Département de Sciences du Langage
Master Ergonomie Cognitive et Ingénierie Linguistique

Mémoire de Master 1

Indices distributionnels pour la comparaison sémantique de dérivés morphologiques

Marine Wauquier
Sous la direction de Cécile Fabre et de Nabil Hathout
2015 - 2016

Remerciements

Je tiens à remercier du fond du cœur les personnes qui m'ont aidée et accompagnée dans la réalisation de ce mémoire.

Je remercie dans un premier temps mes encadrants de mémoire, Cécile Fabre et Nabil Hathout, sans qui ce travail n'aurait pas été possible. Leurs conseils avisés, leur disponibilité et leur accompagnement m'ont été des plus précieux.

Je souhaite aussi remercier le reste de l'équipe enseignante du master LITL, Mai Ho-Dac et Ludovic Tanguy, dont les enseignements et encouragements ont su m'aider à avancer.

Je remercie aussi Michel Roché pour le temps qu'il m'a accordé lors d'un entretien et d'échanges qui ont fortement nourri ma réflexion.

Je tiens par ailleurs à remercier Cécile Fabre, Nabil Hathout, Franck Floricic, Christelle Lesselingue, Edith Galy et Fiammetta Namer pour leur contribution à Lexeur, et Franck Sajous pour la préparation du corpus et pour les données qu'il m'a fournies.

Enfin, je souhaite remercier mes camarades de classe, Amélie, Aurore, Céline, Justine et Laura, ainsi que mes proches pour leur soutien inestimable.

Table des matières

Introduction

La production langagière se caractérise par une grande variabilité : il est possible de formuler de plusieurs façons une même idée. La paraphrase, ou globalement tout processus de reformulation, traduit la multiplicité des formes permettant de convoyer un contenu unique. Ainsi, deux segments tels que présentés en (1) diffèrent dans la forme, mais véhiculent un même contenu. De même, deux mots peuvent être similaires au niveau du sens, mais différents au niveau de la forme, tels que les synonymes présentés en (2). Ces segments et ces mots, bien que différents, sont liés entre eux sur le plan sémantique. Ils partagent un même sens. On parle donc de proximité sémantique. Le traitement de la variabilité de l'expression d'une même information et donc la notion de proximité sémantique sont des enjeux essentiels du Traitement Automatique des Langues, et ce dans le cadre de nombreuses applications.

(1) *Le président a été mortellement blessé par les tirs d'une arme à feu.*
Le décès du président fait suite aux blessures infligées par une arme à feu.

(2) *Mort*
Décès

Dans le cadre de ce mémoire, nous nous intéressons à la proximité sémantique des dérivés morphologiques, et plus précisément au lien entretenu entre un verbe et ses noms d'agent et d'action dérivés. L'objectif est de tester l'hypothèse de Roché (2009) selon laquelle le sens d'une base et de son nom d'action dérivé serait identique : l'utilisation de l'une ou de l'autre forme serait simplement la conséquence de la construction syntaxique choisie par le locuteur. Ainsi, utiliser le nom d'action *protection* au lieu du verbe *protéger* répondrait seulement à un besoin syntaxique dans le cadre des phrases telles que (3).

(3) *Le gouvernement confie la protection des populations à l'armée*.
L'armée est chargée par le gouvernement de protéger les populations.

Calculer la proximité sémantique des différents termes impliqués dans la dérivation morphologique est une des pistes envisageables pour confirmer ou infirmer cette hypothèse. Si différents couples de dérivés morphologiques sont concernés par l'idée d'une simple variation catégorielle ou formelle, comme les couples verbes et noms processifs tels que *protéger* et *protection*, noms et adjectifs de relation comme *gouvernement* et *gouvernemental*, ou encore adjectifs et noms de propriété tels que *profond* et *profondeur*, le choix est fait de ne s'intéresser qu'au cas des dérivés nominaux agentifs en *-eur*, *-euse* et *-rice* (suffixes que nous engloberons sous l'étiquette *-EUR* dans la suite de ce travail), et des dérivés processifs issus d'une base (généralement) verbale de type *protecteur*, *protectrice*, *protection* et *protéger*. Une comparaison de la proximité entretenue entre le

verbe et ses différents dérivés est ainsi envisagée pour mettre au jour (ou non) une variation sémantique plus ou moins importante en fonction de la dérivation impliquée.

De nombreuses méthodes permettent déjà de calculer la proximité de différents items, mais la question se révèle plus complexe lorsque les items comparés n'appartiennent pas à la même catégorie grammaticale. Des bases de données en réseau telles que Wordnet, par exemple, ne représentent pas les liens inter-catégoriels, et il faut donc compter avec ce fossé catégoriel. Pourtant, des solutions existent, des méthodes qui ne reposent pas sur les items eux-mêmes mais sur leur environnement proche.

En effet, l'idée d'une meilleure compréhension du sens d'un mot grâce à son contexte a depuis longtemps été formulée, notamment par Firth qui écrivait : « *You shall know a word by the company it keeps* » (Firth, 1957a: 11, cité par Church et Hanks, 1990). Ainsi, s'intéresser au contexte d'apparition des lexèmes permet de comparer des éléments autrement non comparables. Le partage ou non des profils distributionnels des différents lexèmes comparés est donc une des pistes qui est privilégiée pour analyser la proximité entre les noms d'agent en *-EUR*, les noms d'action et les verbes dont ils sont issus.

Dans ce mémoire, nous cherchons à tester l'hypothèse selon laquelle les noms d'action déverbaux sont sémantiquement plus similaires aux verbes de base que ne le sont les noms d'agent déverbaux. Pour ce faire, nous exploitons des indices distributionnels que nous envisageons comme des indices de similarité sémantique. À l'aide de Word2Vec, nous enrichissons une ressource lexicale d'informations distributionnelles. Nous composons notamment des triplets nom d'agent/verbe/nom d'action afin de comparer les trois éléments qui nous intéressent. Au sein de chaque triplet, nous nous servons des profils distributionnels obtenus pour chaque élément afin de déterminer quel élément, entre le nom d'action et le nom d'agent, est le plus proche du verbe sur un plan distributionnel. Nous élargissons alors notre étude aux familles dérivationnelles complètes de chaque triplet afin de voir si ces indices distributionnels représentent bien le comportement sémantique des éléments comparés à l'échelle de la famille globale.

Un premier temps de ce travail sera consacré à l'étude des objets qui nous intéressent, afin de mieux en cerner les caractéristiques. Un deuxième temps sera dédié à l'analyse des enjeux et défis de la proximité sémantique et aux méthodes à notre disposition pour calculer la proximité de nos items. Dans un troisième temps, nous décrirons Lexeur, la ressource lexicale que nous utilisons après son enrichissement dans le cadre de cette étude. Enfin, dans un quatrième temps, nous ébaucherons une analyse distributionnelle sur la base de nos ressources pour tester l'hypothèse de Roché.

1. Les dérivés agentifs et processifs déverbaux

Puisque notre étude porte sur des dérivés morphologiques et le lien qu'ils entretiennent avec leur base, nous allons dans un premier temps nous pencher sur la notion de dérivation morphologique.

1.1. Morphologie et dérivation

La morphologie est la branche de la linguistique qui étudie les mots et les morphèmes. Le mot, appellation parfois un peu floue, est un signe linguistique composé d'une forme et d'un sens. La forme d'un mot se compose elle-même d'une dimension phonologique et d'une dimension morphosyntaxique. La morphologie affecte donc, soit séparément, soit simultanément, ces trois dimensions.

Cette discipline regroupe elle-même diverses branches, telles que la morphologie flexionnelle et la morphologie dérivationnelle, que l'on oppose de par leur sujet d'étude, à savoir la relation entre le lexème et le mot-forme, dans un cadre syntaxique d'une part (comme dans le cadre de la flexion verbale, par exemple), et la relation entre les lexèmes, dans un cadre lexical d'autre part, dans le cadre de la dérivation par exemple (Fradin, 2009; Roché, 2009).

Fradin (2009) définit le lexème comme une unité lexicale abstraite, à savoir qui ne spécifie pas la valeur de ses éventuels traits flexionnels. Le mot-forme, quant à lui, est l'instanciation du lexème qui le subsume, c'est l'unité concrète correspondant au lexème que l'on retrouve dans le discours. Dans le cadre du français par exemple, le lexème ne porte donc pas de marque de nombre (sauf si une certaine valeur de ce trait flexionnel est inhérente au lexème lui-même, à l'image de la marque du pluriel pour le lexème *vacances,* dans son acception *congés*, qui permet de le distinguer du lexème *vacance* dont le sens diffère), et apparaît sous sa forme citationnelle au singulier.

Si un certain flou demeure parfois sur les diverses appellations en concurrence pour qualifier un certain type de morphologie, qui ne serait ni flexionnelle ni grammaticale, Roché (2009) fait la distinction entre trois approches, bien qu'elles partagent un même cadre d'étude, à savoir le lexique : la morphologie lexicale, la morphologie dérivationnelle, et la morphologie constructionnelle.

1.1.1. Morphologie lexicale

Elle s'intéresse au lexique en tant qu'ensemble de lexèmes qu'il faut traiter en fonction de sa spécificité. La notion de lexique dans le cadre de la morphologie lexicale sera plus amplement développée dans la partie 3.1.1. Ce que l'on peut néanmoins retenir concernant la morphologie lexicale dans (Roché, 2009), c'est que l'organisation du lexique en paradigmes (lexicaux ou dérivationnels) est importante pour la morphologie constructionnelle puisque l'appartenance d'un mot construit, à savoir fruit d'une dérivation, à ces deux types de paradigme a un impact sur la formation de

ce mot.

1.1.2. Morphologie dérivationnelle

Elle s'intéresse quant à elle au processus de création d'un nouveau mot à partir d'un mot existant. L'opération de création lexicale accomplie par la morphologie dérivationnelle a pour but non pas de créer du sens, puisque selon Roché (2009), le contenu sémantique reste identique, mais de nommer différentes catégories. Roché (2009) qualifie ce processus de nomination. Au cœur de ce processus se trouve la dérivation, d'où il tire son nom. Parler de dérivation implique de parler de deux éléments clés : la base et le dérivé. Puisque les dérivés, en l'occurrence les noms d'agent en -*eur* et les noms d'action dans le cadre de cette étude, feront l'objet d'un développement dans la partie 1.2 et 1.3., nous nous intéresserons ici uniquement aux bases. Roché (2009) distingue trois notions différentes à mettre en parallèle pour définir ce qu'est une base : la notion de "base" (sur le plan lexical), la notion de "thème" (sur le plan lexématique, et non pas sémantique comme l'opposé de rhème) et la notion de "radical" (sur le plan phonologique).

La notion de "base" réfère au lexème qui sert de point de départ à la dérivation. C'est un objet complexe, car composé des trois éléments que sont sens, forme et catégorie, et souvent polysémique et polymorphe. Il s'agirait dans l'exemple (4) du mot *puits*. Le "thème" désigne l'une des formes phonologiques associées au lexème spécifique. Ces formes existent en tant que telles dans le lexique, indépendamment du processus de dérivation dans lequel elles sont impliquées. Ainsi, le thème */pɥiz/* utilisé en (4) existe en lui-même et n'a pas été créé pour les besoins de la dérivation. Enfin, le radical est la forme particulière prise par la base pour faire l'objet d'une dérivation, par concaténation avec un affixe. La forme peut correspondre à un des thèmes du lexème, ou elle peut être le fruit d'une adaptation, par troncation ou par interfixation, d'un de ces thèmes. Dans le cadre de (4) (exemple tiré de Roché, 2009), le radical est donc formé sur le thème */pɥiz/*.

(4) PUITS /pɥi/
PUISER *puis-* /pɥiz/
PUISATIER *puisat-* /pɥizat/

La distinction entre thème et radical n'est pas toujours clairement établie, notamment en anglais qui utilise un même mot *stem* pour parler de ces deux éléments, et pourtant elle joue un rôle non négligeable dans la description de certains processus morphologiques tels que le supplétisme. Roché (2009) donne ainsi l'exemple du paradigme dérivationnel (sur le plan fonctionnel et lexical) du lexème CHEVAL, qui contient *cavalier*, *hippique* et *équestre* : ces trois mots sont construits sur trois radicaux spécifiques issus de différentes langues, partagent le même sens et le même lien avec le lexème CHEVAL mais ils ne sont clairement pas des thèmes issus de CHEVAL. De même, dans le couple nom-verbe formé par les mots *clou* et *clouter*, on peut noter l'existence d'un seul thème, */klu/*, mais du radical *clout-* permettant la formation du verbe *clouter*, ne correspondant pas à un thème de

clou (thème du verbe *clouer*). La distinction entre radical et thème reste donc nécessaire.

Sur la base de ces distinctions, Roché (2009) décrit la dérivation comme étant une opération constructionnelle intervenant sur quatre niveaux différents : un niveau lexical concernant le lexème de base et le lexème dérivé, un niveau phonologique, où intervient le choix du thème comme radical, un niveau catégoriel où la catégorie du dérivé est définie, et un niveau sémantique pour ce qui est de la construction du sens.

1.1.3. Morphologie constructionnelle

Elle décrit précisément les trois composantes d'une opération constructionnelle : la composante formelle, la composante sémantique et la composante catégorielle. Par définition, une dérivation entraîne la modification, d'une façon ou d'une autre, du mot existant afin de créer un mot nouveau. L'opération constructionnelle modifie donc l'une ou l'autre de ces composantes : par défaut, elle modifie l'ensemble de ces composantes. Ainsi, une conversion comme le passage du nom *bleu* à l'adjectif *bleu* ne toucherait-t-elle qu'à la composante catégorielle (si tant est que l'on considère la conversion comme une opération catégorielle, et non comme une dérivation affixale (Roché, 2009)), et une suffixation en *-ette* comme dans *maison*/*maisonnette* les composantes formelle et sémantique, mais pas catégorielle.

1.1.4. Règles de construction, schèmes et analogie

La morphologie constructionnelle permet, à l'image de la morphologie dérivationnelle, de créer des nouvelles unités, dites construites, à partir d'unités existantes. Les paradigmes formés par les unités lexicales construites montrent la régularité dans la construction du lexique. Cette régularité a été observée et formalisée de différentes façons.

