudure	forgeage
assimilation	fixation	compactage	soudage
acceptation	vitrification	fixation	étirage
perception	compactage	broyage	broyage
mutation	manipulation	brasage	polissage
restriction	réutilisation	pulvérisation	brasage

TABLE 12.4 – Dix plus proches voisins des barycentres des noms d'action techniques et non techniques

Nous analysons dans la section suivante les 100 plus proches voisins de chaque barycentre.

12.2.2 Caractérisation des barycentres

Nous nous penchons sur l'analyse comparative des barycentres Tech0, Tech1, Tech2 et Tech12 afin de faire émerger les différences entre les quatre classes représentées. Il s'agit tout d'abord de voir dans quelle mesure les barycentres diffèrent ou convergent, puis de les caractériser. Cette caractérisation est approchée tant sur le plan de la technicité que de propriétés morphosémantiques, par le biais de l'examen des 100 plus proches voisins.

Un premier indice concernant les différences de technicité des voisinages des barycentres est donné par l'observation du recouvrement entre les voisins d'un barycentre donné et les amorces des autres barycentres (tableau 12.5).

Le tableau 12.5 ébauche une première distinction en termes de technicité des voisinages des barycentres. En effet, on constate qu'on ne trouve aucune amorce technique (quel que soit leur degré de technicité) dans les voisins du barycentre non technique Tech0. De même, on ne trouve aucune amorce plus technique que celles utilisées dans le voisinage du barycentre

Amorces \ Voisins	Tech0	Tech1	Tech2	Tech12
Tech0	**8**	5	2	4
Tech1	0	**5**	5	6
Tech2	0	0	**2**	1
Tech12	0	5	7	**7**

TABLE 12.5 – Recouvrement entre amorces et voisins des barycentres Tech0, Tech1, Tech2 et Tech12

des noms moyennement techniques : on trouve des noms moins techniques (amorces Tech0) et des amorces correspondant au barycentre (Tech1) mais pas d'amorces Tech2. Il semble donc que les voisinages (à l'exception de celui de Tech12, mais qui agrègent Tech1 et Tech2) ne peuvent dépasser – du moins partiellement – leur degré de technicité.

Cela pourrait indiquer que les barycentres captent, du moins partiellement, la notion de technicité, et qu'ils se caractérisent bien par des degrés de technicité différents. Mais les effectifs très déséquilibrés amènent à fortement relativiser ce constat puisque l'on a peu de noms très techniques, et beaucoup de noms non techniques. L'impact de ce déséquilibre semble être conforté par la présence d'amorces non techniques parmi les voisinages des barycentres Tech1 (5), Tech12 (4) et Tech2 (2) d'une part, et la présence d'amorces moyennement techniques parmi les voisins du barycentre Tech2 (5).

Pour confirmer la distinction entre les différents barycentres, nous examinons le recouvrement entre voisins des barycentres dans le tableau 12.6.

	Tech0	Tech1	Tech2	Tech12
Tech0	-	20	4	14
Tech1	20	-	39	81
Tech2	4	39	-	57
Tech12	14	81	57	-

TABLE 12.6 – Nombre de voisins partagés par les barycentres Tech0, Tech1, Tech2 et Tech12

Le tableau 12.6 montre que les barycentres partagent davantage de voisins à mesure que la technicité de leurs amorces augmente, à l'exception de Tech0. Ainsi, les barycentres Tech1 et Tech2 sont-ils plus proches l'un de l'autre qu'ils ne le sont de Tech0.

On constate par ailleurs une variation de 19% entre les voisinages de Tech1 et Tech12, contre 43% pour Tech2 et Tech12. Les barycentres Tech1 et

Tech12 partagent donc plus de voisins que Tech2 et Tech12. Cela signifie que le barycentre Tech12 est plus similaire à Tech1 que Tech2. Cette plus forte proximité de Tech1 et Tech12 est possiblement due aux différences en termes de représentation de la technicité. En effet, 76% des 89 amorces de Tech12 sont des amorces ayant servi à construire Tech1, contre 24% pour Tech2.

Sur un plan plus qualitatif, l'examen des 100 plus proches voisins des barycentres des noms techniques (et donc construits sur des bases sémantiques) montre d'emblée une différence par rapport aux barycentres construits sur des bases formelles. Là où il y avait une très forte homogénéité formelle pour les barycentres des noms d'action en *-age*, *-ion* et *-ment*, cette homogénéité est moindre dans le cas des barycentres des noms techniques. Nous dressons un constat plus précis des propriétés morphologiques des voisins des noms d'action en fonction de leur technicité en analysant leur type morphologique dans le tableau 12.7.

	suffixé	convert	composé	complexe	indéterminé	simple
Tech0	86	9	1	4	0	0
Tech1	96	4	0	0	0	0
Tech12	90	5	2	1	1	1
Tech2	88	3	6	1	1	1

TABLE 12.7 – Type morphologique des voisins des barycentres de noms techniques

Le tableau 12.7 montre tout d'abord que les différences entre les quatre barycentres semblent relativement minimes. Les différences les plus notables s'observent dans la comparaison des barycentres Tech0 et Tech2, le premier présentant un nombre plus important de noms converts et de noms complexes que le second, ce dernier présentant pour sa part un nombre plus important de noms composés.

Le voisinage du barycentre des noms non techniques (Tech0) présente des types morphologiques divers, incluant des noms suffixés (*manipulation*, *traitement*), des noms converts (*rejet*, *saignée*, *saisie*), des noms complexes non construits (*rigueur*, *coercition*, *redondance*, *violence*) et un nom composé (*mastectomie*). Outre les noms suffixés, le barycentre Tech0 semble favoriser de façon plus marquée que les autres barycentres les noms converts et les noms complexes non construits. L'examen de ces noms converts (40) suggère le caractère peu technique de ces noms, à l'exception de *saignée*. On peut faire l'hypothèse que la présence de ces noms est liée à la grande variabilité sémantique de l'*output* de la conversion, qui permet de construire des noms très divers, allant des noms d'agent aux noms d'action.

(40) rejet, analyse, approche, surcharge, saignée, démarche, saisie, pratique, contrainte

Les noms complexes non construits ne sont globalement pas eux-mêmes des noms techniques, puisque trois d'entre eux ne sont pas des noms d'action mais des noms de qualité ou de propriété (*rigueur*, *redondance*, *violence*), seul *coercition* dénotant une action – qui ne répond au demeurant pas à nos critères de nom d'action technique.

On constate que la présence de noms converts et de noms complexes non construits dans les voisinages des barycentres diminue à mesure que la technicité des amorces augmente. Cette diminution se fait au profit des noms composés, que l'on retrouve principalement dans le voisinage du barycentre des noms très techniques (Tech2). Ces derniers dénotent des processus industriels ou chimiques complexes, dont le degré de technicité est élevé. La présence de ce type morphologique dans le voisinage des noms techniques s'explique par l'utilisation importante de la composition – notamment néoclassique – dans le développement du lexique savant (Lasserre 2016), souvent utilisé pour désigner des procédés scientifiques complexes.

(41) électrolyse, hydrométallurgie, pyrolyse, galvanoplastie, pyrométallurgie, photolithographie

Si les voisins converts (*découpe*, *recuit*, *synthèse*) et simple (*cuisson*) du barycentre Tech2 semblent tendre vers une relative technicité, la technicité des voisins complexes non construits (*autoclave*) et morphologiquement indéterminé [3] (*technique*) ne peut être établie sur la base de notre définition puisqu'il ne s'agit pas de noms d'action.

Le voisinage du barycentre des noms moyennement techniques (Tech1) montre une très forte préférence pour les noms suffixés (96%). On retrouve parmi les quatre converts présents des noms relativement techniques (*recuit* ; *synthèse* et *saignée*), et un nom moins technique (*saisie*).

Si le voisinage de Tech12 contient un nombre intermédiaire de converts, leur examen montre que plusieurs sont partagés avec le voisinage de Tech1 comme *recuit*, *saignée* ou *découpe*. Les voisins composés et complexes non construits correspondent à ceux observés dans Tech2 (*pyrolyse*, *électrolyse*, *autoclave*). On retrouve le même voisin simple (*cuisson*) et le même voisin indéterminé (*technique*).

Notons que si l'on retrouve dans des proportions similaires des noms suffixés dans les voisinages des quatre barycentres étudiés, le nombre de suffixés varie néanmoins légèrement avec le degré de technicité des amorces. Le décompte par suffixe pour chaque barycentre est présenté dans le tableau 12.8.

3. L'indétermination se joue entre nom simple et conversion d'adjectif.

	-age	*-ion*	*-ment*	*-ure*	*-ité*	*-(a/e)nce*	*-eur*	*-esse*	*-erie*	*-ique*	*-ange*
Tech0	0	67	5	2	6	4	1	1	0	0	0
Tech1	28	62	3	1	1	0	0	0	0	0	1
Tech12	35	49	4	1	1	0	0	0	0	0	0
Tech2	54	29	1	2	0	0	0	0	1	1	0

TABLE 12.8 – Suffixe des voisins suffixés des barycentres de noms techniques

Le tableau 12.8 montre d'une part qu'un plus grand nombre de suffixes sont impliqués dans la construction des voisins du barycentre des noms non techniques (Tech0) que ceux des noms techniques (Tech1, Tech12 et Tech2). Notons qu'une partie de ces suffixes sont liés à des procédés qui ne forment pas des noms d'action, à l'image de *-(a/e)nce*, *-ité*, *-eur* et *-esse*, qui forment des noms de qualité ou de propriété (42). Cela conforte le constat posé précédemment d'une technicité plus faible des voisins du barycentre des noms non techniques (Tech0) par rapport aux autres barycentres.

(42) nécessité, rigidité, uniformité, complexité, finalité, non-conformité, tolérance, persistance, défaillance, dégénérescence, faiblesse, lenteur

D'autre part, le tableau 12.8 met en lumière la variation des effectifs de noms en *-age*, *-ion* et *-ment* à mesure que la technicité des amorces augmente. On constate en effet que le nombre de noms suffixés en *-age* augmente à mesure que la technicité augmente, et que le nombre de noms suffixés en *-ion* diminue. On observe aussi une diminution du nombre de noms suffixés en *-ment*, mais cette diminution semble moins significative au regard des faibles effectifs. Cela confirme l'hypothèse d'une corrélation entre degré de technicité et suffixation, allant dans le sens d'une plus grande technicité des noms d'action en *-age* et d'une plus faible technicité des noms d'action en *-ion* et *-ment*.