Si le mot complexe, ou construit, est le produit d'opérations morphologiques, Booij (2009) identifie deux approches pour l'analyser : l'approche basée sur les morphèmes, qui voit le mot construit comme la concaténation d'un morphème et d'un suffixe, et l'approche basée sur le mot, qui compare deux groupes d'items pour établir un parallèle entre leurs différences formelles et sémantiques. On peut obtenir une abstraction, à savoir une représentation, de la relation entre les deux sous-ensembles comparés grâce à une projection. Ainsi, à partir de la deuxième approche, lorsque l'on projette la relation paradigmatique qui unit les ensembles de (5), on obtient un schéma (comme l'exemple (6)) que l'on peut ensuite généraliser. Cette généralisation, que Booij (2009) nomme schème abstrait, illustré ici par l'exemple (7), permet alors de construire de nouveaux mots en unifiant le schème, à savoir en l'instanciant à l'aide du remplacement de la variable x par un mot concret. Cette approche dépend donc de l'existence de mots attestés dans un lexique.

(5) *move* *movable*
break *breakable*
attain *attainable*

(6) [[break]$_V$able]$_A$

(7) [[x]$_V$able]$_A$ ‘*that can be V’*

De par l’architecture en trois niveaux du mot mise en avant dans la section 1.1, Booij obtient la représentation (8) en trois parties, tenant compte des dimensions phonologiques, morphosyntaxiques et sémantiques du mot, regroupées entre elles par un même indice. À chaque unité de sens est associé un indice (*i* pour le mot construit final, *j* pour la base, et *k* pour l'affixe), et chaque dimension est décomposée en items élémentaires. Cette représentation nous montre que le mot est composé sur le plan phonologique de trois syllabes, /*breɪ*/, /*kə*/ et /*bl*/, les deux premières étant liées à la base (puisque portant l'indice *j*), et la dernière au suffixe, puisque portant l'indice *k*. Sur le plan morphosyntaxique, cela nous montre la nature des éléments (un nom pour le mot construit, un verbe pour la base, et le suffixe). Enfin, cette représentation nous montre de quelle manière la sémantique du mot construit est formée sur la base du sémantique de la base, ici du verbe.

(8) $\omega_i \leftrightarrow N_i \leftrightarrow$ [that can be $BREAK_j$]$_i$
/ \ \ | \
σ σ σ V_jAff_k
|\ \ | \ |
[breɪ k]$_j$[əbl]$_k$

Une fois cette représentation généralisée, Booij (2009) propose le schème tripartite abstrait sous sa forme finale tel qu’illustré par (9) dans le cadre de la suffixation adjectivale en -*able* en anglais, où le nombre de syllabes et le sens du verbe ne sont pas indiqués, puisqu’ils sont liés au verbe qui instanciera le schème.

(9) $\omega_i \leftrightarrow N_i \leftrightarrow$ [that can be $PRED_j$]$_i$
| | \
[ ]$_j$[əbl]$_k$ V_jAffk

Les Règles de construction de lexème (ou RCL) décrites par Fradin (2009) s’inscrivent aussi dans une démarche de généralisation dans le sens où elles cherchent à exprimer la régularité perceptible dans certaines séries comme dans (5). Là encore, les RCL reposent sur l’existence de mots attestés et du lien régulier qui permet de les regrouper au sein d’un même ensemble. Il s’agira par exemple de distinguer *orage* et *visage* de *lavage* et *meublage*.

La description offerte par les RCL se base là aussi sur l'existence de différents niveaux, au nombre de trois (phonologie, syntaxe et sémantisme), qui sont traduits de façon schématique sous la forme (10), forme qu'il s'agit ensuite d'instancier, de façon à expliciter les différentes caractéristiques des lexèmes. On en retrouve une illustration, tirée de (Fradin, 2009 : 94), pour les noms d'agent déverbaux en *-eur* en (11), où les trois niveaux sont donc représentés.

(10)

LEXÈME 1		LEXÈME 2
Phonologie		Phonologie
Syntaxe	↔	Syntaxe
Sémantique		Sémantique

(11)

LEXÈME 1		LEXÈME 2
(X)		(Xœʁ)
cat:v, ST-ARG<SN0, SN1>	↔	cat:n, ger:mas
(λy. λx. λe.V'(e,y,x)), x=AGT		(λV'. λx. V'(e,y,x))

Sur le plan phonologique, on retrouve ainsi pour le lexème 2 l'ajout des phonèmes /*œʁ*/ à la forme phonologique du lexème 1. Sur le plan syntaxique, on observe le passage d'un lexème appartenant à la catégorie grammaticale des verbes, et possédant donc une structure argumentale précise (ici deux arguments, SN0 et SN1), à un lexème appartenant à la catégorie des noms, et dont le genre est masculin. Enfin, sur le plan du sémantisme, on retrouve pour le lexème 1 une représentation en lambda calcul du verbe de base *V'* ainsi que de ses deux arguments syntaxiques *x* (agent) et *y* (patient) , la variable *e* représentant l'action, l'évenance. La représentation du sémantisme du lexème 2 construit à partir du lexème 1 montre l'effacement de l'argument Patient (*y*) et de l'évenance *e*. Seul est conservé l'argument Agent *x* initial, auquel est ajouté l'argument V'. Cela souligne la construction du sens du lexème 2 à partir de celui du lexème 1, même si le contenu sémantique des deux lexèmes n'est pas identique.

De par l'importance de la comparaison et de la régularité dans le phénomène de création lexicale, les RCL peuvent être comparées à une analogie réussie (Fradin, 2009). En effet, l'analogie constitue une autre façon de former des mots en se basant sur la régularité d'un phénomène (Dal, 2003). Bien que ce processus ait fait l'objet de vives critiques, et qu'il ne soit pas unanimement reconnu comme valide et productif, l'analogie permet, notamment grâce au principe de la quatrième proportionnelle de Saussure illustré en (9) (exemple tiré de (Saussure, 1916) tel que cité par Dal (2003 : 11)), de décrire et de produire des mots. Si l'analogie ne rentre pas autant que les schèmes abstraits de Booij (2009) ou les RCL décrites par Fradin (2009) dans les caractéristiques phonologiques, morphosyntaxiques ou sémantiques des lexèmes, elle permet néanmoins de combler certains lacunes lexicales.

(12) *réaction:réactionnaire = répression:x*
x = répressionnaire

Peu importe la représentation choisie, il est important, pour comprendre ce qu'ils représentent, de pouvoir caractériser de façon précise les noms d'agent, et plus particulièrement les noms agentifs en *-eur, -euse* et *-rice* qui nous intéressent.

1.2. Caractérisation des noms d'agents

Il est communément admis qu'un nom d'agent est un nom dérivé d'un verbe, moins couramment d'un autre nom, comme dans le cas de *bridgeur* qui dérive de *bridge*, par l'ajout d'un suffixe. Huyghe et Tribout (2015:3) 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 ». L'instruction du suffixe ajouté modifie le sens mais aussi la catégorie grammaticale du mot dans le cadre d'une dérivation sur base verbale, puisque l'on passe généralement d'un verbe (plus rarement d'un nom) désignant une action à un nom désignant la personne ou la chose qui réalise l'action. Le nom d'agent est donc une variante sémantique, formelle et la plupart du temps catégorielle de sa base. L'opération de dérivation permettant la formation du nom d'agent touche donc au niveau sémantique, formel et la plupart du temps catégoriel la base à partir de laquelle le nom est dérivé (hormis lorsque la base est un nom, auquel cas il n'y a pas de changement catégoriel).

Le suffixe agentif par lequel la dérivation a lieu peut être multiple et prend différentes formes : on retrouve ainsi *-ant* (bien que le cas d'une dérivation suffixale pour les noms d'agent en *-ant* ne fasse pas l'unanimité, comme en discute Ascombre (2003)), *-ien*, *-ier*, *-iste*, ou encore *-eur*. De nombreux suffixes féminins, généralement des variantes flexionnelles des suffixes agentifs masculins existants, pourraient aussi être cités : *-ienne*, *-ante*, *-ière*, mais aussi *-euse*, ou *-rice*. Leur fonctionnement étant similaire à celui de leurs équivalents masculins, nous ne nous pencherons de façon approfondie que sur le cas du nom d'agent en *-eur*. Nous reviendrons sur le cas des noms d'agent féminin en *-euse* et *-rice* dans le cadre de la section 1.2.2.

Certains noms d'agent peuvent à première vue entrer en compétition pour une même base verbale, et ce de deux façons différentes. On peut d'abord parler des cas de l'usage de deux suffixes complètement différents pour une même base, tel qu'illustré en (13).

(13) *gagner - gagnant - gagneur*
exploiter - exploitant - exploiteur

Nous ne nous pencherons pas dans cette étude sur ce cas précis, bien qu'une réelle différence sémantique entre *gagnant* et *gagneur* puisse être mise en avant. La notion de concurrence suffixale entre alors en jeu. On parle de concurrence lorsqu'il y a synonymie entre deux affixes, autant sur le plan syntaxique que sur le plan sémantique (Tuesday, 2011). On peut notamment citer les suffixes *-eur*,

-ant, *-ier*, *-iste* ou *-oir* pour la suffixation agentive, comme dans le cas de (13), le cas des suffixes *-esque* et *-ien* pour la suffixation adjectivale dans les syntagmes *électorat chiraquien* et *électorat chevénementiste* (Lignon, 2002), ou le cas des suffixes *-eur* et *-eux* en français québecois dans le couple *magouilleur*/*magouilleux*. Cette concurrence se justifie parfois par des contraintes phonologiques (Lignon, 2002), par des tendances régionalisantes, notamment dans le cas du français québecois (Lachance, 1988), mais aussi par des différences quant à la fonction morphologique et sémantique des suffixes. C'est ce que montre Lachance (1988) pour les suffixes *-eur* et *-eux* du français québecois, ces formes pouvant être divisées en six suffixes distincts sur la base de la fonction morphologique et sémantique. Les suffixes *-eur* et *-eux* se distinguent notamment par l'ajout d'un sens péjoratif dans le cas de *magouilleur* et *magouilleux*, *gagnant* et *gagneur* se distinguant quant à eux par la notion de chance pour le premier et d'effort pour le second.

Mais l'on trouve aussi l'existence de paires *a priori* concurrentes dans le cadre d'une dérivation utilisant le même suffixe, mais pas la même base, tel qu'illustré en (14). On peut en effet se demander ce qui distingue réellement *sauveur* de *sauveteur*, deux noms au sens intuitivement proche.

(14) *sauver - sauveur - sauveteur*
donner - donneur - donateur

Bien qu'ayant un fonctionnement relativement similaire, chaque suffixe se caractérise par des règles de suffixation et des instructions spécifiques. Puisque nous nous penchons dans notre étude sur le cas des noms d'agent en *-EUR*, nous allons essayer de caractériser plus précisément leur formation et les contraintes (phonologiques, catégorielles et/ou sémantiques) qui y sont liées.

1.2.1. Noms d'agents en *-eur*

De par les trois opérations morphologiques (formelle, catégorielle et sémantique) impliquées dans la formation des noms d'agent, ces derniers ont des caractéristiques syntaxiques et sémantiques spécifiques qui diffèrent plus ou moins de leur base. Nous cherchons à identifier ces caractéristiques.

1.2.1.1. Caractérisation syntaxique

Sur le plan syntaxique, nous avons vu précédemment que les noms d'agent étaient construits sur une base généralement verbale : le suffixe *-eur* permet effectivement de créer des noms d'agent déverbaux. Booij (2009 : 17) avait ainsi formalisé la déverbalisation du nom agentif anglais en *-er* sous la forme (15). Nous avons aussi vu dans la section 1.1.4 les RCL illustrées en (11) proposées par Fradin (2009) pour les noms d'agent en *-eur*.

(15) $\omega_i \leftrightarrow N_i \leftrightarrow$ [one who $PRED_j$]$_i$
| | \
[ ]$_j$[ər]$_k$ $V_j Aff_k$

Mais le suffixe *-eur* ne se limite pas à cela. D'une part, il ne produit pas que des noms d'agent, mais aussi des lexèmes n'appartenant pas à la classe des substantifs. On retrouve par exemple des adjectifs (possédant aussi une lecture agentive) tels que *songeur* ou *trompeur*, qu'ils soient le fruit d'une dérivation morphologique à part entière, avec un suffixe *-eur* sémantiquement différent du suffixe agentif *-eur*, ou le fruit d'une opération de conversion. D'autre part, et ce à l'image du préfixe *sur-* (Amiot et Dal, 2009), lui aussi à l'origine de lexèmes de diverses catégories, le suffixe *-eur* accepte des bases d'autres catégories. D'autres schèmes et d'autres RCL pourraient donc être élaborés pour décrire la formation sur base substantivale notamment.

Pourtant, s'il existe des noms d'agent en *-eur* dérivés d'une base nominale, ce n'est pas la norme. En effet, les noms d'agent construits sur une base substantivale sont généralement construits par l'ajout du suffixe *-ier* (Plénat, 2009) et non pas *-eur*. Mais l'on a parfois recours à certains phénomènes tels que « l'échangisme » suffixal pour éviter certains écueils, souvent phonologiques, comme dans le cas des mots de (16). Pour contourner une consécution gênante de phonèmes trop proches (voire identiques), l'utilisation d'un autre suffixe peut être requise. Dans le cas de (16), le suffixe agentif *-ier* est donc remplacé par le suffixe *-eur*, afin de ne pas avoir deux yods à proximité. Cela pourra par exemple expliquer la présence de certains noms d'agent en *-eur* issus de substantifs dans notre étude.

(16) *camionneur - ?camionnier*
avionneur - ?avionnier

L'autre explication à apporter sur l'existence de noms agentifs d'origine nominale est à chercher du côté de Ascombre (2001,2003) qui puise des éléments de réponse dans le mécanisme de transposition appliqué aux noms d'agent en *-eur* (Benveniste, 1974, cité dans Anscombre, 2001, 2003). En effet, la transposition ferait toujours, sur le plan sémantique, passer d'un verbe (ou plus précisément d'un groupe verbal) à un nom. Mais le groupe verbal initial, sur le plan morphologique, ne serait pas nécessairement un verbe en tant que tel : dans cette approche, la notion de verbe englobe le verbe en tant que lexème, mais aussi en tant que groupe verbal permettant d'apparenter le nom à un verbe, sous la forme *faire Nom*. Cette vision permet donc de reformuler *sauveur* sous la forme *qui sauve*, et *sauveteur* sous la forme *qui fait des sauvetages*, le groupe verbal *faire des sauvetages* devenant donc le verbe servant de base à *sauveteur*. Représenter la transposition sous la forme *qui Verbe* ou *qui fait Nom* permet ainsi de passer d'une base nominale ou verbale à un nom, sans que cela ne touche au sens. En outre, la représentation initiale du groupe verbal aurait un impact sur les propriétés syntaxiques et sémantiques du dérivé. Cela permettrait notamment d'expliquer la différence

entre les deux homonymes *travailleur* et *travailleur*, selon l'exemple proposé dans (Ascombre, 2001 : 30), où l'un des deux *travailleur* (comme dans *travailleur de nuit*) serait le fruit d'une transposition, contrairement au second *travailleur* (tel qu'on le retrouve dans l'opposition à *capitaliste*) qui serait l'héritage de l'opposition entre leur base, *travail* et *capital*.

Si les verbes sont les bases de prédilection pour la formation de noms d'agent en *-eur*, tous ne permettent pourtant pas cette opération. Fradin (2009) souligne en effet que seuls les verbes intransitifs inergatifs prennent un dérivé agentif. Sont donc exclus des bases potentielles les verbes ergatifs (aussi dits inaccusatifs) tels que les verbes de perception (s'il existe bien des noms en *-eur* formés sur la base de verbes de perception tels que *entendeur* ou *voyeur*, ces noms prennent un sens spécifique, comme nous le verrons dans la section 1.2.1.2.) , certains verbes de mouvement (certains d'entre eux produisent des noms d'agent, comme le montre Ascombre (2001), mais au sens bien spécifique, comme *coureur*), ou encore les verbes météorologiques tels que *pleuvoir* (**pleuveur*).

Fradin (2009) insiste par ailleurs sur les propriétés des arguments de la base. Le premier argument du verbe servant de base doit en effet avoir les propriétés d'un agent, qu'il soit fort (à savoir qu'il a le contrôle de l'évenance, donc de l'action et/ou de son déroulement) ou faible (à savoir qu'on puisse lui attribuer l'évenance). Cependant, l'argument doit avoir comme seules et uniques propriétés celles d'un agent : il ne peut pas être autre chose.

Mais la suffixation déverbale en *-eur* produit parfois des phénomènes problématiques. Ainsi, à l'image de (14), certaines bases verbales produisent plusieurs noms agentifs en *-eur*. Par ailleurs, certains noms d'agent comme *consommateur* ne se forment pas de la façon attendue (**consommeur*). C'est sur le plan sémantique qu'Ascombre (2001) va chercher des explications.

1.2.1.2. Caractérisation sémantique

Comme nous l'avons vu dans la section 1.1.4. avec l'exemple (11), Fradin (2009) souligne que le sens du nom d'agent est directement construit à partir du sens du verbe dont il dérive, mais qu'ils ne sont pas identiques. Du point de vue sémantique, le suffixe *-eur* fait partie de ces suffixes permettant de faire passer une unité lexicale du statut de procès au statut de propriété (ou inversement), à l'image des suffixes *-able* ou *-ant* pour les adjectifs. Ascombre (2001) distingue quatre types de propriétés, opposées deux à deux : il oppose les propriétés intrinsèques (constitutives de l'individu ou du groupe d'individus) et les propriétés extrinsèques (non constitutives de l'individu, propriétés ajoutées) d'une part, et les propriétés essentielles (partagées par tous les individus) et les propriétés accidentelles (que certains ont, mais que tout le monde ne partage pas) d'autre part. Les deux paires de propriétés peuvent se combiner ensemble : on peut ainsi avoir un item aux propriétés intrinsèques accidentelles ou un item aux propriétés extrinsèques accidentelles. Le couple (17) montre une alternance de propriétés accidentelles respectivement intrinsèque et extrinsèque. L'adjectif *maladif* indique une caractéristique constitutive de l'individu dont il est question, alors que l'adjectif *malade* dénote un aspect temporaire, l'aspect non-constitutif. Par ailleurs, expliciter des propriétés intrinsèques essentielles telles que la vision ou l'ouïe (à la fois constitutives de l'individu, et partagées par tous les

individus du groupe) n'a pas de sens, selon Anscombre (2001) de par leur caractère universel. Dès lors, les noms d'agent formés sur des bases aux propriétés intrinsèques essentielles prennent un sens dérivé, à l'image de *entendeur*, *voyeur* ou encore *coureur* : il ne s'agit pas de nommer la personne qui fait l'action dénotée (puisque, normalement, tout individu voit, entend et court, et qu'il n'est donc pas nécessaire de le préciser), mais de qualifier un comportement spécifique.

(17) *maladif - malade*

Anscombre (2003) associe cette distinction extrinsèque/intrinsèque à l'analyse de Benveniste (1975) sur l'origine des suffixes agentifs et processifs. En effet, dans son étude des suffixes indo-européens agentifs, Benveniste (1975) distingue d'une part ce qui est constitutif de l'individu, celui-ci étant caractérisé par l'action dénotée et ce qu'elle soit actualisée ou non, traduit par le suffixe *-*ter*, et ce qui est une propriété ajoutée, l'individu actualisant l'action dénotée à un moment précis, traduit par le suffixe *-*tor* (deux suffixes que l'on ne conserve en français que sous la forme du seul suffixe -*eur*).

À cette notion de propriétés, et pour expliquer certaines irrégularités, Anscombre (2001) rajoute l'idée de thème. Sur la base de la définition donnée dans le chapitre 1.1.2, il distingue, tout comme Plénat (2009), les deux thèmes verbaux pouvant servir de base pour les noms d'agent en -*eur* : le thème de présent et le thème de supin. Le thème de présent est défini comme étant le thème servant à la formation du participe présent en français. Le thème de supin est quant à lui défini comme étant le thème généralement utilisé pour la formation des noms d'action. Le nom d'agent *imprimeur* est ainsi formé sur le thème de présent, et non de supin, puisqu'il partage le même radical que le participe présent *imprimant*, et pas le radical du nom d'action *impression* (Anscombre, 2001 : 37). Outre le radical qu'ils forment, le thème de présent et le thème de supin se distinguent surtout sur le plan aspectuel. Ainsi, le thème de supin traduit un état stable, le fruit de son accomplissement : c'est une relation de type nominal. Le thème de présent traduit quant à lui une relation de type verbal, puisqu'il présente l'évènement que constitue la réalisation de l'action. Cela explique par exemple la représentation que l'on peut se faire de (14) sous la forme *sauveur* = "qui sauve ou a sauvé" (formé sur le thème de *sauvant*) et *sauveteur* = "qui fait des sauvetages" (formé sur le thème de *sauvetage*) suggérée par le mécanisme de transposition évoquée dans la section 1.1.2.1.

Enfin, Benveniste (1975) souligne un autre phénomène sémantique impliqué dans la formation des noms d'agent : la notion d'objectivité et de subjectivité. Reprise par Anscombre (2001), cette distinction 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. Au contraire, la caractérisation sera subjective si elle définit l'individu, si cela abolit son identité. Toujours pour reprendre l'exemple (14), *sauveur* serait de l'ordre de l'objectif, contrairement à *sauveteur*, dont la fonction sert à identifier l'individu.

L'association des thèmes de présent et de supin et de la distinction objectif/subjectif permet d'aboutir à quatre cas théoriques, cependant pas tous rencontrés à parts égales : la valeur processive du thème de présent invite en effet davantage à une lecture objective, et la valeur plus stative du thème de

supin à une lecture subjective. Cela n'empêche cependant pas de retrouver d'autres combinaisons, telles que, pour *sauveur*, le thème de présent associé à une lecture objective dans le cas de *il a sauvé*, mais associé à une lecture subjective dans le cas de *il est le sauveur de l'humanité*. En combinant cela avec la notion de propriétés, on en conclut donc que la lecture objective correspond à une propriété extrinsèque, et que la lecture subjective correspond à une propriété intrinsèque accidentelle.

Huyghe et Tribout (2015) reprennent ces distinctions mises en évidence par Benveniste, et ils délimitent ainsi trois groupes de noms d'agent : les noms de statuts, les noms d'agents occasionnels et les noms dispositionnels. Les noms de statuts se caractérisent 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 (18) (tirés de Huyghe et Tribout, 2015), comme l'effacement possible du déterminant, la non compatibilité avec des arguments spécifiques, la possibilité de la présence d'un complément du nom sous une forme générique et l'inscription dans des syntagmes nominaux génériques, indéfinis existentiels.

(18) *Pierre est brocanteur.*
?le déménageur de ces meubles
un carreuleur de piscine
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. Ils ne permettent en cela pas la désignation d'un référent précis et admettent, voire nécessitent, l'utilisation de compléments spécifiques, comme le montrent les exemples (19) (tirés de Huyghe et Tribout, 2015). Huyghe et Tribout remarquent que certains noms d'agent peuvent cumuler une interprétation statuaire et une interprétation occasionnelle.

(19) **Pierre est agresseur*
l'agresseur de Pierre
??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 statuaire 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. Les compléments qu'ils acceptent doivent avoir une forme générique, et le déterminant ne peut s'effacer. Enfin, l'emploi d'adjectifs de taille amène à une lecture fréquentielle ou intensive de l'action décrite. Les exemples (20) (tirés de Huyghe et Tribout, 2015) illustrent ces emplois.

(20) *Un séducteur de jeunes filles*
**Pierre est séducteur.*
??le séducteur de Sophie
Pierre est un grand séducteur.

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). Pourtant, on peut s'interroger sur la formation des noms d'agent féminins équivalents à ceux en *-eur*, et sur l'irrégularité de certaines formations.

1.2.2. Le cas des féminins *-euse* et *-rice*

Anscombre (2001) recense trois suffixes agentifs féminins : *-euse*, *-rice* et *-eresse*.

Ces suffixes sont souvent vus comme la variante féminine du suffixe *-eur* qui crée des mots masculins. Pourtant, certains couples de noms d'agent comme *pointeur*/*pointeuse* ou *lessiveur*/*lessiveuse* poussent à s'interroger sur cette équivalence. En effet, si *lessiveuse* est un nom d'agent féminin désignant à la fois une personne dont le métier est de laver du linge et un appareil servant à lessiver le linge, *lessiveur* est quasi uniquement employé pour désigner un outil utilisé dans la papetterie. A contrario, si *pointeur* permet de désigner la personne qui contrôle la présence d'employés ainsi que des outils d'artillerie, *pointeuse* n'est que très rarement employé pour désigner une personne contrôlant la présence d'employés, et quasi exclusivement pour désigner la machine permettant d'enregistrer la présence de ces ouvriers. On peut alors se demander si ces suffixes n'ont qu'une simple valeur aspectuelle.

Bien que régulière, la formation du féminin en *-euse* connaît des limitations relativement importantes puisque le suffixe *-euse* ne peut produire un nom d'agent que s'il est ajouté à une base de type verbal, sur le thème de présent. C'est pour cela que l'on pourra avoir *chanteur*/*chanteuse*, mais *professeur*/**professeuse* (puisque *professeur* n'est pas un dérivé de *professer*).

Anscombre (2001) insiste sur la distinction des propriétés objectives (occasionnelles, qui ne définissent pas l'individu) et des propriétés subjectives (qui définissent l'individu) que permet le féminin. Cette distinction passe notamment par l'usage du suffixe *-eresse* et non du suffixe *-euse* sur un thème de présent, comme dans l'opposition *chasseuse*/*chasseresse* : la *chasseuse* sera une femme qui chasse occasionnellement, alors que la *chasseresse* sera caractérisée par sa pratique de la chasse.

Enfin, notons que le suffixe *-rice* sera généralement utilisé pour des thèmes de supin, même si cette formation admet quelques exceptions.

Dal (2003) souligne que la formation de certains noms d'agent en *-rice* est obtenue par analogie, en se basant sur la formation de leurs équivalents masculins, grâce au principe de la 4e proportionnelle évoquée dans le chapitre 1.1.4, expliquant par exemple la formation de *sénatrice* à

l'image de *sénateur* sans passer par un hypothétique verbe **sénater* ou nom d'action **sénation*/**sénage*.

1.2.3. Noms d'agent versus noms d'instrument

Comme nous avons pu le voir dans la section 1.2.1.1, le suffixe *-eur* est relativement productif. En effet, s'il permet à la fois d'obtenir des noms et des adjectifs, les noms déverbaux suffixés en *-eur* sont eux-mêmes de deux types : les noms d'agent, que nous avons caractérisés dans la section 1.2.1., et les noms d'instrument, comme *aspirateur* ou *foreuse*.

1.2.3.1. Caractérisation des noms d'instrument

Huyghe et Tribout (2015:3) définissent les noms d'instrument comme étant des « nom[s] déverba[ux] qui dénote[nt] l'artefact prototypiquement utilisé pour réaliser l'action décrite par le verbe de base ». Cette définition comprend donc un aspect morphologique, puisqu'elle induit que le nom d'instrument est le fruit d'une dérivation sur la base d'un verbe, et un aspect ontologique basé sur la fonction de l'entité décrite. L'aspect morphologique est primordial dans la description des noms d'instrument puisque c'est ce qui les différencie des noms d'artefact, qui sont de simples noms d'objet. Cela en fait donc une classe à part.

Huyghe et Tribout (2015) distinguent bien la notion de rôle et la catégorisation nominale, qui reprennent toutes les deux les notions d'agent et d'instrument. En effet, les rôles thématiques d'agent et d'instrument peuvent tout à fait s'appliquer à des mots qui ne sont pas des noms d'agent ou d'instrument, quand *a contrario*, un nom d'instrument comme *couteau* peut tout à fait prendre d'autres rôles thématiques que celui d'instrument, comme le montre l'exemple (21) (tiré de Huyghe et Tribout, 2015:3). La notion de rôle n'est donc pas toujours suffisante ou pertinente pour caractériser un nom d'instrument.

(21)	*Pierre aiguise un couteau*	patient
	Pierre a posé le couteau sur la table	thème
	Pierre a coupé la corde avec un couteau	instrument

Les noms d'instrument se caractérisent selon Huyghe et Tribout (2015) par trois éléments : leur autonomie descriptive, leur absence de structure événementielle spécifique et leur fonction dénominative. En effet, les noms d'instrument sont autonomes syntaxiquement et sémantiquement. Ils peuvent s'employer seuls, sans complément, sans la présence de contexte spécifique. Les entités qu'ils dénotent ne nécessitent pas l'existence d'entités d'un autre type pour exister elles-mêmes, contrairement à certains mots comme *blancheur* ou *cicatrisation*. Les noms en *-eur* à interprétation instrumentale définissent des classes d'entités noms d'instrument permettent de définir des classes

d'entités, comme le font les noms d'objet. C'est en cela que l'on parle d'autonomie descriptive. Comme ces noms n'héritent pas de la structure argumentale du verbe dont ils dérivent, ils se retrouvent dénués de structure événementielle spécifique : ils sont généralement incompatibles avec des compléments en *de* liés à des événements spécifiques, et lorsqu'ils sont compatibles avec ces compléments, ces derniers prennent une forme générique, non occurrentielle, à l'image de l'exemple (22) (tiré de Huyghe et Tribout, 2015:6). Les noms d'instruments n'expriment pas l'événementialité propre au verbe dont ils dérivent.