En plus des suffixations produisant des noms de qualité présentées précédemment pour Tech0, on retrouve aussi d'autres suffixes parmi les voisins des barycentres Tech1, Tech12 et Tech2. Ainsi, le suffixe *-ité* est représenté pour Tech1 et Tech12 par *traçabilité* (dont l'actionnalité et la technicité peuvent cependant se discuter). Le suffixe *-ure* est présent pour les trois barycentres avec *soudure*, et avec *argenture* pour Tech2, tous deux présentant un certain degré de technicité. Notons la présence du suffixe *-ange* avec *vidange* pour Tech1 (qui va dans le sens d'une technicité relative), et de *-erie* et *-ique* pour Tech2, avec *tuyauterie* et *électrolytique*. Si ces deux voisins ne sont pas des noms d'action (le premier étant un nom d'artefact et le second un adjectif), ils sont associés à une certaine technicité.

Pour évaluer de façon plus objective encore la technicité de ces voisins,

nous reprenons les différents scores qui sont à notre disposition.

Le score de technicité perçue souffre une nouvelle fois d'une couverture faible : 8% des voisins pour Tech0, 10% pour Tech1, 11% pour Tech12 et 9% pour Tech2 été annotés par les locuteurs. Ces voisins sont présentés en (43) par ordre de proximité décroissant, en (43a) pour Tech0, (43b) pour Tech1, (43c) pour Tech12, et (43d) pour Tech2. Les scores de technicité perçue sont donnés entre parenthèses.

(43) a. simplification (0), utilisation (0), acceptation (0), modification (0), dispersion (0), justification (0), usure (0), aggravation (0)

b. utilisation (0), vitrification (1), humidification (0), simplification (0), lubrification (1), quantification (1), optimisation (1), cristallisation (1), identification (0), clouage (0)

c. vitrification (1), utilisation (0), humidification (0), lubrification (1), simplification (0), quantification (1), optimisation (1), clouage (0), carbonisation (1), cristallisation (1), filage (2)

d. filage (2), carbonisation (1), assemblage (0), vinification (2), vulcanisation (1), lubrification (1), vitrification (1), utilisation (0), rabotage (1)

Bien que calculés sur de petits échantillons, les scores de technicité perçue des voisins des quatre barycentres tendent à confirmer une plus forte technicité de Tech2 (valeur moyenne de 1) et une technicité nulle de Tech0 (valeur moyenne de 0), les deux autres barycentres affichant des scores moyens intermédiaires (valeur moyenne de 0.5 pour Tech1, et de 0.73 pour Tech12).

Le recouvrement est un peu plus conséquent avec les critères annotés automatiquement dans la phase initiale (respectivement 40%, 51%, 54% et 57% pour les voisinages des barycentres Tech0, Tech1, Tech12 et Tech2). Une nouvelle phase d'annotation n'est cependant pas ici pertinente en l'état puisque tous les voisins ne sont pas des noms d'action, et ne peuvent donc pas être évalués au regard de notre définition et de nos critères. Une annotation des voisins pertinents mériterait à terme d'être menée. Nous donnons néamoins dans le tableau 12.9 les valeurs moyennes des critères pour les voisins annotés des barycentres des noms techniques Tech0, Tech1, Tech12 et Tech2.

Les scores moyens présentés dans le tableau 12.9 tendent à indiquer que les voisinages sont de plus en plus techniques. On observe ainsi globalement que le ratio des fréquences relatives (RATIO_FREQR) et la proportion de noms faisant l'objet d'une entrée dans le corpus *Wikipedia2018* (PAGE_W18) augmentent à mesure que l'on passe des barycentres Tech0 à Tech2, alors que le nombre de marqueurs de domaines (NB_DOM_G et NB_DOM_T), de définitions (NB_DEF_G et NB_DEF_T), et de synonymes (NB_SYN) diminue,

Critères	Tech0	Tech1	Tech12	Tech2
Ratio_freqr	0.81	3.28	3.72	8.14e+05
Nb_cat_w18	0.43	1.14	1.39	1.9
Nb_dom_t	5.6	3.73	3.24	2.75
Nb_dom_g	1.73	1.06	0.91	0.68
Lst (%)	32.5	11.8	9.26	3.51
Page_w18 (%)	82,5	86.2	88.8	90
Nb_syn	26.5	6.48	6.67	5.4
Nb_def_t	10.38	6.49	5.98	4.65
Nb_def_g	4.38	2.67	2.33	1.97
Nss (%)	5	0	0	0

Table 12.9 – Valeurs moyennes des critères des voisins des barycentres des noms d'action techniques Tech0, Tech1, Tech12 et Tech2

tout comme la proportion de noms dans le Lst et parmi les noms sous-spécifiés (Nss). Seul le nombre de catégories ne suit pas nos hypothèses, et augmente à mesure que la technicité des amorces des barycentres augmente.

Globalement, les différences affichées par les critères en fonction des suffixes sont toutes significatives sauf Nb_dom_t, Nss et Page_w18. Les valeurs pour Lst sont significatives uniquement pour Tech0 et Tech2. Pour tous les autres critères, la différence est toujours significative entre Tech0 et Tech2 d'une part, et Tech0 et Tech12 d'autre part. Dans le cas de Nb_cat_w18, la différence est également significative pour Tech1 et Tech2, et pour Nb_syn et Nb_def_g, la différence entre Tech0 et Tech1 est elle aussi significative. Cela constitue un argument supplémentaire en faveur de la validation de ces critères.

Ces résultats montrent que l'opposition entre technicité nulle, avec le barycentre Tech0, et technicité maximale, avec Tech2, est particulièrement marquée sur le plan distributionnel, et qu'elle se traduit de façon significative au niveau des critères de technicité. *A contrario*, les niveaux intermédiaires de technicité sont moins marqués. Ces constats sont cependant issus de données limitées, et ne sont donc à ce stade qu'indicatif.

En l'absence d'une description quantitative complète, et pour approfondir l'examen des barycentres des noms techniques, nous en proposons une analyse plus qualitative. Nous examinons à ce titre les voisins des barycentres à l'aune de leur type ontologique, défini comme la catégorie ontologique référentielle dénotée par le nom (Huyghe 2015), à l'image des noms de propriété, d'émotion, d'objets, d'humain, d'événement, etc. En faisant l'hypothèse que certains types ontologiques sont intrinsèquement plus techniques (noms d'action relevant des sciences ou de l'industrie) que d'autres (noms de propriété,

noms abstraits), la comparaison des types ontologiques représentés dans les différents voisinages nous permet d'appuyer la thèse d'une plus grande technicité du barycentre Tech2 et d'une technicité moindre du barycentre Tech0.

En effet, comme évoqué précédemment, on trouve parmi les voisins de Tech0 des noms qui ne sont pas des noms d'action, à l'image des noms de qualités, qui ne peuvent donc être caractérisés en termes de technicité (*tolérance*, *persistance*, *faiblesse*, *lenteur*, *rigidité*, *uniformité*, *redondance*). Parmi les voisins qui sont des noms d'action, on retrouve des processus cognitifs (*perception*, *compréhension*, *appréhension*, *appropriation*) ainsi que des noms généraux voire sous-spécifiés (*analyse*, *approche*, *fonctionnement*, *démarche*, *amélioration*, *ajustement*, *réduction*, *limitation*, *action*, *réaction*). On trouve enfin de façon plus surprenante – mais marginale – des noms techniques issus du domaine scientifique (*contamination*, *saignée*, *mastectomie*).

La présence de noms liés aux sciences est encore plus marquée dans le voisinage du barycentre Tech1 (médecine (44a) ou chimie (44b)), et implique pour tous un certain degré de technicité.

(44) a. désinfection, traitement, décontamination, cicatrisation, stérilisation, détartrage, intubation

b. vitrification, pasteurisation, polymérisation, dessication, cristallisation, réticulation, phytoremédiation, lixiviation

Le voisinage du barycentre des noms moyennement techniques contient aussi des noms relatifs à l'industrie et à l'artisanat, tels que *soudure*, *meulage*, *usinage*, et *extrusion*. On retrouve uniquement *traçabilité* comme nom de propriété, et quelques noms non techniques (*standardisation*, *application*, *récupération*).

On retrouve parmi les voisins du barycentre Tech12, qui partage 81 voisins avec le barycentre Tech1, les mêmes catégories ontologiques à savoir le médical (*désinfection*), la chimie (*vitrification*) et l'industrie (*usinage*). Les 19 voisins spécifiques au barycentre Tech12 tendent à être un peu plus associés à l'industrie et à l'artisanat (*soudage*, *étuvage*, *forgeage*, *pyrolyse*, *électrolyse*, *carbonisation*, *lyophilisation*) que les voisins spécifiques à Tech1, qui sont surtout liés aux sciences et à la médecine (*dénaturation*, *contamination*, *stérilisation*, *intubation*, *réticulation*, *ponction*) et assez peu à des activités manuelles (*vidange*, *élagage*).

Les noms d'action relevant des sciences deviennent marginaux dans le voisinage du barycentre des noms très techniques Tech2 (*désinfection*), au profit de procédés industriels ou artisanaux (*cardage*, *rabotage*, *galvanisation*, *argenture*, *pétrissage*, *pyrométallurgie*, *galvanoplastie*, *pyrolyse*, *vinification*, *chromage*, *laminage*), qui semblent caractéristiques de ce voisinage.

Globalement, les voisinages deviennent de plus en plus techniques à mesure que les amorces deviennent techniques. Notons que le degré de technicité tel qu'il s'exprime dans les modèles distributionnels semble être fortement lié au domaine ontologique des noms. Ainsi, le degré intermédiaire de technicité semble être orienté vers les sciences, et plus précisément la médecine et la chimie, alors que le degré le plus élevé renvoie davantage à l'industrie. Si l'on croise ces résultats avec le lien établi entre technicité et suffixation, cela confirme l'association préférentielle entre le suffixe *-age* et l'industrie d'une part, et le suffixe *-ion* et les sciences d'autre part (Dubois 1962).