(22) *un adoucisseur d'eau*
**l'adoucisseur de cette eau*

Enfin, on parle de fonction dénominative puisque les noms d'instrument servent à désigner les entités auxquelles ils font référence : ils peuvent être employés pour identifier les référents qu'ils dénotent, à l'aide de formules comme *ça s'appelle un X*.

1.2.3.2. Distinction entre agent et instrument

Outre les connaissances du monde du locuteur, il n'est pas toujours évident de distinguer les noms d'agent des noms d'instrument. Comme le montrent les résultats de la tâche d'annotation réalisée par Huyghe et Tribout (2015) sur les noms en *-eur*, *-euse* et *-rice* issus de Lexique 3, la proportion de noms d'instrument et de noms répondant à la fois aux critères de nom d'agent et de nom d'instrument n'est pas négligeable : ainsi, sur les 1547 noms en *-eur* conservés, 1 077 d'entre eux sont des noms d'agent, soit 70 % des noms , 135 sont des noms d'instrument (soit 9%), 154 répondent à la fois aux critères des deux catégories (soit 10%), et 181 ne sont pas définis (soit 11%). Sont considérés comme indéfinis les noms en *-eur* n'étant pas des noms d'agent car ne référant pas à des animés, mais ne répondant pas aux critères de nom d'instrument, par exemple, ou nécessitant l'emploi de tournures causatives pour répondre aux critères, tels que les noms *successeur* ou *détonateur.*

Huyghe et Tribout (2015) proposent un test pour distinguer le nom d'instrument du nom d'agent. Il s'agit de tester l'instrumentalité d'un nom déverbal en *-eur* grâce à deux constructions présentées en (23) et (24) (Huyghe et Tribout, 2015:4).

(23) NP_0 Vb NP_1 {avec/grâce à/à l'aide de/au moyen de} Det N_{eur}
J'aspire la poussière à l'aide d'un aspirateur.

(24) NP_0 Vb NP_1 avec Det N_{eur} ↔ NP_0 utilise Det N_{eur} pour Vb NP_1.
Je perce le mur avec une perceuse ↔ J'utilise une perceuse pour percer le mur.

À ces deux tests permettant d'affirmer le caractère instrumental ou non d'un nom peuvent être

associés des tests certifiant le caractère agentif (ou non) d'un nom. Peuvent être testés le caractère agentif d'un nom, grâce aux tournures illustrées dans les exemples (25), ainsi que le caractère dynamique des verbes de base, grâce aux tournues illustrées en (26).

(25) Det N_{eur} {a décidé/a choisi} de Vb (NP_1)
Le chanteur a décidé de chanter cette chanson.
Det N_{eur} (de NP) VB {volontairement/délibérément/consciemment/intentionnellement} (NP_1)
Le chanteur a délibérément chanté cette chanson.

(26) NP_0 est en train de Vb (NP_1)
Jean est en train de chanter cette chanson.
NP_0 vient de Vb (NP_1)
Jean vient de chanter cette chanson.

Ces tests permettent de confirmer que les noms d'agent en -*eur* répondent bien à la définition fournie au début de la section 1.2.

Outre la distance entre la base (verbale ou nominale) et son nom dérivé agentif ou instrumental, nous nous intéressons aussi à la proximité entre cette même base et son ou ses noms d'action dérivés. Nous allons donc, à l'image du nom d'agent, chercher à caractériser le nom d'action.

1.3. Caractérisation des noms d'action

La catégorie des noms d'action représente une portion majeure du lexique nominal dans la typologie établie : on dénombrerait plus de 10 000 noms d'action, selon Huyghe (2014), ce qui en ferait les noms les plus nombreux. Leur définition précise reste pourtant encore problématique, puisque la délimitation de cette classe de noms est relativement floue, car généralement basée sur l'intuition. Pourtant, des critères existent, bien que certains problèmes subsistent.

1.3.1. Définition du nom d'action

Huyghe (2014) définit la classe des noms d'action comme « l'ensemble des noms qui dénotent des actions, i.e. des situations temportelles dynamiques, causant un changement » (Huyghe, 2014:2). Pour tenter de délimiter plus précisément la classe des noms d'action, Huyghe (2014) évoque trois critères distincts : le critère morphologique, la description temporelle, et l'aspect dynamique.

D'un point de vue morphologique, le nom d'action peut être un nom simple ou construit, auquel cas le nom processif est généralement déverbal, à l'image du nom d'agent et du nom d'instrument, et construit au moyen d'une dérivation morphologique : cette construction peut être une

suffixation ou une conversion. Dans le premier cas, l'ajout d'un suffixe sur une base initiale, généralement verbale, construit ainsi une nouvelle unité dont le sens et la catégorie se retrouvent modifiés. Huyghe (2014) comptabilise ainsi près de 9393 couples dans le lexique Verbaction où le nom d'action est lié morphologiquement et sémantiquement au verbe. Cela se justifie notamment par le fait que l'action est fondamentalement liée au verbe. C'est ce que Croft (1991) met en avant et il corrèle de façon prototypique les trois catégories syntaxiques que sont les verbes, les noms et les adjectifs aux deux niveaux qu'il identifie, à savoir la classe sémantique et la fonction pragmatique. Les noms réfèreraient donc prototypiquement à des objets, avec un rôle de référence. Les adjectifs réfèreraient quant à eux à des propriétés, avec un rôle de modification. Enfin, les verbes réfèreraient à des actions, avec un but de prédication. Mais les différentes fonctions pragmatiques mises en avant (référence, modification, prédication) peuvent très bien être instanciées par une autre classe sémantique que celle qui leur est prototypiquement associée : la fonction pragmatique de référence peut ainsi être réalisée par des items issus de la classe sémantique des actions (prototypiquement représentée par les verbes), et ce par le biais de noms d'action, de compléments, d'infinitifs ou de gérondifs. Selon cette représentation, le nom d'agent est par ailleurs vu comme la réalisation de la fonction pragmatique de prédication par des items issus de la classe sémantique des objets.

Pourtant, le critère morphologique n'est pas suffisant, et n'est pas toujours pertinent. Huyghe souligne qu'il n'y a pas toujours d'héritage des propriéts aspectuelles. Ainsi, tout nom déverbal dont le verbe de base dénote une action n'est pas systématiquement un nom d'action. Il existe par ailleurs certains noms d'action qui ne sont pas morphologiquement lié à un verbe, comme *bal*, *conférence* ou encore *stage*. Huyghe (2014) souligne donc l'importance d'un deuxième critère, celui de la description temporelle.

La description temporelle est une des propriétés fondamentales de sémantisme des noms d'action. Huyghe (2014) distingue trois traits différents propre à la description temporelle que sont l'ancrage, le repérage et l'extension. Tout nom d'action doit répondre à au moins un de ces trois traits, même si les traits d'ancrage et de repérage semblent plus largement partagés au sein des noms d'action que le trait d'extension. L'extension se caractérise chez les noms d'action par la description de situations duratives, dans des formules illustrées dans les exemples (27), (28) et (29).

(27) Det N a duré x temps
Le voyage a duré trois mois.

(28) Un N x temps
Une discussion de dix minutes

(29) x temps de N
Deux heures de bricolage

L'ancrage temporel correspond à la possibilité d'une localisation dans le temps de l'entité désignée par le nom d'action. Des tournures comme celles illustrées dans les exemples (30) et (31) permettent de mettre en avant cet ancrage.

(30) Il y a (eu) un N à tel moment
Il y a un entraînement à 14h.

(31) {la date/le moment/l'instant} du N
la date de l'inauguration

Le repérage temporel repose lui aussi sur l'idée de localisation dans le temps, mais il se distingue de l'ancrage temporel en cela que l'entité désignée par le nom d'action devient un repère temporel pour localiser d'autres éléments. L'emploi de prépositions temporelles telles que *lors de*, *pendant*, *avant* ou *depuis* permet de mettre en avant ce trait chez les noms d'action employées avec ces prépositions.

Si tout nom d'action doit présenter au moins un de ces trois traits, ces traits ne sont pas spécifiques aux noms d'action, comme on peut le voir dans des formules comme une chanson de trois minutes ou deux heures de joie.Le critère de description temporelle ne suffisant donc pas, Huyghe (2014) met en jeu le critère de dynamicité, propre à la définition même du nom d'action. Huyghe (2014) propose pour mettre en avant l'aspect dynamique d'un nom d'action l'emploi de construction avec des verbes supports actionnels ou événementiels, comme *accomplir*, *procéder à* , *avoir lieu* ou *se produire*. D'autres tournures, comme les expressions *une opération de N* ou *un mode de N*, ou la périphrase *en voie de N*, permettent aussi de caractériser la dynamicité des noms d'action.

Sur le plan sémantique, notons que Roché (2009) voit en la formation du nom d'action une opération formelle et catégorielle, mais pas sémantique, contrairement aux noms d'agent. En effet, dans le cadre de nominalisations processives, l'opération sémantique ne serait pas activée, comme dans le cadre du passage de *protéger* à *protection*. Cela signifie que le contenu sémantique du verbe est conservé lors de la création du nom d'action. Le nom d'action et le verbe ont donc un même contenu sémantique, contrairement au nom d'agent et au verbe, dont la dérivation morphologique entraîne une modification du contenu sémantique.

Nous l'avons vu dans la partie 1.2.1.2, les noms d'action sont généralement formés sur le thème de supin, puisque c'est ainsi que Anscombre (2001) définit l'un des thèmes servant à la formation des noms en *-eur*. Là encore, différents suffixes peuvent permettre la formation d'un nom d'action : citons notamment les suffixes *-age*, *-ment*, *-ure* et *-(a/i)tion*. Benveniste (1975) établit un parallélisme entre les noms d'agent en **-ter* et **-tor* et les noms d'action en **-tu* et **-ti* de l'indo-européen. En effet, ces suffixes partageraient les mêmes caractéristiques, à savoir objectivité et subjectivité. Les noms d'action marqueraient aussi cette distinction entre ce qui est intrinsèque et extrinsèque. Les noms en **-tu* marqueraient ainsi la capacité ou la manière d'accomplir, et seraient la

manifestation de l'agent. Il s'agirait davantage de noms d'opération. Les noms en *-*ti* marqueraient quant à eux le fait objectif de l'accomplissement, et l'actualisation de l'action. Il s'agirait en cela davantage de noms d'activité.

1.3.2. Une typologie des noms d'action

Nous venons de le voir, Benveniste distinguait parmi les noms d'action différentes sous-classes : les noms d'activité et les noms d'opération, en l'occurrence. En effet, à l'image des noms d'agent qui peuvent se diviser en trois classes telles que vu dans la section 1.2.1.2., les noms d'action (dans l'acception la plus globale) peuvent être de différents types, avec chacun leurs caractéristiques propres.

L'établissement d'une typologie des noms d'action est pourtant délicate : les délimitations floues de la catégorie des noms d'action rend en effet une sous-catégorisation plus complexe. Sur la base de l'hypothèse d'un héritage des propriétés aspectuelles du verbe en cas de nominalisation, on distingue parmi les noms d'action les sous-catégories d'activité, d'accomplissement et d'achèvement. Mais l'attribution d'une de ces trois étiquettes aux noms d'action n'est pas toujours aisée, et d'autres catégories se rencontrent régulièrement dans la littérature. Huyghe (2014) souligne notamment la forte proximité entre noms d'action (au sens strict) et noms statifs (comme *résignation* ou *dévouement*), que l'on retrouve pourtant couramment catégorisés comme noms d'action (dans son acception la plus générale). De même, une distinction peut être faite au sein des noms d'action entre noms de mouvement ou de déplacement comme *incursion*, noms de création ou de redescription comme *fabrication* ou *transcription*, noms de changement d'état comme *acroissement*, noms d'événement social comme *manifestation*, noms de spectacle comme *audition*, ou encore noms d'événement météorologique comme *orage*. Kerleroux (1999) distingue quant à elle les noms déverbaux résultatifs, comme *commémoration*, les noms d'activité (comme *construction* dans *la construction marche bien en ce moment*) et les noms d'événement, qu'elle divise en deux sous-groupes en fonction de la présence ou non d'une structure argumentale (l'absence de structure argumentale rapprochant les noms d'activité des noms d'événement dit simple). Huyghe (2014) discute lui aussi de la distinction entre noms d'événement et noms d'action en tant que tels lorsqu'il analyse le critère de dynamicité dont il est question dans la section 1.3.1. Il sépare ainsi les noms permettant de décrire des actions dites saturées (qui ne nécessitent pas de spécifications extérieures, donc pas de structure argumentale), i.e. les noms d'action au sens restreint du terme, de ceux nécessitant une spécification extérieure pour permettre la description de situations complètes sur le plan sémantique, à savoir les noms d'événement.

Au lieu de reprendre les différentes étiquettes que l'on attribue régulièrement aux noms d'action, Huyghe propose de se baser sur plusieurs distinctions sémantiques pour caractériser les noms d'action. Il suggère ainsi quatre oppositions, selon des critères d'occurrentialité, de durée, de télicité, et de fortuité.

Il distingue ainsi les noms d'actions occurrentielles des noms d'actions non occurrentielles, qu'il met en parallèle du caractère comptable ou massique de ces noms. Les noms comptables comme *cambriolage* ou *expulsion* peuvent en effet dénoter des occurrences d'action, que Huyghe définit comme « des situations dynamiques intrinsèquement individuées » (2014:9). Les noms massifs comme *jardinage* ou *natation* ne peuvent quant à eux pas dénoter des événements puisque les situations qu'ils décrivent ne sont pas bornées. Il n'y a pas d'instanciation de l'action, ce qui se traduit par l'agrammaticalité de la deuxième phrase illustrée dans l'exemple (32), contrairement à la première phrase, avec l'usage de la tournure *avoir lieu* mise en avant dans la description du critère de dynamicité dans la section 1.3.1.

(32) *Le cambriolage a eu lieu dans l'après-midi.*
**Le jardinage a eu lieu dans l'après-midi.*

Cette opposition basée sur l'occurrentialité confirme l'existence d'une distinction entre noms d'événement d'une part et d'autres noms d'action d'autre part au sein de la catégorie globale des noms d'action.

Nous l'avons vu dans la section 1.3.1., la temporalité intervient dans deux des trois traits permettant d'identifier un nom d'action. Elle intervient aussi dans la deuxième opposition que Huyghe (2014) établit entre actions duratives et actions non-duratives. Les noms d'action non occurrentielle sont duratifs. Il est donc possible de les utiliser dans des constructions impliquant des verbes aspectuels comme *commencer* ou *s'interrompre*, ou dans des tournures comme *x temps de N*. Les noms occurrentiels sont quant à eux soit duratifs, soit non duratifs. Une partie d'entre eux relevent de la catégorie exprimant à divers degrés l'achèvement par le biais d'une réalisation ponctuelle comme *entrée*, d'un changement précédé d'une préparation comme *élimination*, ou d'une action qui s'achève sur un état statif duratif comme *coupure*.

La troisième opposition mise en avant par Huyghe (2014) repose sur l'existence d'un « point culminant qui actualise l'action décrite » (2014:10), ou *telos*, dans la structure de certains noms d'action occurrentiels, notamment ceux duratifs. Ce critère de télicité peut être testé à l'aide du paradoxe imperfectif qui repose sur l'idée que l'interruption d'une action culminante vient empêcher la réalisation de cette action. Ainsi les exemples illustrés en (33) (tiré de Huyghe, 2014) montrent respectivement un événement non culminant et un événement culminant.

(33) *La manifestation a été interrompue* implique *Ils ont manifesté.*
La réparation du vélo a été interrompue n'implique pas *Il a réparé le vélo.*

Si par nature les noms d'action non occurrentiels sont atéliques et les noms d'action non duratifs téliques, Huyghe (2014) souligne que la télicité de certains noms d'action peut varier dans certains cas.

Huyghe (2014) propose comme dernière grande opposition celle basée sur le caractère fortuit ou non d'une action. Cette opposition s'applique uniquement aux noms d'action occurrentiels et peut

être testée à l'aide de constructions employant *se produire*, qui mettent en avant des événements accidentels sur lesquels au moins une partie des participants n'a pas le contrôle, comme *séisme* ou *explosion*. Ces événements fortuits ne peuvent par ailleurs pas être employé en tant que complément de lieu. Les noms d'action non fortuits peuvent quant à eux être employés dans des constructions basées sur *être prévu* ou *être reporté*, qui permettent de décrire des actions programmées. Les noms compatibles à la fois aux constructions propres aux noms fortuits et à celles des noms non fortuits décrivent quant à eux des événements prémédités, mais jugés accidentels et imprévisibles pour les participants, à l'image de *braquage* ou *attentat*.

Sur la base de ces quatre opposition, Huyghe propose de représenter la structure sémantique d'un nom d'action sous la forme (34) (tiré de Huyghe, 2014).

(34) *jardinage* : [-occurrentiel][+duratif][-culminant]
réparation : [+occurrentiel][+duratif][+culminant][-fortuit]

Huyghe (2014) met en avant d'autres oppositions, plus mineures, comme l'opposition entre noms d'action agentifs et non agentifs, que l'on peut tester à l'aide d'une construction utilisant le verbe *effectuer*, ou l'opposition entre noms d'action intensifs et non intensifs, que l'on peut tester avec l'emploi antéposé de *fort* et *faible* ou de *un degré de*.

1.4. Polysémie et réseaux secondaires

Si certains noms d'action peuvent être clairement catégorisés selon les étiquettes ou critères précédents détaillés, ce n'est pas le cas pour l'ensemble des noms d'action. En effet, on dénombre de nombreux cas où un même nom d'action peut être interprété de différentes façons en fonction de sa structure argumentale ou simplement de la phrase dans laquelle il est employé. On parle alors de polysémie. C'est notamment le cas de *construction*, comme l'illustre l'exemple (35).

(35)	*La construction du pont a pris 5 ans.*	action
	La construction est un secteur en pleine croissance.	activité
	Cette construction est un chef d'œuvre.	objet résultatif

Le cas de polysémie des noms d'action le plus courant est celui d'une interprétation résultative, comme l'illustre par exemple (35). Le mot-type *construction* propose deux acceptions, une acception processive, qui correspond à l'action (ou l'événement complexe, selon la terminologie de Kerleroux, 1999), et une acception objectuelle, qui correspond au résultat de l'action décrite, à l'entité créée par cette action. Ce résultat est ici matériel, puisqu'il s'agit d'un objet, mais il peut aussi être informationnel, comme dans le cas de *traduction* (qui associe en l'occurrence trois acceptions : action, contenu informationnel et support matériel). L'interprétation résultative s'applique aussi aux états, les

noms d'action comme *humiliation* ou *satisfaction* décrivant alors l'action dénotée par le verbe de base et l'état produit par cette action. D'autres liens polysémiques peuvent unir les différentes acceptions d'un nom d'action, comme dans le cas des noms d'action dénotant une action et son instrument, à l'instar de *jeu* ou de *convocation*, mais aussi le cas des noms d'action dénotant une action et son moyen, à l'instar de *éclairage*, le cas des noms d'action dénotant une action et son agent, à l'instar de *rébellion*, ou encore des noms d'action dénotant une action et son lieu, à l'instar de *croisement*. Certaines relations sont plus courantes que d'autres, comme l'association du mouvement et du lieu pour les noms de déplacement comme *parcours*, du repas et de l'aliment pour les noms de repas comme *déjeuner*, des interventions orales et des supports matériels pour les noms d'intervention orale comme *exposé*, ou encore l'association de l'action, du résultat et de l'agent collectif pour un nom comme *rédaction*. Cette relation ne se limite pas toujours à deux acceptions, certains noms d'action comme *traduction* ou *rédaction* dont nous parlions précédemment alliant trois acceptions.

Mais le phénomène de polysémie des noms d'actions ne s'arrête pas à une question d'interprétation. Ainsi, dans son étude sur le phénomène d'apocope en français qui permet de former par troncation de nouveaux mots comme *labo* pour *laboratoire* ou *manif* pour *manifestation*, Kerleroux (1999) met en avant l'impact de la polysémie de certains noms sur le phénomène d'apocope, un même mot-type n'acceptant pas de la même façon l'apocope en fonction du contexte, comme l'illustre l'exemple (36) (tiré de Kerleroux, 1999) pour le mot-type *manifestation*.

(36) **La manif de la vérité aura pris cinquante ans.*
La manif des étudiants a duré cinq heures.

Outre la différence d'interprétation, *manifestation* étant respectivement interprété dans (36) comme un nom dénotant un procès et un nom dénotant un résultat, on peut s'interroger sur les raisons de cette limitation du phénomène d'apocope. La question se pose alors du véritable lien entre ces deux occurrences de *manifestation* ou s'il ne s'agit pas plutôt de deux mots distincts. Lorsque l'on recherche *manifestation* dans le *Trésor de la Langue Française informatisé*[1] (maintenant TLFi), on constate qu'il y a deux entrées distinctes. La première entrée définit un premier lexème *manifestation* comme un substantif féminin décrivant « l'action, le fait de révéler », le « signe et (la) forme sensible d'un principe immatériel », ou une « présentation publique ». La deuxième entrée définit quant à elle un deuxième lexème *manifestation* comme un substantif féminin décrivant un « rassemblement de personnes [...] dans le but de faire connaître une opinion ». Il est intéressant de noter que la première entrée renvoie à une première entrée *manifester*, qui définit un verbe transitif dénotant le procès de « faire connaître », de « révéler » quelque chose, qui est distincte d'une deuxième entrée *manifester*, qui définit quant à elle un verbe intransitif dénotant le procès de « faire une manifestation, participer à une manifestation ». S'il est clair que le second lexème *manifester* découle largement sur le plan sémantique du premier lexème *manifester*, ils sont désormais bien distincts. On se retrouve donc dans le cadre de *manifestation* avec deux familles dérivationnelles distinctes, toutes deux basées sur un lexème *manifester* distinct. Si l'on voit une famille dérivationnelle comme un réseau, on peut

1 Disponible à l'adresse http://atilf.atilf.fr/tlf.htm

s'interroger sur l'existence d'un « réseau secondaire »[2] que constituerait l'une de ces deux familles dérivationnelles.

Si la présence de deux réseaux semble assez limpide dans le cas de *manifestation*, comme nous l'avons vu dans la section 1, c'est parfois moins évident et plus questionnable dans d'autres cas. Prenons par exemple la famille dérivationnelle basée sur le verbe *transporter*, qui forme notamment le nom d'action *transport* et le nom d'agent *transporteur*. Le TLFi ne propose pour ces trois mots qu'une seule entrée chacun. Pourtant, on peut noter deux interprétations de *transport* : une acception processuelle, et une acception d'activité, à l'instar de *construction*. On peut alors se demander s'il n'y a pas deux noms d'agent *transporteur* distincts, un qui serait lié à l'action, et l'autre à l'activité, et s'il n'y a pas ici deux réseaux distincts. Le réseau qui se baserait sur l'activité ne contiendrait pas de verbe. On observe la même chose avec le triplet *danseur*/*danser*/*danse*, où *danseur* peut dénoter à la fois l'agent occasionnel et le nom statuaire (selon la terminologie évoquée dans la section 1.2.1.2.), et *danse* à la fois l'action et l'activité. Deux réseaux sémantiques coexisteraient sur la base des mêmes termes, mais l'un correspondant à l'action et l'autre à l'activité.

Nous avons vu dans cette première section de nombreuses caractéristiques propres aux noms d'agents, aux noms d'action, mais aussi aux noms d'instruments. On peut s'interroger sur l'impact de ces caractéristiques. En effet, on peut se demander si le thème de supin ou de présent à la base de la formation d'un nom d'agent se traduit par un comportement sémantique spécifique. De même, on peut s'interroger sur la charge sémantique des suffixes *-euse* et *-rice*, et se demander si la différence entre un nom d'agent féminin et un nom d'agent masculin formés sur la même base est simplement de l'ordre du genre, ou s'il y a des conséquences sémantiques plus lourdes liées au choix du suffixe.

Les descriptions proposées sont autant d'outils utilisables pour caractériser les données que nous allons analyser dans ce mémoire. Tous ne seront pas utilisés dans le cadre de ce travail, mais cela pose tout de même la question de l'application de ces descriptions. Si l'intuition est un outil non négligeable, on peut cependant s'interroger sur l'automatisation de cette caractérisation. Notons que si tous les tests ne sont pas forcément aisément applicables, des pistes sont envisageables quant à une généralisation de la description. La projection sur corpus des patrons mis en évidence, notamment pour les noms d'instrument ou les noms d'action, pourrait notamment éclairer de façon automatique de la nature instrumentale ou agentive ou processive de certains noms. La question se pose alors de la nature du corpus à choisir, entre corpus représentatif des usages actuels ou des définitions académiques. L'utilisation de ressources morphologiques pourrait là aussi éclairer la formation des noms d'agent afin de les caractériser selon les traits mis en avant plus tôt.

2 Terminologie employée par Michel Roché lors de différents échanges

2. Indices distributionnels de proximité sémantique

Dans la mesure où les noms d'agents en -*EUR* et les noms d'action appartiennent à un même paradigme dérivationnel, ils partagent certaines informations sémantiques. Nous l'avons vu, Roché (2009) soutient que la formation du nom d'action n'implique pas d'opération sémantique, contrairement à la formation du nom d'agent qui active quant à elle cette opération. Le contenu sémantique est donc modifié. Pour tester cette hypothèse, il nous faut donc comparer l'écart de sens qu'il y a véritablement entre le nom d'agent en -*EUR* et le verbe d'une part, à l'écart de sens qu'il y a entre le nom d'action et le verbe d'autre part. Pour cela, nous devons donc commencer par analyser la proximité sémantique de deux types d'objet différents : des noms et des verbes.

2.1. Similarité et catégories

La notion de proximité sémantique a été approchée de nombreuses façons, tant elle est essentielle au TAL. En effet, nombre d'applications reposent sur les relations sémantiques entretenues par différents items, comme la synonymie, l'hyperonymie ou l'opposition (Fabre, 2010). La recherche d'information se base sur les relations sémantiques pour étendre et enrichir une recherche ; les systèmes de questions-réponses reposent sur les liens de parenté pour déterminer que deux énoncés partagent un même contenu sémantique ; la traduction et le résumé automatiques dépendent de bases de données riches en mots sémantiquement liés. La variabilité étant une des caractéristiques essentielles du langage naturel, le Traitement Automatique des Langues et ses applications dépendent donc beaucoup de leurs capacités à traiter cette variabilité.

Les relations morphologiques ne sont qu'un lien de parenté parmi ceux cités plus haut. Elles représentent à la fois un outil intéressant, puisque la proximité sémantique est doublée d'une proximité formelle, ce qui en facilite l'exploitation, et un obstacle majeur dans le cadre du calcul de la proximité sémantique puisque, de par une dérivation morphologique impliquant une opération catégorielle telle que la formation des noms d'agent en -*eur*, l'absence d'unicité catégorielle empêche *a priori* toute comparaison. En effet, les deux entités n'étant pas sur le même plan, elles ne sont pas comparables en tant que telles.

Cette idée de séparation par catégorie se retrouve notamment dans des projets comme WordNet. WordNet est une base de données lexicale de l'anglais (Miller, 1995) liant les différents noms, adjectifs, adverbes et verbes de l'anglais grâce à les liens et relations sémantiques, afin de créer un réseau. Six relations sémantiques ont ainsi été représentées, toutes choisies pour leur forte présence dans la langue anglaise, et par la facilité de compréhension de ces relations par quasi n'importe quel locuteur : la synonymie, qui s'applique sur les quatre catégories de mots représentées, l'antonymie, s'appliquant elle aussi sur l'ensemble des catégories, l'hyponymie, la méronymie (ces deux types de relation ne s'appliquant qu'aux noms), la troponomie et l'implication, ne s'appliquant qu'aux verbes.

Mais ces relations ne s'appliquent que sur deux items appartenant à la même catégorie. On se retrouve donc avec plusieurs réseaux lexicaux bien distincts, un verbal, un nominal, et ainsi de suite, malgré les liens sémantiques qu'entretiennent concrètement certains noms et verbes entre eux.