Ce résultat émerge d'une approche à la fois d'une formelle et d'une sémantique des noms d'action, et contribue à expliquer la coexistence de ces suffixes, et la sélection préférentielle de l'un ou de l'autre sur un plan sémantique et ontologique.

12.2.3 Comparaison avec les barycentres des suffixes

Nous venons de voir que les barycentres des noms techniques captent la dimension technique telle que nous la décrivons, et que les suffixes *-age*, *-ion* et *-ment* présentent des degrés de technicité distincts. Réciproquement, nous nous demandons dans quelle mesure les barycentres des noms d'action en *-age*, *-ion* et *-ment* coincident avec les barycentres des noms techniques, et par conséquent dans quelle mesure la technicité explique les distinctions que l'on observait dans le chapitre 10.

Pour répondre à cette question, nous comparons les barycentres constitués sur une base formelle (les suffixes) et ceux constitués sur une base sémantique (la technicité des amorces) en évaluant les voisins partagés. Nous faisons l'hypothèse que si les barycentres des noms d'action en *-age*, *-ion* et *-ment* sont bien conditionnés par la technicité, indépendamment du suffixe, ils partageront un nombre important de voisins avec les barycentres de noms techniques. Conséquemment, nous faisons l'hypothèse que le barycentre des noms en *-age* partagera un nombre plus important de voisins avec le barycentre des noms très techniques Tech2 qu'avec celui des noms non techniques Tech0, et *a contrario* que le barycentre des noms d'action en *-ion* partagera un plus grand nombre de voisins avec le barycentre des noms d'action non technique Tech0 qu'avec celui des noms très techniques Tech2. Nous donnons les taux de recouvrement des différents barycentres dans le tableau 12.10.

Les résultats du tableau 12.10 vont dans le sens de notre hypothèse. En effet, on constate que le nombre de voisins partagés par les barycentres des noms techniques avec le barycentre des noms en *-age* augmente significativement (p-value < 0.01) à mesure que la technicité des amorces techniques augmente (passant de 2 pour Tech0 à 53 pour Tech2). *A contrario*, le nombre

	-age	*-ion*	*-ment*
Tech0	2	**66**	12
Tech1	32	35	8
Tech12	41	24	6
Tech2	**53**	5	1

TABLE 12.10 – Nombre de voisins partagés par les barycentres des noms d'action techniques Tech0, Tech1, Tech12 et Tech2, et les barycentres des noms d'action suffixés en *-age*, *-ion* et *-ment*

de voisins partagés diminue pour le barycentre des noms en *-ion*, passant de 66 pour Tech0 à 5 pour Tech2. On observe aussi une diminution du nombre de voisins partagés par le barycentre des noms d'action en *-ment* à mesure que la technicité augmente, ce qui semble indiquer une faible technicité des noms en *-ment*. On remarquera cependant que ce barycentre est relativement distant des différents barycentres des noms techniques, le nombre de voisins qu'il partage avec les barycentres de noms techniques étant globalement faiblee.

Ces résultats suggèrent que les barycentres construits sur des bases formelles coincident partiellement avec les barycentres construits sur des bases sémantiques. Plus précisément, le barycentre du suffixe *-age* semble partager une certaine proximité sémantique avec le barycentre des noms techniques (Tech2). Le barycentre des noms d'action en *-ion* semble quant à lui relativement proche du barycentre des noms non techniques (Tech0). Nous avons vu précédemment que *-ment* présentait un degré intermédiaire de technicité, mais cela ne se traduit pas au niveau de la comparaison des barycentres. Le nombre de voisins partagés par le barycentre des noms d'action en *-ment* avec les autres barycentres suggère que le barycentre du suffixe *-ment* est plus proche du barycentre des noms non techniques Tech0 (12 voisins partagés) que des barycentres des noms techniques Tech1, Tech12 ou Tech2 (respectivement 8, 6 et 1 voisins partagés). Cette proximité avec le barycentres des noms non techniques semble moindre comparée à celle observée pour le barycentre des noms d'action en *-ion* : le barycentre des noms d'action en *-ment* se caractérise ainsi par un niveau de technicité (très) faible.

Ces résultats soutiennent l'hypothèse d'un lien distributionnel entre technicité et suffixation. Comme le montre le tableau 12.3, la proportion de noms en *-ion* diminue peu (on passe de 48% pour Tech0 à 43% pour Tech2), mais celle de noms en *-age* augmente beaucoup (de 10% pour Tech0 à 42% pour Tech2).

Nous avons exploré dans ce chapitre la pertinence d'une distinction des

noms relativement à leur technicité dans les modèles distributionnels. La comparaison de barycentres caractérisés par des degrés de technicité distincts nous a permis de mettre en avant l'existence de différences distributionnelles sur la base du trait sémantique de la technicité, et d'identifier, sur le plan morphologique et ontologique, des propriétés des noms d'action techniques. Ce chapitre complète l'étude présentée dans la chapitre 10 dans la mesure où les noms d'action y sont considérés non pas sur une base formelle, mais sur une base sémantique.

Plus largement, cette partie a permis d'explorer la différenciation sémantique des noms d'action en *-age*, *-ion* et *-ment* à l'aide de l'association de méthodes distributionnelles et empiriques. Cette étude nous a amené à proposer une définition et une évaluation – à la fois introspective et quantitative – de la technicité appliquée aux noms d'action, mais qu'il serait envisageable d'appliquer à d'autres types de noms, et notamment les noms d'action. Nous avons montré, par le biais de la mise en place d'une évaluation quantitative de la technicité dans le chapitre 11, que cette discrimination était morpho-sémantiquement corrélée, dans la mesure où les suffixes présentent des spécificités en termes de technicité. Plus précisément, cette étude confirme à grande échelle et de façon quantitative la préférence du suffixe *-age* pour les prédicats techniques relevant du domaine ontologique de l'industrie, là où le suffixe *-ion* penche pour le domaine scientifique, moins technique au vu de nos critères.

Bien que nous n'ayons pas creusé plus l'approche – d'où le fait que nous n'en rendions pas compte ici – le clustering des noms d'action en *-age*, *-ion* et *-ment* confirme la tendance des espaces distributionnels d'une part à regrouper les noms en *-age*, et d'autre part à regrouper les noms présentant un degré plus élevé de technicité. Cela corrobore donc, sous un autre angle, nos résultats.

Conclusion et perspectives

Nous avons dans cette thèse exploré plusieurs façons d'utiliser les modèles distributionnels pour analyser sémantiquement des dérivés morphologiques, au travers de l'étude plus spécifique des noms déverbaux en *-eur*, *-euse*, *-rice* d'une part, et en *-age*, *-ion* et *-ment* d'autre part.

Une première approche explorée consiste en l'opérationnalisation de l'hypothèse linguistique d'une plus grande proximité sémantique de certains membres de familles dérivationnelles. Par l'évaluation de la proximité distributionnelle des membres des familles dérivationnelles deux à deux, nous avons pu montrer à grande échelle que le verbe tend à être plus proche du nom d'action que du nom d'agent. Cette méthode simple repose sur la traduction de l'hypothèse linguistique en des termes mathématiques, que les modèles distributionnels nous permettent ensuite de valider empiriquement. L'analyse globale offre donc une méthode de validation efficace à faible coût. L'analyse locale rend quant à elle compte des phénomènes à l'œuvre dans le corpus, tels que la polysémie, la lexicalisation, ou la fréquence, confirmant que les modèles distributionnels sont un bon outil de représentation des usages en corpus.

Une autre approche explorée repose sur l'utilisation des modèles distributionnels comme outil d'exploration pour la caractérisation de groupes de mots dont on fait l'hypothèse d'une homogénéité sémantique. Nous pouvons accéder aux spécificités sémantiques de ces groupes de mots à l'aide d'une représentation unifiée dans les espaces vectoriels, le barycentre, que l'on décrit au travers des propriétés de ses voisins distributionnels. Nous avons notamment utilisé cette approche pour comparer les noms en *-eur*, *-euse* et *-rice*, afin de faire émerger les différences axiologiques qui existent entre les noms en *-euse* et *-rice* les premiers étant porteurs d'une connotation dépréciative absente des seconds. Nous avons aussi étudié la catégorie lexicale des noms d'agent à l'aide de ses représentants prototypiques. Cela nous a permis de confirmer heuristiquement la diversité morphologique des noms d'agent d'une part, mais aussi de faire émerger leurs spécificités morphosémantiques en fonction du type (référentiel et ontologique) des agents dénotés. Les deux études illustrent la possibilité d'utiliser les modèles distributionnels à différentes granularités, et avec différents degrés de contrôle et d'analyse des amorces et des voisins. Elles montrent la finesse et la diversité des phénomènes que l'on peut étudier, et illustrent le fort degré d'adaptabilité de la méthode.

Enfin, nous avons aussi montré qu'il était possible d'intégrer les modèles distributionnels dans un dispositif plus large, en amont en tant qu'outil d'exploration puis en aval pour confirmer les résultats. Cette approche des espaces vectoriels est mise en œuvre par l'analyse comparative des noms d'action déverbaux en *-age*, *-ion* et *-ment*. La comparaison de leurs profils distribution-

nels a permis de confirmer l'existence d'une spécificité sémantique forte de ces suffixations, et d'identifier la technicité comme étant un facteur de cette différenciation, les noms en *-age* semblant plus techniques que ceux en *-ion* ou *-ment*. La description de cette différence de technicité a été approfondie à l'aide d'une étude statistique du phénomène. La combinaison des méthodes distributionnelles et statistiques offre une description du phénomène que la sémantique distributionnelle a permis de découvrir, et en propose une exploration plus complète.

Plus largement, cette thèse montre les possibilités d'adaptation des espaces vectoriels pour l'analyse linguistique, tant sur le plan descriptif qu'expérimental. Les modèles distributionnels permettent de décrire un grand nombre de régularités sémantiques. Nous les avons mises en évidence dans le cadre de l'étude des dérivés morphologiques, mais elles existent à d'autres niveaux linguistiques. La flexibilité de ces méthodes en fait un outil d'exploration riche. Elles bénéficient des apports des travaux de linguistique descriptive, par le biais de l'identification des amorces permettant le repérage des régularités, et contribuent à l'enrichissement de la description linguistique.