Pourtant, certaines initiatives telles que l'EuroWordNet se sont déjà attachées à combler ce manque. EuroWordNet est le fruit de l'extension de la base de données WordNet à une échelle multilingue, associant les WordNet ou autres réseaux lexicaux équivalents de l'anglais, de l'italien, de l'espagnol et du néerlandais. Il fonctionne donc sur le même principe que WordNet (Miller, 1995), impliquant les mêmes relations entre mots, et la mise en place de liens intra-catégoriels. Mais EuroWordNet propose en plus l'ajout de relations inter-catégorielles dans le but premier de combler les différences entre les langues (certains termes ou concepts n'appartenant pas à la même catégorie grammaticale d'une langue à l'autre, ou certaines langues ne faisant pas cette distinction), mais aussi pour permettre le regroupement jusque là impossible de mots sémantiquement similaires mais non rapprochés du fait de leur différence catégorielle. Enfin, d'un point de vue applicatif, cela permet d'améliorer l'utilité de l'EuroWordNet comme outil pour la recherche d'information en permettant la mise en correspondance d'un contenu donné avec un plus grand nombre de structures syntaxiques. Ont donc été ajoutées les *explicit cross-part-of-speech relations* (Vossen, 1997), c'est-à-dire les relations explicites de synonymie ou d'hyponymie entre deux termes n'appartenant pas à la même catégorie, à l'image de l'exemple (37) (tiré de Vossen, 1997). Ont aussi été ajoutées les relations de cause, pouvant lier un verbe et son résultat, et les relations de rôle, liant un événement et ses participants, à l'image des exemples (38) (tirés de Vossen, 1997).

(37) {to kill} V CAUSES {death} N
{death} N IS_CAUSED_BY {to kill} N *reversed*

(38) {hammer} N ROLE_INSTRUMENT {to hammer} V
{to hammer} V INVOLVED_INSTRUMENT {hammer}N *reversed*

Une autre façon de passer outre la notion de catégorie liée à l'objet étudié est de se baser sur autre chose que l'objet lui-même. Ainsi, nous l'avons vu dans l'introduction, l'idée que le contexte d'un mot peut nous apprendre beaucoup sur le mot lui-même est soutenue depuis longtemps. C'est notamment l'hypothèse soutenant les méthodes distributionnelles, qui décrivent des mots par le biais de leurs distributions.

2.2. Proximité distributionnelle et proximité sémantique

L'approche distributionnelle établit une corrélation entre la similarité distributionnelle et la similarité sémantique (Sahlgren, 2008). Ainsi, des mots sémantiquement similaires partageraient des voisins distributionnels, c'est-à-dire des mots qui partagent des cooccurrents syntaxiques. Une autre

façon de formuler la chose est la suivante : « words with similar meanings will occur with similar neighbors if enough text material is available » (Schütze et Pedersen, 1995, cité par Sahlgren, 2006). Notamment théorisée par Harris et Bloomfield, cette approche voudrait que la différence de sens soit exprimée par la différence de distribution. Harris propose de quantifier en effet la différence de sens de la façon suivante : « The amount of meaning correspond[s] roughly to the amount of difference in their environments. » (1954:157). Harris (1954) définit la distribution d'un élément comme étant l'ensemble de ses « environnements », à savoir un ensemble de ses cooccurrents. Calculer la proximité sémantique de deux termes reviendrait à comparer la distribution de chaque terme. Moins deux termes partagent de voisins distributionnels, moins ils sont similaires. La notion de voisins distributionnels est le fruit de cette approche. Ainsi, l'approche distributionnelle permet de dire que *maison* et *bâtiment* sont des voisins distributionnels car ils partagent comme cooccurrents syntaxiques *édifier*, *rénover*, *charpente de* ou encore *démolition de*.

Les modèles distributionnels représentent les mots sous la forme de vecteurs, dont il suffit alors de calculer la distance entre eux pour calculer la similarité des mots comparés, à l'image de la recherche d'information. Ces vecteurs sont définis à partir de la distribution, et donc des contextes, des mots. Plusieurs types de contextes, au nombre de trois, peuvent être considérés, que Fabre et Lenci (2015) décrivent.

Des mots peuvent tout d'abord être rapprochés d'autres mots s'ils apparaissent au sein d'un même paragraphe ou document. On parle alors de modèles « document-based » (Fabre et Lenci, 2015). Ces modèles seraient davantage performants dans l'identification des voisins appartenant à un même thème sémantique que les autres modèles.

Les modèles « word-based » s'intéressent quant à eux plus précisément aux cooccurrents graphiques des mots cibles dans une certaine fenêtre autour des mots cibles, selon une approche dite « sac-de-mots ». Ces modèles tendraient à mieux identifier les relations d'association que les autres modèles (Fabre et Lenci, 2015). La similarité mise au jour par ces modèles est dite « attributional », selon les termes de Turney et Pantel (2010), signifiant que la similarité des deux mots comparés dépend du degré de correspondance entre les propriétés de ces mots. Turney et Pantel (2010) donnent ainsi l'exemple de *chien* et *loup* comme étant deux mots ayant un fort degré de similarité attributionnelle.

Les modèles « syntax-based » utilisent quant à eux les relations de dépendance des mots pour les comparer. Ces modèles identifieraient quant à eux davantage les voisins distributionnels liés sur un plan ontologique, comme des co-hyponymes. Turney et Pantel (2010) parlent pour les modèles « syntax-based » de similarité relationnelle, à savoir la similarité entre deux paires de mots. L'exemple de *chien*:*aboyer* et *chat*:*miauler* est notamment donné pour illustrer un fort degré de similarité relationnelle.

Sahlgren (2008) utilise les termes de modèle paradigmatique et de modèle syntagmatique pour désigner les deux derniers modèles décrits par Fabre et Lenci (2015). Le modèle syntagmatique repose sur la cooccurrence, et le modèle paradigmatique sur les voisins distributionnels. Dans le cadre d'un modèle basé sur la cooccurrence, la notion de fenêtre de recherche est importante, puisqu'elle

aura un impact crucial sur les résultats. En effet, plus la fenêtre sera étendue, plus le premier terme de la comparaison aura de chance de cooccurrer avec l'autre terme de la comparaison. Mais une fenêtre trop restreinte limiterait sans doute les cooccurrences pertinentes. Le modèle paradigmatique concerne les mots qui partagent souvent les mêmes mots de contexte. Non seulement la taille de la fenêtre sera importante, puisqu'elle augmentera ou diminuera le nombre de voisins potentiels, mais la position des voisins va aussi jouer un rôle.

Dans une matrice distributionnelle, chaque vecteur est une représentation de la distribution d'un lexème dans un contexte précis. Cela se traduit donc par un nombre très important de vecteurs, ce qui alourdit et ralentit tout calcul. Les matrices sont donc généralement réduites, afin de limiter le nombre de dimensions impliqués dans les calculs. Cette réduction est opérée à l'aide de différentes méthodes visant toutes à optimiser l'espace vectoriel, 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). Tous ces traitements sont donc appliqués sur des données que l'utilisateur ne voit pas, et les seules données auxquelles il peut accéder ne correspondent plus à la représentation explicite des contextes, mais à une schématisation sémantique. L'utilisateur ignore sur quelles bases le modèle a réduit de la sorte telle ou telle matrice, ce qui réduit la lisibilité des résultats que la matrice fournit.

Mais la réduction de la matrice ne constitue que la troisième des quatre grandes étapes de création d'une matrice par un système de sémantique distributionnelle. En effet, le système commence d'abord par parcourir le corpus, et pour chaque mot cible, il collecte les contextes de ce mot, contextes qui sont comptabilisés. Cela permet d'obtenir des fréquences qui vont permettre de caractériser l'importance de certains contextes et donc la plus grande significativité de ces contextes par rapport à d'autres. Sur cette première matrice a donc ensuite lieu l'opération de réduction. La nouvelle matrice obtenue fait alors l'objet d'un calcul de la similarité de ses différents éléments.

Comme nous l'avons dit précédemment, on estime la similarité des mots d'une matrice sur la base de la distance entre leurs vecteurs. Pour cela, on calcule le cosinus de l'angle entre les vecteurs, comme l'illustre le schéma 1 qui simplifie la représentation d'un espace multidimensionnel à un espace à deux dimensions. Le calcul du cosinus permet de réduire l'importance de la fréquence de chaque mot, qui a elle un impact sur la longueur des vecteurs.

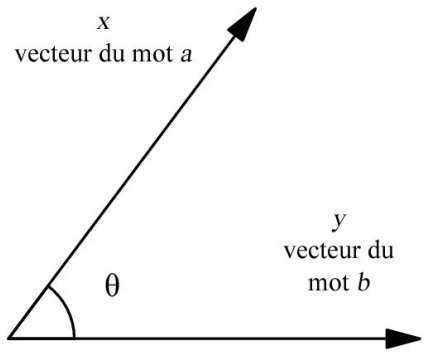


Schéma 1 *– Représentation du cosinus de θ*

Le schéma 1 illustre la comparaison des vecteurs *x* et *y* des deux mots *a* et *b*. Le cosinus de l'angle θ séparant les vecteurs des mots *a* et *b* correspond à la distance entre les deux mots. Si le cosinus est égal à 1, l'angle θ est de 0 degrés, signifiant que les deux vecteurs *x* et *y* sont géométriquement identiques : les mots *a* et *b* sont donc distributionnellement identiques. *A contrario*, si le cosinus de θ est égal à 0, l'angle θ est de 90° degrés, ce qui signifie que les vecteurs sont orthogonaux : les mots a et b sont donc distributionnellement différents. Le degré de similarité distributionnelle est donc défini par la valeur prise par le cosinus de θ. Cette mesure de cosinus peut alors être convertie de diverses façons en une mesure de similarité (Turney et Pantel, 2010).

Les modèles distributionnels se basent sur des corpus pour représenter le sémantisme en contexte des mots, le choix du corpus va donc avoir un effet non négligeable sur les résultats de l'analyse distributionnelle (Fabre et Lenci, 2015). Une analyse distributionnelle nécessite des données conséquentes sur le plan quantitatif, comme l'ont souligné Schütze et Pedersen (1995). En effet, plus le corpus sera important, plus la similarité de certains mots sera perceptible parmi le bruit que représentent de simples cooccurrences non significatives (Rychlý et Kilgarriff, 2007). Ainsi, plus on augmente la taille du corpus, plus les résultats seront précis et pertinents. Par ailleurs, utiliser des corpus aussi larges et variés que possible permet de couvrir un plus grand champs lexical, de façon plus complète, ce qui améliore les performances des systèmes distributionnels.

2.3. Introduction à Word2Vec

Word2Vec[3] est un outil d'apprentissage non supervisé mis au point par Tomas Mikolov et son équipe (Mikolov, T., Chen, K., Corrado, G., et Dean, J., 2013) ayant pour but d'apprendre les relations sémantiques entre les mots. Cet outil permet de créer une représentation vectorielle de l'ensemble des mots d'un texte et d'exploiter cette représentation, à l'aide de différents modules.

L'utilisation de Word2Vec commence par la création d'une matrice qui est spécifique au corpus que l'on fournit à l'outil. Cette matrice, qui contient la représentation vectorielle du sens des mots du corpus, peut être au format binaire ou au format texte. Word2Vec parcourt le texte qu'on lui fournit avec une certaine fenêtre dans le but d'apprendre pour chaque mot son contexte. Intervient alors un filtrage des mots appris sur la base de la fréquence, pour ne conserver que les mots ayant une fréquence minimum dans le corpus. Différents paramètres peuvent être modifiés et personnalisés pour la création d'une matrice. L'utilisateur peut notamment choisir entre une architecture en sac de mots, ou CBOW, où l'ordre des mots n'a pas d'impact, et qui permet de prédire le mot actuel en fonction du contexte, et une architecture en skip-gram, basée sur le principe des *n*-grammes (fruit du découpage d'une chaîne donnée en sous-séquences de *n* éléments) mais permettant d'unir des éléments qui ne sont pas directement consécutifs à l'intérieur d'une certaine fenêtre, et qui permet quant à lui de prédire le contexte à partir du mot actuel (Mikolov, Chen, Corrado et Dean, 2013). L'algorithme d'apprentissage

3 L'outil Word2Vec et sa documentation sont disponibles à l'adresse https://code.google.com/archive/p/word2vec/.

est aussi personnalisable, le choix étant laissé entre le *hierarchical softmax*, jugé plus performant pour les mots peu fréquents, et le *negative sampling*, jugé quant à lui plus performant pour les mots fréquents ou pour des mots aux vecteurs avec un faible nombre de dimensions. Le nombre de dimensions est lui aussi personnalisable, tout comme la taille de la fenêtre, un seuil maximal de fréquence d'apparition des mots ou les frontières de mots.

Le fonctionnement propre à Word2Vec dans la réduction opérée lors de la création de la matrice n'étant pas directement accessible, il faut donc rester prudent quant à l'analyse des résultats. En effet, on ignore les rapprochements qui ont pu être faits par l'algorithme ou les raccourcis faits par la matrice. Puisque nous ne pouvons décrire précisément le processus de création des vecteurs, les résultats obtenus grâce à Word2Vec ne pourront être directement comparés à des résultats obtenus par des méthodes d'analyse distributionnelle classiques. Ces résultats ne seront valables que pour la représentation vectorielle créée par Word2Vec. Il faudra aussi prendre en compte le fait que nous avons utilisé le modèle « word-based » de Word2Vec, qui construit sa matrice à partir de cooccurrents graphiques et non pas syntaxiques. Il n'y a pas de réelle analyse morphosyntaxique du corpus d'entrée. L'outil construit une représentation vectorielle des tokens, et ne tient donc pas compte de la catégorie grammaticale des mots.

Des modules nous permettent d'interroger cette matrice non interprétable en l'état et de nous fournir des informations plus faciles à interpréter, toujours en gardant la distance nécessaire. Différents modules sont à notre disposition pour exploiter la représentation vectorielle du sens des mots apprise par Word2Vec.

Le module *distance* permet ainsi d'obtenir les voisins distributionnels d'un mot cible ainsi que la distance qui les sépare sur la base de la mesure cosinus : l'angle séparant les vecteurs des mots est calculé à l'aide du cosinus, la valeur de cet angle oscillant entre 0 et 1 en fonction de la proximité des mots. Le module *analogy* permet quant à lui de trouver par analogie le quatrième membre d'un groupe tel que *king-man-woman* (qui serait ici *queen*). Un autre module, *classes*, permet de classer les mots qu'on lui fournit en les regroupant par classe.

Par ailleurs, la distance séparant deux mots sur la base du cosinus de leur vecteur n'est analysable que dans le cadre d'une comparaison entre plusieurs paires, la lecture d'un cosinus n'étant pas significative en elle-même (à part lorsque le cosinus vaut 1, auquel cas les vecteurs des deux mots sont identiques, et donc les deux mots sémantiquement identiques). Il sera fait référence dans la suite de ce mémoire à ce chiffre, appelé *Cosine distance* par l'outil, sous l'étiquette « indice de proximité » pour plus de lisibilité.

3. La démarche expérimentale

Ce mémoire consiste en une comparaison sémantique des dérivés morphologiques processifs et agentifs en *-EUR* à l'aide d'indices distributionnels. Cette analyse repose sur une ressource lexicale, Lexeur, que nous allons enrichir d'informations distributionnelles grâce à l'outil Word2Vec présenté en 2.3.

3.1. Les données initiales

Parmi les données initiales, on compte un lexique des noms d'agent en *-EUR* et des familles dérivationnelles auxquelles ils appartiennent, et des corpus textuels.

3.1.1. Lexeur

Le lexique sur lequel se base dans un premier temps la démarche est un lexique descriptif fini du nom de Lexeur. Lexeur regroupe 5974 noms en *-EUR*, classés par ordre lexicographique, issus dans leur grande majorité du TLFi. Les autres noms en *-EUR* présents dans Lexeur sont des formes dont la présence a été attestée sur Internet. Lexeur est le fruit de la fusion de deux bases de données constituées il y a une dizaine d'années par deux annotateurs différents. L'ensemble a été unifié pour former un seul et même ensemble.

Sont associés aux noms d'agent masculins en *-eur* listés le nom d'agent féminin correspondant en *-euse* et/ou en *-rice*, le verbe ou le nom dont ils dérivent, la catégorie grammaticale de la base, ainsi que les noms d'action potentiels associés à la base nominale ou verbale. C'est donc une partie de la famille constructionnelle de la base qui est reconstituée. Chaque famille constructionnelle est affichée sur une même ligne. Les entrées sont présentées selon la forme illustrée par les exemples regroupés dans le tableau 1 suivant.

De plus, une première annotation des lexèmes est proposée, à savoir un étiquetage grammatical des entrées, comprenant sa catégorie (*Nc* pour nom commun, *V* pour verbe), ainsi que son nombre et son genre (*m* pour masculin, *f* pour féminin, *s* pour singulier). Puisque les lemmes sont présentés sous leur forme citationnelle (*i.e.* forme non fléchie), tous les noms sont donc au singulier et les verbes à l'infinitif. Par convention vis-à-vis du système d'annotation choisi, la forme citationnelle des verbes est annotée *Vmn*.

Nom d'agent M.	Nom d'agent F.	Base	Cat.	Dérivés
abatteur/Ncms	abatteuse/Ncfs	abattre/Vmn----	Vb	abat/Ncms abattement/Ncms abatture/Ncfs abattage/Ncms abattis/Ncms
culteur/Ncms	culteuse/Ncfs cultrice/Ncfs	culte/Ncms	Nb	Ø
endoscopeur/Ncms	endoscopeuse/Ncfs	Ø		Endoscopie/Ncfs
fraudeur/Ncms	fraudeuse/Ncfs	frauder/Vmn----	Vb	fraude/Ncfs
frauduleur/Ncms	frauduleuse/Ncfs	frauder/Vmn----	Vb	fraude/Ncfs
wheeleur/Ncms	wheeleuse/Ncms	wheel/Ncms	Nb	Ø

Tableau 1 *– Exemple de six entrées de Lexeur*

Le tableau 1 illustre la grande diversité des familles dérivationnelles regroupées dans Lexeur. Toutes les familles dérivationnelles ne contiennent en effet pas les mêmes éléments. Ainsi, si l'ensemble des familles dérivationnelles présente un nom d'agent masculin et un nom d'agent féminin, 82 familles dérivationnelles proposent les deux variantes féminines *-euse* et *-rice* du noms d'agent, à l'image de l'entrée *culteur* dans le tableau. Notons que 43 de ces 82 familles sont des noms issus de champs lexicaux techniques propres à la culture et à l'élevage, ayant comme base *cultiver*, utilisant comme radical le thème *cult-*, à l'image des noms d'agent féminins *acériculteuse/acéricultrice*, *colombiculteuse/colombicultrice* ou encore *hélicicultrice/héliciculteuse*. Notons que le suffixe *-euse* est beaucoup plus largement représenté que le suffixe *-rice*, à raison de 4542 noms d'agent féminins suffixés en *-euse* (y compris les 82 familles contenant les deux variantes féminines) et de 1514 noms d'agent féminins suffixés en *-rice* (y compris une nouvelle fois les 82 familles contenant les deux variantes).

Lexeur regroupe sans distinction des noms d'agent et des noms d'instrument : on retrouve ainsi indifféremment des termes comme *réfrigérateur*, *compacteur* ou *autorégulateur*, mais aussi des termes comme *commentateur* ou *organisateur*. Aucune indication ne permet de distinguer *a priori* les noms d'agent et noms d'instrument dans Lexeur, et il s'agira donc lors de l'exploitation de Lexeur d'identifier nous-mêmes les deux types de de noms grâce aux tests décrits dans la section 1.2.3.2.

De même, le tableau 1 illustre les cas où l'on observe deux noms d'agent masculins différents dérivés d'un même verbe, à l'image de *fraudeur* et *frauduleur*, et qui ravivent la question de la concurrence suffixale évoquée dans la section 1.2.

Toutes les familles ne proposent pas a priori de base. C'est notamment le cas de l'entrée *endoscopeur* dans le tableau, et c'est plus généralement le cas de 443 familles dérivationnelles qui ne proposent pas de base directe dans Lexeur (en l'occurrence, pour lesquelles il n'y a pas d'annotations dans la colonne indiquant la catégorie de la base). Notons que pour les bases annotées, 4676 d'entre elles sont des verbes, et 855 d'entre elles sont des noms, soit 78% des familles dérivationnelles de Lexeur construites sur la base d'un verbe. Cela confirme le caractère majoritairement verbal de l'origine des noms d'agent en *-EUR* mis en avant dans la section 1.2.

De même, toutes les familles ne proposent pas de noms d'action, à l'image de l'entrée *wheeleur*. Les noms d'action sont eux-mêmes très divers, et l'on retrouve aussi bien des noms d'activité comme *couvrage* que d'objets ou d'états résultatifs comme *fondation* ou *espoir*. Les différents types de noms d'action recensés ne sont pas caractérisés selon la typologie évoquée dans la section 1.3.2., ce qui nécessitera des observations et des analyses de notre part lors de l'exploitation de Lexeur.

Lexeur est une ressource très riche, mais à toujours manipuler avec un certain recul. Ainsi, on retrouve certains défauts dans sa constitution. On peut constater des erreurs typographiques comme par exemple l'entrée illustrée en (39), à savoir le premier mot-forme *débiter* ici supposé être nom d'agent, ou tout simplement dans les annotations morphosyntaxiques dans la colonne attribuée à la catégorie de la base. On retrouve ainsi 24 annotations *Xb*, à la place de l'annotation *Nb*, comme dans le cas de *collision*, et 1 cas où l'annotation *Vn* remplace *Vb*, le cas de *auto-exciter*.

(39) *débiter/Ncms débiter/Ncfs Ø débit/Ncms*

De même, l'unification de deux travaux distincts a amené à la conservation accidentelle de doublons, à l'image de l'entrée illustrée en (40) que l'on retrouve deux fois à l'identique (et d'affilée) dans Lexeur.

(40) *énumérateur/Ncms énumératrice/Ncfsénumérer/Vmn Vb énumération/Ncfs*
énumérateur/Ncms énumératrice/Ncfsénumérer/Vmn Vb énumération/Ncfs

Mais il y a aussi des erreurs résiduelles lors de la constitution des familles dérivationnelles comme dans le cas de l'entrée illustrée (41), où le nom d'action *ouvrage* semble hors de propos, *ouvrage* dérivant non pas du verbe *ouvrir* mais du verbe *ouvrer*, et n'appartenant donc pas à cette famille dérivationnelle.

(41) *ouvreur/Ncms ouvreuse/Ncfs ouvrir/Vmn Vb ouverture/Ncfs ouvrage/Ncms*

Enfin, on peut s'interroger sur la constitution des familles dérivationnelles, à relier à la notion de réseau secondaire telle que vue dans la section 1.4. Si l'absence d'un nom d'agent en -*eur* dérivé de *manifester* nous évite de nous interroger sur l'intégration des deux familles dérivationnelles fondées sur *manifester*, la présence de *transporteur* sous la forme d'une famille dérivationnelle unique peut porter à débat. Notons que *danseur* fait quant à lui l'objet de deux familles dérivationnelles, l'une basée sur le verbe *danser*, l'autre sur le nom *danse*.

Comme nous avons décidé d'utiliser une approche distributionnelle décrite dans la section 2.2. pour comparer sémantiquement les noms d'agent déverbaux en -*EUR*, les noms d'action déverbaux et les verbes dont ils dérivent, nous avons besoin de corpus qui permettent l'instanciation des lexèmes contenus dans Lexeur, et ce par le biais de corpus.

3.1.2. Les corpus

L'approche distributionnelle nécessite des données textuelles volumineuses pour porter ses fruits, comme nous l'avons vu dans la section 2.2. Deux corpus en particulier ont ainsi été choisis : le corpus Wiki et le corpus LM10.

Le premier et principal corpus, dont il sera question par la suite sous l'intitulé corpus Wiki, est un corpus constitué par Franck Sajous à partir du *dump* de la version française de l'encyclopédie en ligne Wikipédia en date du 28 mars 2013. Il regroupe plus de 765 346 articles, pour un total de 254 857 216 mots. Ce corpus a été choisi pour sa grande variété, tant dans les productions linguistiques que dans les thèmes et champs lexicaux proposés, mais aussi pour la grande quantité de données qu'il offrait.

Le corpus auquel il sera fait par la suite référence sous l'appellation corpus LM10 est un corpus constitué à partir des articles du journal français *Le Monde* publiés entre les années 1991 et 2000. Il contient près de 200 millions de mots. Ce corpus, qui fut longtemps l'un des corpus de référence pour le Traitement Automatique des Langues en français, est plus petit et spécifique que le corpus Wiki. Il permettra cependant d'envisager la stabilité de certains phénomènes par une comparaison de certaines données entre les deux corpus.

Pour chacun de ces corpus, la fréquence brute de chacun des mots a été calculée.

3.2. Enrichissement des données

Sur la base de Lexeur et des corpus à notre disposition, il nous faut maintenant, selon l'approche que nous avons choisie, analyser la distribution des noms d'agent et des noms d'action. Nous allons pour cela enrichir les ressources que nous avons. De nombreux outils existent pour enrichir des données en vue d'une analyse distributionnelle. Le choix a été fait d'un outil prêt à l'emploi, Word2Vec, que nous avons présenté dans la section 2.3, et dont nous présentons les enrichissements réalisés ci-dessous.

3.2.1. Enrichissement de Lexeur à l'aide de Word2Vec

Pour créer les matrices utilisées dans notre étude, les paramètres par défaut ont été conservés. L'architecture par défaut est le Skip-gram, l'algorithme d'apprentissage par défaut est le *hierarchical softmax*, le seuil de fréquence maximale pour le sous-échantillonnage aléatoire (ou décimation) des mots fréquents est fixé à 1e-3, le nombre d'itérations est de 5, le nombre de tâches se déroulant en

parallèle est fixé à 12, le nombre minimum d'occurrence des mots est fixé à 5, le nombre de mots qui peuvent être sautés dans le cadre du skip-gram est de 5, et le nombre de dimensions des vecteurs est fixé à 100. Les frontières de mots sont par défaut l'espace, la tabulation et l'EOL (délimiteur de fin de ligne).

À l'aide des matrices créées et des différents modules à notre disposition, nous allons enrichir les données de Lexeur avec les informations distributionnelles que nous fournit Word2Vec. Cet enrichissement va se faire en deux étapes. Word2Vec prend donc en entrée un corpus pour s'entraîner et créer une matrice (ou modèle). Une fois le modèle créé, il suffit de l'interroger. L'entrée varie en fonction du module : un mot pour le simple calcul des voisins et de leurs distances, un syntagme si l'on veut chercher des expressions plus complexes, des triplets de mots si l'on cherche à produire une analogie… Deux types de données ont ainsi pu être obtenus à partir des différents modules : une liste des voisins distributionnels de chaque lexème de Lexeur, et une liste de triplets nom d'agent/verbe/nom d'action.

3.2.2. Triplets nom d'agent/verbe/nom d'action

L'utilisation de Word2Vec permet d'enrichir une liste de triplets nom d'agent/verbe/nom d'action constitués à l'aide de Lexeur avec des indices de proximité liant les éléments deux à deux. Puisque l'on s'intéresse tout particulièrement à la proximité entre le nom d'agent et le verbe d'une part et le nom d'action et le verbe d'autre part, former ainsi des triplets nous permet de constituer un cadre de travail plus adapté que les familles dérivationnelles complètes telles que présentées dans Lexeur.

Il n'est pas directement possible de comparer ensemble deux mots à l'aide de Word2Vec. On peut soit interroger la matrice pour un mot, dans le cadre du module *distance*, soit interroger la matrice sur la base de trois mots, dans le cadre du module *analogy*, mais il n'est pas possible en l'état d'entrer nos deux mots directement. Il nous faut pour cela un programme. À l'aide d'un programme fourni par Nabil Hathout, chaque ligne de Lexeur est parcourue, afin d'extraire de chaque famille dérivationnelle les différents triplets nom d'agent/verbe/nom d'action possibles. Au sein de chacun de ses triplets, les trois couples ainsi possibles (nom d'agent/verbe, nom d'agent/nom d'action, et verbe/nom d'action) sont interrogés via le module *similarity* de Gensim (module pas directement interrogeable, mais qui peut être intégré à des programmes python), qui fournit la distance séparant les deux éléments proposés. Ainsi, l'exemple (42a) nous montre une famille dérivationnelle issue de la ressource Lexeur. L'exemple (42b) liste les différents triplets nom d'agent/verbe/nom d'action que l'on peut théoriquement former à partir de la famille (42a). L'exemple (42c) montre quant à lui les triplets conservés suite au traitement du programme appliqué à la matrice créée à partir du corpus Wiki.