De nombreux aspects méthodologiques restent à explorer ou à approndir, et notamment ceux relatifs à l'entraînement des espaces vectoriels. Par exemple, nous avons fait le choix dans ce travail d'entraîner nos modèles avec l'implémentation CBOW de Word2Vec. Au vu de nos données lexicales et de leur fréquence relativement faible, l'utilisation de l'architecture Skip-gram pourrait préciser ou compléter la description sémantique des dérivés morphologiques considérés. L'approche par barycentre pourrait aussi bénéficier d'autres implémentations. Là où nous avons, dans ce travail, décidé de sonder les barycentres à l'aide de leurs 100 plus proches voisins, il serait intéressant de voir ce que d'autres méthodes de sondage peuvent nous apprendre sur les barycentres d'une part, et sur les classes considérées d'autre part. Une option serait d'examiner uniquement les voisins qui ne sont pas des amorces, et dont le sens n'a donc pas servi dans la construction de la représentation moyenne. Une autre option serait, pour un barycentre de noms suffixés donné, d'observer les noms porteurs de ce suffixe les plus distants du barycentre, dont on peut faire l'hypothèse qu'ils n'expriment pas ou peu le sens capté par le barycentre. D'autres approches compositionnelles pourraient par ailleurs être creusées, à l'image des vecteurs de différence, qu'il s'agirait d'explorer de façon plus systématique. Enfin, la question de la significativité des voisins reste posée. Comme évoqué dans le chapitre 5, nous n'avons pas de moyens empiriques d'évaluer la valeur de la présence d'un mot dans un voisinage donné, et il n'existe pas de seuil à partir duquel on peut dire que les voisins ne sont plus pertinents. Une piste de recherche à explorer serait donc le développement d'un degré de significativité des voisins, que l'on pourrait par

exemple calculer sur la base des scores cosinus et des rangs dans les voisinages. Cette réflexion se heurte cependant à la grande variabilité de la notion de pertinence des voisins en fonction de l'objet étudié.

Au-delà des considérations méthodologiques, plusieurs questions restent en suspens à l'issue de nos études.

Tout d'abord, nous estimons que l'exploration de la notion de technicité doit être approfondie et étendue à d'autres catégories que les noms d'action, comme par exemple les noms d'instrument (*balai* par opposition à *fraiseuse*). Comme évoqué dans le chapitre 11, la poursuite de la tâche d'annotation par des locuteurs semble essentielle pour conforter les observations menées. Du fait de leur inadéquation, certains critères empiriques de technicité doivent être repris, et d'autres peuvent être développés. La concrétude des noms, l'existence d'un instrument associé à l'action dénotée, ou encore le TF-IDF dans *Wikipedia2018* sont autant de pistes. Une perspective diachronique peut aussi être envisagée, à l'image de ce qui est fait dans (Bonami et Thuilier 2019). Dans la mesure où seul le suffixe *-age* est productif désormais, *-ion* n'étant plus productif (Dal *et al.* 2008), on peut se demander quelle est l'influence de la productivité sur la technicité. Autrement dit, deux hypothèses possiblement complémentaires peuvent découler du postulat que l'on ne crée plus que des noms techniques (hypothèse qui reste cependant à vérifier). On peut tout d'abord envisager la productivité du suffixe en *-age* comme une conséquence de sa technicité intrinsèque, faisant de ce suffixe le seul à même de former ces nouveaux noms. À l'inverse, on peut aussi faire l'hypothèse que le suffixe *-age* tire sa technicité du fait qu'il est le seul suffixe productif et donc disponible pour la formation de nouveaux noms d'action. Enfin, nous souhaitons développer un score de technicité unique, qui agrège les informations des critères et du score de technicité perçue, pour permettre une évaluation plus directe des noms d'action, sans pour autant perdre d'information du fait d'une importance variable des différentes dimensions de la technicité, ou de la variation inter-locuteurs.

Par ailleurs, à l'image de ce qui a été fait pour la classe lexicale des noms d'agent dans le chapitre 9, nous aimerions explorer l'organisation de la classe des noms d'action, et notamment la typologie présentée par Huyghe et Marín (2007) ou Balvet *et al.* (2011) en termes d'achèvement, d'accomplissement et d'événements. Une première exploration, basée sur les données issues de la ressource Nomage (Balvet *et al.* 2011) a montré des résultats prometteurs, dans la mesure où les classes qui partagent le plus de traits définitoires sont rapprochées dans l'espace distributionnel. L'étude mériterait cependant d'être reprise et étendue, sur la base d'une quantité plus importante de données lexicales et avec un soin tout particulier porté sur leur sélection.

Concernant les noms en *-eur*, *-euse* et *-rice*, deux points méritent d'être

creusés. Tout d'abord, il y a la différenciation des noms en *-euse* et *-rice* relativement à la lemmatisation du corpus. Nous avons en effet signalé que les différences axiologiques observées avaient été faites à partir d'un nombre relativement faible de noms féminins correctement lemmatisés. Cela pousse donc à s'interroger sur la représentativité de ces noms. Leur non-lemmatisation implique des spécificités qui distinguent ces noms des autres noms féminins.

Par ailleurs, nous avons constaté dans la partie IV des différences de comportement sur le plan distributionnel entre noms d'agent et noms d'instrument, mais qui n'ont pas pu être systématisées faute d'une annotation plus large des noms d'agent et d'instrument. Il serait intéressant de comparer, à partir de leurs barycentres, les classes de noms d'agent et de noms d'instrument monosémiques afin de définir leurs propriétés sémantiques distinctives. On peut notamment se demander si la différence relève uniquement de la présence ou absence du trait humain, comme certaines définitions le suggèrent, ou si d'autres propriétés sont en jeu. On peut par ailleurs se demander si les propriétés des noms d'instrument diffèrent en fonction du suffixe à partir duquel ils sont formés, qu'il s'agisse des suffixes *-eur*, *-euse* et *-rice*, ou plus largement d'autres suffixes de noms d'instrument comme *-oire*.

Bibliographie

Aleksandrova, A. (2013). *Noms humains de phase : problèmes de classifications ontologiques et linguistiques*. Thèse de doctorat, Université de Strasbourg, Strasbourg.

Alexiadou, A. et Schäfer, F. (2010). On the syntax of episodic vs. dispositional -*er* nominals. *In* Alexiadou, A. et Rathert, M., éditeurs : *The syntax of nominalizations across languages and frameworks*, pages 9–38. Mouton de Gruyter, Berlin.

Amiot, D. et Dal, G. (2007). Integrating combining forms into a lexeme-based morphology, u. *In Fifth Mediterranean Meeting on Morphology (MMM5)*.

Amiot, D. et Dal, G. (2008). La composition néoclassique en français et l'ordre des constituants. *La composition dans une perspective typologique. Arras : Artois Presses Université*, pages 89–113.

Anscombre, J.-C. (2001). A propos des mécanismes sémantiques de formation de certains noms d'agent en français et en espagnol. *Langages*, 143:28–48.

Anscombre, J.-C. (2003). L'agent ne fait pas le bonheur : agentivité et aspectualité dans certains noms d'agent en espagnol et en français. *Thélème, Revista Complutense de Estudios Franceses*, pages 11–27.

Antoniak, M. et Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Aronoff, M. et Lindsay, M. (2016). Competition and the lexicon. *In Livelli di analisi e fenomeni di interfaccia. Atti del XLVII congresso internazionale della Società di Linguistica Italiana*, pages 39–52.

Avraham, O. et Goldberg, Y. (2017). The interplay of semantics and morphology in word embeddings. *arXiv preprint arXiv :1704.01938*.

Balvet, A., Barque, L., Condette, M. H., Haas, P., Huyghe, R., Marin, R. et 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.

Barker, C. (2008). Possessives and relational nouns. *In* von Maienborn, C., Heusinger, K. et Portner, P., éditeurs : *Semantics : an international handbook of natural language meaning*. Mouton de Gruyter, Berlin.

Baroni, M., Bernardi, R. et Zamparelli, R. (2014a). Frege in space : A program of compositional distributional semantics. *LiLT (Linguistic Issues in Language Technology)*, 9.

Baroni, M. et Bisi, S. (2004). Using cooccurrence statistics and the web to discover synonyms in a technical language. *In Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*, Lisbon.

Baroni, M., Dinu, G. et Kruszewski, G. (2014b). Don't count, predict ! a systematic comparison of context-counting vs. context-predicting semantic vectors. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 238–247.

Baroni, M. et 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)*, pages 1183–1193, Cambridge. Association for Computational Linguistics.

Benveniste, E. (1975). *Noms d'agent et noms d'action en indo-européen*. Maisonneuve, Paris.

Bernier-Colborne, G. et Drouin, P. (2016). Évaluation des modèles sémantiques distributionnels : le cas de la dérivation syntaxique. *In Proceedings of the 23rd French Conference on Natural Language Processing (TALN)*, pages 125–138.

Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing*, 8(4):243–257.

Bobkova, N. et Montermini, F. (2020). Suffix rivalry in russian : what low frequency words tell us. *In* Audring, J., Koutsoukos, N. et Manouilidou, C., éditeurs : *Rules, patterns, schemas and analogy, MMM12 Online Proceedings*, volume 12, pages 1–17.

Bojanowski, P., Grave, E., Joulin, A. et Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv :1607.04606*.

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:213–234.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. et 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)*, pages 4349–4357, Barcelone.

Bonami, O. et Boyé, G. (2019). Paradigm uniformity and the french gender system. *Perspectives on morphology. Edinburgh : Edinburgh University Press, to appear*.

Bonami, O. et Thuilier, J. (2019). A statistical approach to rivalry in lexeme formation : French -iser and -ifier. *Word Structure*, 12(1):4–41.

Botha, J. et Blunsom, P. (2014). Compositional morphology for word representations and language modelling. *In International Conference on Machine Learning*, pages 1899–1907.

Boulanger, J.-C. (1996). Les dictionnaires généraux monolingues, une voie royale pour les technolectes. *TradTerm*, 3:137–151.

Buchta, C., Kober, M., Feinerer, I. et Hornik, K. (2012). Spherical k-means clustering. *Journal of Statistical Software*, 50(10):1–22.

Bullinaria, J. A. et Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics : A computational study. *Behavior research methods*, 39(3):510–526.

Bussmann, H. et Hellinger, M. (2003). Engendering female visibility in german. *Gender across languages : The linguistic representation of women and men*, 3:141–174.

Caliskan, A., Bryson, J. J. et Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Cartoni, B., Namer, F. et Lignon, S. (2015). A cross-linguistic insight on agentive noun formation in Italian and French. *Selected Papers from the 8th Décembrettes : Morphology in Bordeaux, Carnets de Grammaire*, 22:81–98.

Cerquiglini, B. (2018). *Le Ministre est enceinte*. Le Seuil, Paris.

Chomsky, N. (1970). *Remarks on nominalization*, volume R. Jakobs and P. Rosenbaum (Eds.), Readings in English Transformational Grammar, pages 184–221. Ginn, Waltham.

Condette, M.-H., Marin, R. et Merlo, A. (2012). La structure argumentale des noms déverbaux : du corpus au lexique et du lexique au corpus. *In SHS Web of Conferences*, volume 1, pages 845–858. EDP Sciences.

Corbin, D. (1987). *Morphologie dérivationnelle et structuration du lexique*, volume 1. Max Niemeyer Verlag, Tübingen.

Corbin, D. et Corbin, P. (1991). Un traitement unifié du suffixe *-ier(e)*. *Lexique*, (10):61–145.

Croft, W. (1991). *Syntactic categories and grammatical relations : The cognitive organization of information*. University of Chicago Press, Chicago.

Cruse, D. A. (1973). Some thoughts on agentivity. *Journal of Linguistics*, 9:11–23.

Dal, G. (2003). Analogie et lexique construit : quelles preuves ? *Cahiers de grammaire*, 28:9–30.

Dal, G., Fradin, B., Grabar, N., Namer, F., Lignon, S., Plancq, C., Zweigenbaum, P. et Yvon, F. (2008). Quelques préalables au calcul de la productivité des règles constructionnelles et premiers résultats. *In Congrès Mondial de Linguistique Française*, page 142. EDP Sciences.

Dal, G., Hathout, N., Lignon, S., Namer, F. et Tanguy, L. (2018). Toile versus dictionnaires : Les nominalisations du français en *-age* et en *-ment*. *In* Neveu, F., Harmegnies, B., Hriba, L. et Prevost, S., éditeurs : *SHS Web of Conferences*, volume 46, page 08003. EDP Sciences.

Dal, G., Hathout, N. et Namer, F. (1999). Construire un lexique dérivationnel : théorie et réalisations. *Actes de la 6ème conférence sur le Traitement Automatique des Langues Naturelles (TALN1999)*, 99:115–124.

Dal, G. et Namer, F. (2015). La fréquence en morphologie : pour quels usages ? *Langages*, (1):47–68.

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

DeLancey, S. (1984). Notes on agentivity and causation. *Studies in Language*, 82:181–213.

Dendien, J. et Pierrel, J.-M. (2003). Le trésor de la langue française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatiques des Langues*, 44(2):11–37.

Devlin, J., Chang, M.-W., Lee, K. et Toutanova, K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Drouin, P. (2003). Term Extraction Using Non-technical Corpora as a Point of Leverage. *Terminology*, 9(1):99–115.

Drouin, P. (2010). *Extracting a Bilingual Transdisciplinary Scientific Lexicon*, volume eLexicography in the 21st Century : New Challenges, New Applications, pages 43–53.

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*. Larousse, Paris.

Dubois, J. et Dubois-Charlier, F. (1999). *La dérivation suffixale en français*. Paris : Nathan.

El Khamissy, R. (2016). Les verbes causatifs dans les textes scientifiques : essai de typologie. *Thélème. Revista Complutense de Estudios Franceses*, 31(1):55–79.

Erk, K. et Padó, S. (2008). A structured vector space model for word meaning in context. *In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 897–906, Honolulu.

Fabre, C. et Lenci, A. (2015). Distributional semantics today introduction to the special issue. *Traitement Automatique des Langues*, 56(2):7–20.

Fábregas, A. (2007). A syntactic account of affix rivalry in spanish nominalisations. *The Syntax of Nominalizations across Languages and Frameworks. Berlín : De Gruyter*.

Faruqui, M. (2016). *Diverse Context for Learning Word Representations*. Thèse de doctorat, University of Trento.

Ferret, K., Soare, E. et Villoing, F. (2010). Les noms d'événement en-age et en-ée : une différenciation fondée sur l'aspect grammatical. *2ème Congrès Mondial de Linguistique Française*, page 063.

Ferret, K. et Villoing, F. (2012). L'aspect grammatical dans les nominalisations en français : les déverbaux en-age et-ée. *Lexique*, 20:73–127.

Ferret, O. (2015). Réordonnancer des thésaurus distributionnels en combinant différents critères. *Traitement automatique des langues*, 56(2).

Fillmore, C. (1968). The case for case. *In* Bach, E. et Harms, R. T., éditeurs : *Universals in linguistic theory*, pages 1–88. Holt, Rinehart & Winston, New York.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *In* Firth, J. R., éditeur : *Studies in Linguistic Analysis*, pages 1–32. Basil Blackwell, Oxford.

Flaux, N. et Van de Velde, D. (2000). *Les noms en français : esquisse de classement*. Editions Ophrys, Paris.

Fleischman, S. (1980). *The French Suffix -age : its Genesis, Internal Growth, and Diffusion*. Ann Arbor : University Microfilms International.

Fradin, B. (2012). Les nominalisations et la lecture 'moyen'. *Lexique*, (20):125–152.

Fradin, B. (2014). La variante et le double. *In* Villoing, F., Leroy, S. et David, S., éditeurs : *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*, pages 109–147. Presses Universitaires de Paris Ouest, Paris.

Fradin, B. (2016). L'interprétation des nominalisations en N-age et N-ment en français. *In Actes du XXVIIe congrès international de linguistique et philologie romanes (Nancy, 15-20 juillet 2013)*, volume 3, page 53–66. Société de linguistique romane / Eliphi.

Fradin, B. et Kerleroux, F. (2003). Troubles with lexemes. *In* Booij, G., Cesaris, J., Scalise, S. et Ralli, A., éditeurs : *Topics in Morphology. Selected Papers from the Third Mediterranean Morphology Meeting*, pages 177–196, Barcelona. IULA-Universitat Pompeu Fabra.

Fradin, B., Montermini, F. et Plénat, M. (2009). Morphologie grammaticale et extragrammaticale. *Aperçus de morphologie du français*, pages 21–45.

Frassinelli, D., Naumann, D., Utt, J. et m Walde, S. S. (2017). Contextual characteristics of concrete and abstract words. *In IWCS 2017—12th International Conference on Computational Semantics—Short papers*.

Fuentes, A. C. (2001). Lexical behavior in academic and technical corpora : Implications for esp development. *Language Learning and Technology*, 5(3):106–129.

Gilquin, G. et Gries, S. T. (2009). Corpora and experimental methods : A state-of-the-art review. *Corpus linguistics and linguistic theory*, 5(1):1–26.

Godard, D. et Jayez, J. (1993). Types nominaux et anaphores : le cas des objets et des événements. *In* De Mulder, W., Tasmowski-De Ryck, L. et Vetters, C., éditeurs : *Anaphores temporelles et (in-)coherence, Cahiers Chronos*, volume 1, pages 41–58. Rodopi, Amsterdam.

Goldberg, Y. et Levy, O. (2014). Word2vec Explained : Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method. *arXiv preprint arXiv :1402.3722*.

Gonen, H., Kementchedjhieva, Y. et Goldberg, Y. (2019). How does grammatical gender affect noun representations in gender-marking languages ? *arXiv preprint arXiv :1910.14161*.

Grimm, S. (2011). Semantics of case. *Morphology*, 21(3-4):515–544.

Grimshaw, J. (1990). *Argument structure.* the MIT Press.

Gruber, J. S. (1967). Look and see. *Linguistics*, 43:937–947.

Gupta, A., Boleda, G., Baroni, M. et Padó, S. (2015). Distributional vectors encode referential attributes. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 12–21.

Guzmá n Naranjo, M. et Bonami, O. (2020). Distributional assessment of derivational semantics. Communication orale présentée à SLE 2020.

Haas, P., Huyghe, R. et Marin, R. (2008). Du verbe au nom : calques et décalages aspectuels. *In Congrès Mondial de Linguistique Française (CMLF)*, pages 2015–2065, Paris.

Habert, B., Naulleau, E. et Nazarenko, A. (1996). Symbolic word clustering for medium size corpora. *In Proceedings of the 16th conference on Computational Linguistics (COLING)*, pages 490–495, Copenhagen.

Halliday, M. A. et Hasan, R. (1976). *Cohesion in English.* Longman, London.

Han, J., Pei, J. et Kamber, M. (2011). *Data mining : concepts and techniques.* Morgan Kaufmann Publishers, Burlington.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Haspelmath, M. et Sims, A. (2013). *Understanding morphology.* Routledge.

Hathout, N. et Namer, F. (2014). Démonette, a french derivational morpho-semantic network. *Linguistic Issues in Language Technology*, 11(5):125–168.

Hathout, N. et Namer, F. (2016). Giving lexical resources a second life : Démonette, a multi-sourced morpho-semantic network for french. *In* Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. et Piperidis, S., éditeurs : *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1084–1091, Portorož.

Hathout, N., Namer, F. et Dal, G. (2002). An Experimental Constructional Database : the MorTAL Project. *In* Boucher, P., éditeur : *Many morphologies*, pages 178–209. Cascadilla Press, Sommerville.

Hathout, N. et Sajous, F. (2016). Wiktionnaire's wikicode glawified : a workable french machine-readable dictionary. *In* Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. et Piperidis, S., éditeurs : *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1369–1376, Portorož.

Hatier, S. (2016). *Identification et analyse linguistique du lexique scientifique transdisciplinaire. Approche outillée sur un corpus d'articles de recherche en SHS*. Thèse de doctorat, Université Grenoble Alpes.

Hellinger, M. et Bussmann, H. (2001). *Gender Across LLanguage : The Linguistic Representation of Women and Men*, volume 1. John Benjamins Publishing Company.

Herbelot, A. et Baroni, M. (2017). High-risk learning : acquiring new word vectors from tiny data. *arXiv preprint arXiv :1707.06556*.

Ho-Dac, L.-M., Miletić, A., Wauquier, M. et Fabre, C. (2020). Approches outillées pour l'étude des noms sous-spécifiés ou noms capsules. *Le Français Moderne - Revue de linguistique Française*, (1).

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

Huyghe, R. (2014). La sémantique des noms d'action : quelques repères. *Cahiers de Lexicologie*, 102:181–201.

Huyghe, R. (2015). Les typologies nominales : présentation. *Langue Française*, (1):5–27.

Huyghe, R. (2018). Noms généraux et noms sous-spécifiés : des relations à préciser. Communication lors des journées d'étude S'Caladis : *Les noms sous-spécifiés en français : du lexique au discours*. Toulouse.

Huyghe, R., Barque, L., Haas, P. et Tribout, D. (2017). The semantics of underived event nouns in french. *Italian Journal of Linguistics*, 29: 117–142.

Huyghe, R. et Marín, R. (2007). L'héritage aspectuel des noms déverbaux en français et en espagnol. *Faits de langues*, 30(1):265–273.

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

Huyghe, R. et Wauquier, M. (2020). What's in an agent ? a distributional semantics approach to agent nouns in french. *Morphology*, 30:185–218.

Huyghe, R. et Wauquier, M. (À paraître). Une étude distributionnelle des noms d'agent en *-ant*, *-eur*, *-ien*, *-ier* et *–iste*. *Verbum*.

Huyghe, R. et Wauquier, M. (soumis). Distributional semantics insights on agentive suffix rivalry in french.

Josselin-Leray, A. (2005). *Place et rôle des terminologies dans les dictionnaires généraux unilingues et bilingues : étude d'un domaine de spécialité : volcanologie*. Thèse de doctorat, Lyon 2.

Kelling, C. (2001). Agentivity and suffix selection. *In* Butt, M. et King, T. H., éditeurs : *Proceedings of the Lexical Functional Grammar '01 Conference (LFG'01)*, pages 147–162, Hong Kong.

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

Kipper Schuler, K. (2005). *VerbNet : a broad-coverage, comprehensive verb lexicon*. Thèse de doctorat, University of Pennsylvania, Philadelphia.

Kisselew, M., Rimell, L., Palmer, A. et Padó, S. (2016). Predicting the direction of derivation in english conversion. *In Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 93–98, Berlin.

Kleiber, G. (1988). Prototype, stéréotype : un air de famille ? *DRLAV. Documentation et Recherche en Linguistique Allemande Vincennes*, 38(1): 1–61.

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

Knittel, M. L. (2017). French derived nominals in *-ant* : semantic properties. Communication orale présentée à ISMo 2017.

Kocourek, R. (1982). *La langue française de la technique et de la science*. Oscar Brandstetter Verlag, Wisbaden.

Koehl, A. et Lignon, S. (2014). Property nouns with *-ité* and *-itude* : formal alternation and morphopragmatics or the sad-itude of the a*ité*$_{\text{N}}$. *Morphology*, 24(4):351–376.

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

Köper, M. et Im Walde, S. S. (2016). Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 german lemmas. *In* Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. et Piperidis, S., éditeurs : *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2595–2598, Portorož.

Laca, B. (2001). Derivation. *Language Typology and Language Universals : An International Handbook*, 2:1214–1227.

Lakoff, G. (1977). Linguistic gestalts. *In Papers from the thirteen regional meeting, Chicago Linguistic Society*, volume 13, pages 236–287, Chicago. University of Chicago.

Lakoff, G., Johnson, M. *et al.* (1999). *Philosophy in the flesh : The embodied mind and its challenge to western thought*, volume 640. Basic Books, New York.

Landauer, T. K. et Dumais, S. T. (1997). A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Lapesa, G., Kawaletz, L., Plag, I., Andreou, M., Kisselew, M. et Padó, S. (2018). Disambiguation of newly derived nominalizations in context : A distributional semantics approach. *Word Structure*, 11(3): 277–312.

Lapraye, A. (2017). Une approche statistique de la concurrence entre procédés constructionnels. la dérivation en -age et en -ment en français. Mémoire de D.E.A., Université Paris Diderot, Paris.

Lasserre, M. (2016). *De l'intrusion d'un lexique allogène : l'exemple des éléments néoclassiques*. Thèse de doctorat, Université Toulouse Jean Jaurès, Toulouse.

Lasserre, M. et Montermini, F. (2014). Pour une typologie des lexèmes construits : Entre composition, composition néoclassique et affixation. *In* Neveu, F., Blumenthal, P., Hriba, L., Gerstenberg, A., Meinschaefer, J. et Prévost, S., éditeurs : *Actes du 4 e Congrès Mondial de Linguistique Française – CMLF 2014 (SHS Web of Conferences 8)*, pages 1797–1812. ILF, Paris.

Lazaridou, A., Marelli, M., Zamparelli, R. et Baroni, M. (2013). Compositional-ly derived representations of morphologically complex words in distributional semantics. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1517–1526, Sofia.

Le Draoulec, A. et Péry-Woodley, M.-P. (2016a). « auteure », « écrivaine », suite... `http://bling.hypotheses.org/1503`. Repéré le 20 avril 2017.

Le Draoulec, A. et Péry-Woodley, M.-P. (2016b). La femme de l'écrivain. `http://bling.hypotheses.org/1405`. Repéré le 20 avril 2017.

Le Draoulec, A., Péry-Woodley, M.-P. et Rebeyrolle, J. (2014). Glissements progressifs de « sémantique ». *Le discours et la langue*, 6(1):109–126.

Legallois, D. et Grea, P. (2006). L'objectif de cet article est de... construction spécificationnelle et grammaire phraséologique. *Cahiers de praxématique*, 46:151–171.

Lemay, C., L'Homme, M.-C. et Drouin, P. (2005). Two methods for extracting "specific" single-word terms from specialized corpora : Experimentation and evaluation. *International Journal of Corpus Linguistics*, 10(2):227–255.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, 4:151–171.

Lenoble-Pinson, M. (2008). Mettre au féminin les noms de métier : résistances culturelles et sociolinguistiques. *Le français aujourd'hui*, (4):73–79.

Lerat, P. (1984). Grammaire des noms d'agent en *-ant* en français contemporain. *Cahiers de Lexicologie*, 24:23–29.

Levy, O. et Goldberg, Y. (2014a). Dependency-based word embeddings. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 302–308, Baltimore.

Levy, O. et Goldberg, Y. (2014b). Linguistic regularities in sparse and explicit word representations. *In Proceedings of the 18th Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor.

Levy, O., Goldberg, Y. et Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Lignon, S. (2000). *La suffixation en -ien : aspects sémantiques et phonologiques*. Thèse de doctorat, Université Toulouse 2, Toulouse.

Lignon, S. (2007). Les noms de spécialistes en *-iste* et en *-ien* : le chimiste perturbé ou comment le physicien se réajuste. *In* Vaxélaire, B., Sock, R., Kleiber, G. et Marsac, F., éditeurs : *Perturbations et Réajustements. Langue et langage*, pages 287–295. Publications de l'Université Marc Bloch - Strasbourg 2, Strasbourg.

Lignon, S. (2013). Les suffixations en *-iser* et en *-ifier* : vérifier les données pour vériser les hypothèses ? *In Décembrettes 7. Colloque international de Morphologie à Toulouse*, pages 119–132.

Lignon, S., Namer, F. et Villoing, F. (2014). De l'agglutination à la triangulation ou comment expliquer certaines séries morphologiques. *In SHS Web of Conferences*, volume 8, pages 1813–1835. EDP Sciences.

Lindsay, M. (2012). Rival suffixes : synonymy, competition, and the emergence of productivity. *In Mediterranean Morphology Meetings*, volume 8, pages 192–203.

Lombard, A. et Huyghe, R. (À paraître). Les néologismes en -*age* en français contemporain : héritage verbal et polysémie. *Journal of French Language Studies*.

Luong, T., Socher, R. et Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. *In Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia.

Mahlberg, M. (2005). *English general nouns : a corpus theoretical approach*, volume 20. John Benjamins Publishing, Amsterdam & Philadelphia.

Manguin, J.-L., François, J., Eufe, R., Fesenmeier, L., Ozouf, C. et Sénéchal, M. (2004). Le dictionnaire électronique des synonymes du CRISCO : un mode d'emploi à trois niveaux. *Cahiers du CRISCO*, 34.

Marcato, G. et Thüne, E.-M. (2002). Gender and female visibility in Italian. *In* Hellinger, M. et Hadumod, B., éditeurs : *Gender Across Languages : The Linguistic Representation of Women and Men*, volume 2, pages 187–217. John Benjamins Publishing Company, Amsterdam.

Marelli, M. et Baroni, M. (2015). Affixation in semantic space : Modeling morpheme meanings with compositional distributional semantics. *Psychological review*, 122(3):485.

Martin, F. (2010). The semantics of eventive suffixes in french. *In* Alexiadou, A. et Rathert, M., éditeurs : *The Semantics of Nominalizations across Languages and Frameworks*, pages 109–141. Mouton de Gruyter, Berlin/New-York.

Maurice, F. (2001). Deconstructing gender – the case of Romanian. *Gender across Languages : The linguistic representation of men and women*, 1:229–252.

Mel'čuk, I. (1994). *Cours de Morphologie Générale, 2ème Partie : Significations Morphologiques*. Les Presses Universitaires de Montréal, Montréal.

Mickus, T., Bonami, O. et Paperno, D. (2019). Distributional effects of gender contrasts across categories. *In Proceedings of the Society for Computation in Linguistics*, volume 2, pages 174–184.

Mickus, T., Paperno, D., Constant, M. et van Deemter, K. (2020). What do you mean, bert ? *In Proceedings of the Society for Computation in Linguistics 2020*, pages 235–245.

Mikolov, T., Chan, K., Corrado, G. et Dean, J. (2013a). Efficient estimation of word representations in vector space. *In Proceedings of International Conference on Learning Representations (ICLR)*, Scottsdale.

Mikolov, T., Le, Q. V. et Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv :1309.4168*.

Missud, A. (2019). Modélisation quantitative de la rivalité entre la suffixation en -age et la conversion de verbe à nom. Mémoire de D.E.A., Université Paris Nanterre, Paris.

Missud, A. et Villoing, F. (2019). French *-age* suffixation versus verb to noun conversion : quantitative approaches on surface and underlying properties. Communication orale présentée à ISMo 2019.

Missud, A. et Villoing, F. (2020). The morphology of rival *-ion*, *-age* and *-ment* selected verbal bases. *Lexique*, 26:29–52.

Mitchell, J. et Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Mudraya, O. (2006). Engineering english : a lexical frequency instructional model. *English for Specific Purposes*, 25(2):235–256.

Naccarato, C. (2019). Agentive (para) synthetic compounds in Russian : a quantitative study of rival constructions. *Morphology*, 29(1):1–30.

Namer, F. (2009). *Morphologie, Lexique et Traitement Automatique des Langues*. TIC et sciences cognitives. Hermès-Lavoisier. ISBN 978-2-7462-2363-9.

Namer, F. et Villoing, F. (2015). Sens morphologiquement construit et procédés concurrents : les noms de spécialistes en *–logue* et *–logiste*. *Revue de Sémantique et de Pragmatique*, 35:7–26.

Nazar, R., Vilvaldi Palatresi, J. et Wanner, L. (2012). Co-occurrence graphs applied to taxonomy extraction in scientific and technical corpora. *Procesamiento del Lenguaje Natural*, 49:67–74.

New, B., Pallier, C., Brysbaert, M. et Ferrand, L. (2004). *Lexique 2* : a new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516–524.

Padó, S., Palmer, A., Kisselew, M. et Šnajder, J. (2015). Measuring semantic content to assess asymmetry in derivation. *In Workshop on Advances in Distributional Semantics*.

Padó, S., Snajder, J. et Zeller, B. D. (2013). Derivational smoothing for syntactic distributional semantics. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 731–735, Sofia.

Partee, B. H. et Borschev, V. (2003). Genitives, relational nouns, and argument-modifier ambiguity. *In* Lang, E., Maienborn, C. et

Fabricius-Hansen, C., éditeurs : *Modifying Adjuncts*, volume 4, pages 67–112. Mouton de Gruyter, Berlin.

Pennington, J., Socher, R. et Manning, C. (2014). Glove : Global vectors for word representation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. et Zettlemoyer, L. (2018). Deep contextualized word representations. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Pierrejean, B. (2020). *Qualitative evaluation of word embeddings : investigating the instability in neural-based models.* Thèse de doctorat, University Toulouse Jean Jaurès, Toulouse.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Rappaport Hovav, M. et Levin, B. (1992). *-Er* nominals : implications for the theory of argument structure. *In* Stowell, T. et Wehrli, E., éditeurs : *Syntax and semantics*, volume 26, pages 127–153. Academic Press, New York.

Rastier, F. (2008). Doxa et sémantique de corpus. *Langages*, (2):54–68.

Rehurek, R. et Sojka, P. (2010). Software framework for topic modelling with large corpora. *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta.

Roché, M. (1997). Briard, bougeoir et camionneur : dérivés aberrants, dérivés possibles. *In D. Corbin, B. Fradin, B. Habert, F. Kerleroux, et M. Plénat, éds., Mots possibles et mots existants. 1res rencontres du forum de morphologie*, volume 1, pages 241–250.

Roché, M. (2003). Catégorisation et recatégorisation en morphologie dérivationnelle : le cas de la dérivation en *-ier(e)*. *In* Coll, G. et Régis, J.-P., éditeurs : *Morphosyntaxe du lexique : Catégorisation et mise en discours, Actes du Colloque de Tours, 7-8 juin 2002, Travaux Linguistique du CerLiCO*, volume 16, pages 75–92. Presses Universitaires de Rennes, Rennes.

Roché, M. (2004). Mot construit ? mot non construit ? quelques réflexions à partir des dérivés en *-ier*(e). *Verbum*, 26(2):459–480. An optional note.

Roché, M. (2009). Pour une morphologie lexicale. *Mémoires de la Société de Linguistique de Paris*, 13(Nouvelle série n ° 17):65–87.

Roché, M. (2011). Quel traitement unifié pour les dérivations en *-isme* et en *-iste* ? *In* Roché, M., Boyé, G., Hathout, N., Lignon, S. et Plénat, M., éditeurs : *Des unités morphologiques au lexique*, pages 69–143. Hermès, Paris.

Rosenberg, M. (2008). *La formation agentive en français : les composés [VN/A/Adv/P]N/A et les dérivés V-ant, V-eur et V-oir(e).* Thèse de doctorat, University of Stockholm, Stockholm.

Roy, I. et Soare, E. (2012). L'enquêteur, le surveillant et le détenu : les noms déverbaux de participants aux événements, lectures événementielles et structure argumentale. *Lexique*, 20:207–231.

Rychlỳ, P. et Kilgarriff, A. (2007). An efficient algorithm for building a distributional thesaurus (and other sketch engine developments). *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 41–44.

Sahlgren, M. (2006). *The Word-Space Model : Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* Thèse de doctorat, Stockholm University, Stockholm.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies*, 20:33–53.

Salton, G., Wong, A. et Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Schafroth, E. (2001). *Gender in French Structural properties, incongruences*, volume 3, pages 87–117. John Benjamins Publishing.

Schlesinger, I. M. (1989). Instruments as agents : on the nature of semantic relations. *Journal of Linguistics*, 25(1):189–210.

Schnedecker, C. et Aleksandrova, A. (2016). Les noms d'humains en *-aire* : essai de classification. *In* Neveu, F., Bergounioux, G., Côté, M.-H.and Fournier, J.-M., Hriba, L. et Prévost, S., éditeurs : *Congrès Mondial de Linguistique Française 2016.* Institut de Linguistique Française, Paris.

Schütze, H. et Pedersen, J. (1995). Information retrieval based on word senses. *In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

Siddiqui, T., Ren, X., Parameswaran, A. et Han, J. (2016). Facegist : Collective extraction of document facets in large technical corpora. *In Proceedings of the 25th ACM International on Conference on Information and Knwoledge Management*, pages 871–880, Indianapolis.

Simondon, G. (1958). *Du mode d'existence des objets techniques*. Paris : Editions Aubier-Montaigne.

Soricut, R. et Och, F. J. (2015). Unsupervised morphology induction using word embeddings. *In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 1627–1637, Denver.

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

Tanguy, L., Sajous, F. et Hathout, N. (2015). Évaluation sur mesure de modèles distributionnels sur un corpus spécialisé : comparaison des approches par contextes syntaxiques et par fenêtres graphiques. *Traitement Automatique des Langues*, 56(2):103–127.

Tissier, J., Gravier, C. et Habrard, A. (2017). Dict2vec : Learning word embeddings using lexical dictionaries. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 254–263.

Tribout, D. (2010). *Les conversions de nom à verbe et de verbe à nom en français*. Thèse de doctorat, Université Paris Diderot, Paris.

Tribout, D., Barque, L., Haas, P. et Huyghe, R. (2014). De la simplicité en morphologie. *In* Neveu, F., Blumenthal, P., Hriba, L., Gerstenberg, A., Meinschaefer, J. et Prevost, S., éditeurs : *SHS Web of Conferences*, volume 8, pages 1879–1890. EDP Sciences.

Turney, P. D. et Pantel, P. (2010). From frequency to meaning : Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Urieli, A. (2013). *Robust French Syntax Analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Thèse de doctorat, Université Toulouse Jean Jaurès, Toulouse.

Uth, M. (2010). The rivalry of french *-ment* and *-age* from a diachronic perspective. *In* Alexiadou, A. et Rathert, M., éditeurs : *The Semantics of Eventive Suffixes in French*, pages 215–244. Mouton de Gruyter, Berlin/New-York.

Van Valin, R. et LaPolla, R. (1997). *Syntax : structure, meaning and function*. Cambridge University Press, Cambridge, MA.

Varvara, R., Lapesa, G. et Padó, S. (2016). Quantifying regularity in morphological processes : an ongoing study on nominalization in German. *In ESSLLI DSALT Workshop : Distributional Semantics and Semantic Theory*, Bolzane.

Vendler, Z. (1957). Verbs and times. *The philosophical review*, pages 143–160.

Verhoeven, B., Daelemans, W. et 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)*, pages 121–125, Pretoria.

Vikner, C. et Jensen, P. A. (2002). A semantic analysis of the English genitive : interaction of lexical and formal semantics. *Studia Linguistica*, 56(2):191–226.

Villoing, F. et Fiammetta, N. (2014). Composition néoclassique en *-logue* et en *-logiste* : les noms *-logue* sont-ils encore des noms de spécialistes ? *Verbum*, 2(34):213–231.

Wagner, C., Garcia, D., Jadidi, M. et 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)*, pages 454–463, Oxford.

Wang, X., Lo, D., Jiang, J., Zhang, L. et Mei, H. (2009). Extracting paraphrases of technical terms from noisy parallel software corpora. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP) Conference Short Papers*, pages 197–200, Singapore.

Wauquier, M. (2018). Analyse des noms agentifs dans les espaces vectoriels distributionnels. *In Actes de la Conférence RJC 2018, conjointe à TALN et CORIA 2018*, Rennes.

Wauquier, M., Fabre, C. et Hathout, N. (2018). Différenciation sémantique de dérivés morphologiques à l'aide de critères distributionnels. *In* Neveu, F., Harmegnies, B., Hriba, L. et Prevost, S., éditeurs : *SHS Web of Conferences*, volume 46, pages 1156–1170. EDP Sciences.

Wauquier, M., Fabre, C. et Hathout, N. (À paraître). Différenciation des noms d'action dérivés : le facteur de technicité étudié en corpus. *In* Frérot, C. et Pecman, M., éditeurs : *Des corpus numériques à la modélisation linguistique en langues de spécialité*. UGA Editions, Grenoble.

Wauquier, M., Hathout, N. et Fabre, C. (2020a). Contributions of distributional semantics to the semantic study of French morphologically

derived agent nouns. *In* Audring, J., Koutsoukos, N. et Manouilidou, C., éditeurs : *Rules, patterns, schemas and analogy, MMM12 Online Proceedings*, volume 12, pages 111–121.

Wauquier, M., Hathout, N. et Fabre, C. (2020b). Semantic discrimination of technicality in French nominalizations. *Zeitschrift für Wortbildung / Journal of Word Formation*, 4(2).

Winther, A. (1975). Note sur les formations déverbales en *-eur* et *-ant*. *Cahiers de lexicologie*, 26:35–54.

Zeller, B., Šnajder, J. et Padó, S. (2013). DErivBase : Inducing and evaluating a derivational morphology resource for german. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, volume 1, pages 1201–1211, Sofia.

Zeller, B. D., Padó, S. et Snajder, J. (2014). Towards semantic validation of a derivational lexicon. *In Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 1728–1739, Dublin.

Annexes

Annexe A

Guide d'annotation de la technicité des noms d'action

Cette tâche d'annotation consiste en l'attribution d'un score de technicité, allant de 0 (technicité nulle) à 5 (technicité maximale) à une série de noms d'action. Nous définissons un nom d'action technique de la façon suivante :

> *Un nom d'action technique est un nom peu transparent pour un public non initié, dénotant une action précise complexe, dont la réalisation et la connaissance nécessite un savoir acquis et qui est spécifique à un domaine particulier. Les noms d'action techniques appartiennent typiquement aux domaines de l'industrie, de l'agriculture et de l'artisanat.*

L'annotation se fait *a priori*, hors de tout contexte, sur la base de l'intuition de l'annotateur, et par rapport à la définition précédente. Plus un mot répond à cette définition, plus il a un score élevé. Un nom ne répondant pas à cette définition se voit annoté 0. Si l'annotateur ne connaît pas un mot, il peut en chercher le sens dans des dictionnaires en ligne (TLFi, Wiktionnaire...) ou des occurrences dans un concordancier de type NoSketchEngine [1]. Le recours à un moteur de recherche d'images est aussi admis. Si le nom reste inconnu à l'issue de sa recherche, l'annotateur l'indiquera en lui attribuant le score de -1.

Pour aider à la décision, l'annotateur peut se servir des critères suivants pour préciser son jugement :

— **Agentivité :** Les noms ayant un fort degré de technicité dénotent des actions réalisées de façon volontaire et consciente par un agent humain (dont le nom sera parfois, mais pas nécessairement, morphologiquement lié au nom d'action). Cela est illustré par des noms comme reliure, ac-

1. Accessible à l'adresse www.clarins.si/noske.

tion réalisée par un agent humain (*relieur*). *A contrario*, l'action désignée par un nom comme disparition n'implique pas obligatoirement un agent pour son déroulement.

— **Complexité :** Les noms ayant un fort degré de technicité dénotent des actions complexes, dont la réalisation demande un certain degré d'expertise, acquis par le biais d'une formation spécifique. L'existence d'un métier associée à cette action est donc un indice de sa technicité. L'annotateur pourra ainsi s'interroger sur la possibilité pour tout un chacun, y compris soi-même, de réaliser cette action. Cela opposera par exemple exorcisme et danse : on dit de quelqu'un qu'il danse dès qu'il bouge de façon consciente au rythme d'un musique, peu importe son niveau de formation en danse, alors que tout le monde ne peut pas pratiquer un exorcisme. L'action dénotée nécessite un apprentissage particulier, et le nom est parfois, mais pas nécessairement, peu transparent. La familiarité du nom d'action aux yeux de l'annotateur pourra être un autre indice de son degré de technicité. L'action désignée par le nom ébreuillage (qui dénote le fait de vider un poisson de ses boyaux) repose sur des compétences acquises, que l'on utilise pas dans la vie quotidienne car elles relèvent d'une expertise particulière. Le nom est par ailleurs a priori peu transparent pour une personne qui ne serait pas familière de ce type d'activités. A contrario, l'action dénotée par le nom *fléchage* (action de jalonner un chemin de flèches pour indiquer la direction à suivre) ne nécessite pas un savoir acquis à l'issu d'un apprentissage particulier et semble un peu plus transparent.

— **Outils :** Les noms ayant un fort degré de technicité dénotent des actions qui nécessiteront, pour leur réalisation, des outils dont le nom sera parfois, mais pas nécessairement, morphologiquement lié au nom d'action, et démontrant eux-mêmes un certain degré de complexité, dont la manipulation nécessite une formation. On peut ainsi prendre comme exemple le nom alésage, dont l'action implique l'utilisation d'un alésoir, ou le nom linogravure, dont la réalisation nécessite l'usage de gouges. La complexité de l'outil pourra orienter sur le degré de technicité du nom broyage qui implique au niveau industriel des machines très complexes. A contrario, l'action dénotée par un nom comme agrandissement ne nécessite pas de façon intrinsèque un outil en particulier, et l'outil en question variera en fonction de l'objet (matériel ou immatériel) que l'on cherchera à agrandir.

— **Spécificité :** Un nom d'action ayant un fort degré de technicité est considéré comme spécifique puisqu'il dénote une action précise. Les noms ayant un fort degré de technicité auront donc tendance à être peu

polysémiques. Ils auront à ce titre un nombre plus limité de définitions, de synonymes et d'hyponymes que des noms non techniques. Le nom drayage illustre bien l'idée de spécificité, puisqu'il dénote une opération précise, relative au tannage, et ne présente qu'une seule définition, et un nombre de synonymes faible, à savoir dérayage dans Wiktionnaire. A contrario, le nom habillage relève de très nombreux domaines (boucherie, cuisine, horticulture, marine, etc.), désignant dans chacun de ces domaines des actions différentes. On compte ainsi de nombreux synonymes pour habillage : habillement, présentation, ornement, décoration, disposition, etc. Notons que l'annotation doit se faire sur la base de la technicité du nom et pas celle du domaine dans lequel il s'intègre. À l'image de aimance, un nom spécifique à un domaine n'est pas nécessairement technique : bien qu'il soit issu d'un domaine a priori assez technique (celui de la psychologie) et qu'il dénote un phénomène bien précis, impliquant un agent, et peu familier des non-experts, l'action elle-même (mode d'aimer propre à l'enfant) n'est pas technique, et n'implique pas d'outil ni un apprentissage particulier. Enfin, le fait que l'action dénotée par le nom se déroule dans un cadre spécifique, et non pas dans la vie quotidienne, par une personne bien particulière et pas par n'importe qui (reprenant l'idée de complexité) est un indice de la technicité du nom. L'annotateur pourra ainsi s'interroger sur la facilité d'intégrer le nom dans une phrase, et sur les contraintes sémantiques et ontologiques sur cela implique. Cela opposera par exemple danse et couplage : le premier pourra facilement s'intégrer dans des phrases génériques comme j'aime la danse ou la danse me permet de me changer les idées, alors que l'on pourra difficilement faire la même chose pour couplage.

Pour être technique (score de 1 à 5), un nom d'action doit nécessairement remplir au moins un des critères, mais pas nécessairement l'ensemble des critères. Un nom ne remplissant aucun critère se voit attribuer un score de 0. Les différents critères permettent de nuancer le degré de technicité des noms. Un nom validant un ou des critères n'est cependant pas nécessairement technique, à l'image de *aimance*.

Cette annotation repose autant que possible sur l'intuition de l'annotateur. Dans le cas de noms polysémiques, comme *presse*, l'annotateur doit annoter la technicité de l'acception qu'il juge principale. Les auteurs de ce guide jugent que l'acception principale de *presse* est liée au monde du journalisme, et celle de *blanchissage* liée au nettoyage du linge, et considèrent donc *presse* et *blanchissage* comme des noms respectivement non et faiblement techniques. Mais si le sens premier de ces noms pour es annotateurs corres-

pond à l'action de presser quelque chose dans un pressoir pour presse, et aux malversations financières pour blanchissage, ils doivent annoter la technicité du nom sur la base de ces sens spécifiques. Soulignons enfin que l'annotation doit se faire sur la base du nom et exclusivement du nom ; l'annotateur ne doit pas se référer aux mots qui lui seraient proches notamment par la forme. En effet, *accouvage* et *couver* n'ont par exemple pas le même degré de technicité.

Nous illustrons nos propos ainsi que différents cas de figure à l'aide des exemples suivants.

Le nom ***calandrage*** est jugé très technique (5) selon notre définition. Il s'agit d'un nom peu transparent, dénotant une action précise et complexe, spécifique à un domaine, celui de l'industrie, et nécessitant un savoir-faire acquis. Il nécessite un agent qui réalise l'action, à l'aide d'un outil morphologiquement lié au nom d'action (calandre).

Le mot ***ciselure*** est jugé technique (4) selon nos critères. Il relève en effet du domaine de l'artisanat et de l'industrie, et est le fruit d'une action réalisée intentionnellement par un agent (ciseleur) à l'aide d'outils spécifiques (ciselets). L'action nécessite un savoir-faire acquis. Il n'a qu'un seul synonyme (gravure), et un nombre limité de définitions. Il est cependant considéré comme moins technique que calandrage du fait de sa plus grande transparence et de la complexité moindre de l'action dénotée.

Le mot ***tondaison*** est jugé moyennement technique (3) selon nos critères. Il relève en effet du domaine de l'agriculture et de l'artisanat, et est le fruit d'une action réalisée intentionnellement par un agent à l'aide d'outils spécifiques (tondeuse). L'action nécessite un certain savoir-faire, acquis, pour être bien réalisée. Le nom n'a qu'un seul synonyme (tonte) et un nombre limité de définitions. L'action dénotée semble cependant moins technique que celle dénotée par les noms ciselure et calandrage, et le nom est transparent.

Les noms ***baguage*** et ***inventorisation*** nous semblent faiblement techniques (2) en cela que les compétences nécessaires à leur réalisation semblent d'une complexité relativement faible, surtout en comparaison de tondaison. L'outillage et les connaissances nécessaires pour la réalisation de l'action qu'ils dénotent sont à la portée du plus grand nombre..

Les noms ***crayonnage*** et ***danse*** sont jugés comme des noms très faiblement techniques (1). S'ils impliquent tous les deux un agent pour la réalisation de l'action qu'ils dénotent, celle-ci reste peu complexe, à la portée de tout le monde, et ne nécessitant aucune expertise particulière, quand bien même elle peut être faite de manière professionnelle. Par ailleurs, elles n'impliquent pas d'outils, ou alors des outils très génériques qui ne sont pas spécifiquement dédiés à cette action en particulier.

Le degré de technicité du mot ***déclin*** est jugé nul (0). Il s'agit d'un mot

relativement générique, non spécifique au domaine technique, et il ne dénote pas une action réalisée intentionnellement par un agent, mais plutôt une action subie. Il présente de nombreux synonymes (abaissement, affaiblissement, baisse, chute, décroissance, diminution, etc.) et un nombre assez important de définitions. Sa réalisation ne nécessite pas un outil ou un savoir-faire particulier.