(42a) *éventreur – éventreuse – éventrer – éventrage – éventrement – éventration*

(42b) *éventreur – éventrer – éventrage*
éventreur – éventrer – éventrement
éventreur – éventrer – éventration
éventreuse – éventrer – éventrage
éventreuse – éventrer – éventrement
éventreuse – éventrer – éventration

(42c) *éventreur – éventrer – éventration*

Pour une même famille dérivationnelle, on peut donc obtenir un nombre variable de triplets, en fonction du nombre de noms d'action appartenant à la famille. Certaines familles peuvent ainsi être représentées par plusieurs triplets, ce qui n'est pas le cas pour notre famille (42a) dont on observe que seul un triplet est finalement conservé en (42c). Les cinq autres triplets théoriques n'ont donc pas été ramenés par le programme. Cela signifie qu'un ou plusieurs des lexèmes des triplets non ramenés n'étaient pas représentés dans la matrice Word2Vec, soit parce qu'ils n'étaient pas assez fréquents dans le corpus, soit parce qu'ils n'étaient pas présents du tout dans le corpus. Ainsi, si les tokens *éventreur*, *éventrer* et *éventration* ont respectivement une fréquence brute de 275, 197 et 26 dans le corpus Wiki, les tokens *éventrage* et *éventrement* n'apparaissent quant à eux que 1 et 4 fois (moins que le nombre d'occurrences minimum par défaut de 5). Il n'y a, de plus, aucune occurrence du lexème *éventreuse* dans le corpus Wiki. Cela explique donc que seul un des six triplets théoriques ait été conservé.

On obtient donc en sortie une liste de triplets nom d'agent/verbe/nom d'action accompagnés des indices de proximité liant deux à deux les trois entités du triplet, comme le montrent les exemples dans le tableau 2. Comme cela a été décrit en 2.2., l'indice de proximité est obtenu à l'aide du calcul du cosinus des vecteurs, et est donc compris entre 0 et 1, 1 signifiant la grande proximité des deux éléments comparés.

Nom d'agent	Verbe	Nom d'action	Indice AgVb	Indice AgAc	Indice VbAc
directeur	diriger	direction	0,44391475	0,51821282	0,26974376
directeur	diriger	directoire	0,44391475	0,35822141	0,13566181
discriminateur	discriminer	discrimination	0,14538923	0,15078802	0,48874008
diseur	dire	dire	0,07516962	0,07516962	1

Tableau 2 *– Exemple de quatre entrées de la liste de triplets nom d'agent/verbe/nom d'action*

Le tableau 2 nous montre quatre exemples de triplets, dont deux issus d'une même famille et partageant le même nom d'agent (*directeur*/*diriger*) et deux autres triplets comme uniques

représentants des familles dont ils sont issus. Les indices de proximité entre le nom d'agent et le verbe AgVb, entre le nom d'agent et le nom d'action AgAc, et le verbe et le nom d'action VbAc seront par la suite respectivement désignés sous les notations *iAgVb*, *iAgAc* et *iVbAc*. Notons que si cet indice n'est jamais égal à 0, il est à plusieurs reprises égal à 1, à l'image du triplet *diseur*/*dire*/*dire*. On retrouve ainsi dans la liste des triplets projetée sur la matrice créée à partir du corpus Wiki 22 triplets pour lesquels un des indices est égal à 1. Dans la totalité de ces cas, l'indice égal à 1 est celui impliquant le verbe et le nom d'action, ce qui signifie que ces verbes et ces noms d'action sont deux à deux distributionnellement identiques.

Nom d'agent	Verbe	Nom d'action	Indice AgVb	Indice AgAc	Indice VbAc
baiseur	baiser	baiser	0.22172747	0.22172747	1
goûteur	goûter	goûter	0,22766364	0,22766364	1
lanceur	lancer	lancer	0.26520752	0.26520752	1
toucheur	toucher	toucher	0,22935323	0,22935323	1

Tableau 3 *– Exemples issus de la liste de triplets nom d'agent/verbe/nom d'action*

Le tableau 3 illustre d'autres cas de triplets pour lesquels *iVbAc* (l'indice de proximité entre le verbe et le nom d'action) est égal à 1. On remarque que le verbe et le nom d'action sont tous graphiquement identiques, bien qu'il s'agisse dans tous ces cas respectivement d'un verbe et d'un nom d'action bien distinct. Cette confusion est une conséquence du fonctionnement de Word2Vec dont nous parlions dans la section 3.2.1., et qui n'intègre pas d'analyse morphosyntaxique. Cela se traduit par deux mots distincts considérés comme un seul et même mot dans la matrice (puisque le cosinus de leur vecteur est égal à 1, leurs vecteurs sont strictement identiques). Cela réduit donc notre triplet à un couple de mots, ce qui explique que *iAgVb* et *iAgAc* (les indices de proximité entre le nom d'agent et le verbe d'une part, et le nom d'agent et le nom d'action d'autre part) soient identiques. Puisque le verbe et le nom d'action sont identiques, on ne compare plus le nom d'agent qu'à un seul et même élément. Bien que minoritaires, ce genre de cas nous montre qu'il faut analyser avec un certain recul les résultats fournis par Word2Vec.

On obtient ainsi une liste de 1 945 triplets et leurs indices de proximité associés. Notons que sur ces 1945 triplets, 11 sont des doublons, à l'image des triplets *édificateur*/*édifier*/*édification* ou *gaffeur*/*gaffer*/*gaffe* que l'on retrouve donc en double dans la liste produite par le programme. Cela s'explique par la présence en double de certaines familles dérivationnelles dans Lexeur, comme cela a été évoqué dans la section 3.1.1. Ces doublons n'étant cependant par la suite traités qu'une seule fois, cela réduit donc notre liste effective à 1 934 triplets. Si l'on rapporte ce nombre de triplets au nombre de familles contenues dans Lexeur (5 974, cf 3.1.1.), on constate que moins d'un tiers des familles se voient représentées par un triplet. Mais la répartition des triplets par famille est loin d'être homogène,

et l'on observe de grandes disparités, comme le montre le tableau 4 ci-dessous. Sont comptabilisés dans ce tableau les 11 doublons.

Nombre de triplets pour un même verbe	1	2	3	4	5	6	8	9	10	12
Nombre de cas	664	329	39	61	8	17	6	2	3	2

Tableau 4 *– Nombre de cas en fonction du nombre de triplets issus d'une même famille dérivationnelle*

Ce tableau comptabilise le nombre de triplets issus d'une même famille dérivationnelle, et a été obtenu en comparant le nombre de triplets partageant un même verbe. C'est en cela une simplification, puisqu'un verbe peut n'appartenir qu'à une seule famille dérivationnelle dans Lexeur, mais se retrouver dans deux réseaux sémantiques distincts, tout comme deux familles dérivationnelles de Lexeur peuvent se partager un même verbe, comme nous l'avons vu dans la section 1.4. et 3.1.1. Cette hypothèse nous permet cependant d'obtenir un aperçu général de la répartition des triplets.

Pour chaque verbe identifié, il a été compté le nombre de fois où il apparaissait dans un triplet : on obtient ainsi une fourchette allant de 1 à 12, un même verbe pouvant apparaître dans un triplet, deux triplets, et ainsi de suite jusqu'à 12 triplets. Cela correspond à la première ligne du tableau 3. Puis a été quantifié le nombre de verbes appartenant à chaque catégorie : 668 verbes apparaissent ainsi dans un unique triplet, 328 verbes apparaissent dans deux triplets différents, et ainsi de suite. Cela correspond à la deuxième ligne du tableau 3.

Ainsi, si la majorité des familles dérivationnelles sont représentées par un ou deux triplets seulement, certaines familles se distinguent par le nombre plus important de triplets qu'elles produisent, à l'image des familles *batteur/batteuse/battre* et *porteur/porteuse/porter* qui produisent chacune 12 triplets. Au total, c'est donc près de 1130 familles dérivationnelles, soit un cinquième de Lexeur (plus, si l'on garde en tête la simplification faite pour ce calcul) qui sont ainsi représentées, dont 462 par plusieurs triplets, ce qui laisse quelques 4800 familles exclues de cette première analyse à cause de l'absence de certains de leurs constituants dans la matrice créée par Word2Vec. Il faudra donc garder en tête dans les manipulations à venir dans la section 4. que les résultats obtenus ne concernent que l'échantillon représenté par notre liste, et pas l'ensemble de Lexeur. Notons par ailleurs que cette liste a été constituée sur la base de Lexeur dans son état initial, à savoir non nettoyé des défauts de constitution que nous avions mis en avant plus tôt.

3.2.3. Voisins distributionnels

Comme nous l'avons vu en 3.2.1., le module *distance* permet d'obtenir pour un mot cible donné une liste de ses voisins les plus proches dans la matrice créée par Word2Vec. À ces voisins sont aussi associées les distances séparant chaque voisin du mot cible initial. Puisque nous penchions pour une approche distributionnelle dans la section 2.3., il nous faudra nous intéresser aux voisins distributionnels de nos noms d'agent, verbes et noms d'action.

Le module nous permet donc de les obtenir. La commande

./distance model.bin

nous permet d'interroger la matrice (ou modèle) voulue. L'utilisateur entre alors le mot cible dont il souhaite obtenir les plus proches voisins. Le nombre de voisins renvoyés par Word2Vec peut être paramétré, et est par défaut réglé à 40. L'exemple (43) nous montre l'*output* fourni par Word2Vec pour le mot *sauveur.*

(43) *Enter word or sentence (EXIT to break): sauveur*

Word: sauveur Position in vocabulary: 10851

Word	*Cosine distance*
rédempteur	*0.732513*
consolateur	*0.689293*
intercesseur	*0.649364*
christ	*0.648511*
jésus	*0.637197*
ressuscité	*0.632098*
dieu	*0.615086*
bénissant	*0.607997*
prophète	*0.604830*
miséricordieux	*0.601006*
repentant	*0.596062*
paphnuce	*0.593738*
libaire	*0.591662*
budoc	*0.591088*
marie-madeleine	*0.588178*
séraphin	*0.587334*
bienheureuse	*0.587191*
gandolfi-scheit	*0.584788*
miséricorde	*0.582564*
serviteurs	*0.581109*
thaumaturge	*0.580844*
bénisse	*0.576613*
guénolé	*0.575485*
méryem	*0.575249*
euthyme	*0.574720*

uriel	*0.574048*
oint	*0.573454*
apôtre	*0.573012*
notre-seigneur	*0.569502*
martyre	*0.566768*
nade	*0.566759*
cénéré	*0.565242*
bienheureux	*0.565096*
énimie	*0.564416*
ange	*0.563738*
sépulcre	*0.563667*
trinité	*0.562653*
pécheur	*0.561991*
recrucifier	*0.559657*
restitute	*0.558452*

Dans l'exemple (43), on peut ainsi constater que le mot *rédempteur* est le voisin distributionnel le plus proche de *sauveur* dans la matrice créée par Word2Vec à partir du corpus Wiki. Le cosinus séparant les vecteurs de *sauveur* et de *rédempteur* a une valeur de 0,732513, valeur relativement proche de 1 signifiant que ces deux voisins sont donc relativement similaires sur le plan sémantique. Ils sont du moins plus similaires que *sauveur* et *pécheur*, dont le cosinus a une valeur de 0,561991. On remarque qu'une grande partie des voisins proposés renvoient à une signification religieuse du mot *sauveur*, mais l'on constate la présence de noms plus « obscurs » dont on comprend moins instinctivement le rapprochement, comme le mot anglais *restitute*, ou les mots *gandolfi-scheit* ou *paphnuce*.

Lorsqu'un mot n'est pas représenté dans la matrice, pour les raisons vues dans les sections 2.3. et 3.2.1., l'output n'affiche aucun voisin, et l'encart *Position in the vocabulary* affiche une valeur de -1 ainsi que la phrase *Out of dictionary word!*.

La commande ne permettant d'interroger la matrice que pour un seul mot à la fois, et puisque l'on souhaitait obtenir les voisins distributionnels de tous les lexèmes de Lexeur, un programme a été conçu pour automatiser l'interrogation de la matrice via le module *distance* et ce afin ne pas répéter manuellement l'opération plus de 5974 fois minimum. Le code de ce module a été modifié afin de rendre l'*output* plus facilement traitable par ce programme. Le nombre de voisins renvoyés par le module a aussi été modifié, passant de 40 à 500 afin d'obtenir un plus large spectre de voisins. Ce dernier parcourt chaque ligne de Lexeur, et pour chaque famille dérivationnelle, il récupère les différents lexèmes. La matrice est alors interrogée tour à tour pour chacun de ces lexèmes. Une fois les voisins ainsi que l'indice de proximité récupérés pour l'ensemble des lexèmes d'une famille dérivationnelle, la ligne suivante de Lexeur est à son tour parcourue, et ainsi de suite.

On obtient ainsi un fichier au format texte comportant quatre colonnes : une première colonne indiquant les mots cible, une seconde colonne indiquant le rang du voisin, une troisième colonne affichant les voisins, et une dernière colonne indiquant la distance entre les deux premiers éléments, à l'image de l'exemple (44).

(44) *absoluteur* *0*

absoudre	*1*	*excommunier*	*0.750697*
absoudre	*2*	*expier*	*0.699005*
absoudre	*3*	*calomnier*	*0.688371*
absoudre	*4*	*offenser*	*0.680091*
absoudre	*5*	*confesser*	*0.676277*

L'exemple (44) nous indique donc que *absoluteur* n'est pas dans la matrice, et que *absoudre* a comme voisin le plus proche *excommunier*, et comme deuxième voisin le plus proche *expier*, et ainsi de suite. Notons que certains mots cible n'ont pas de voisins, et sont directement marqués d'un 0 dans la deuxième colonne : il s'agit de mots non représentés dans la matrice créée par Word2Vec, et initialement marqués d'un -1 par le module. Les listes obtenues sur la base du corpus Wiki et du corpus LM10 ne sont donc pas identiques, puisque les lexèmes de Lexeur n'y sont pas présents de la même façon dans les deux matrices créées.

Pour espérer mieux exploiter les voisins distributionnels de Lexeur, le module de distance de Word2Vec a aussi été intégré à un autre programme dans le but d'obtenir une liste des voisins communs aux éléments d'une même entrée de Lexeur, que l'on considère pour le moment comme une famille dérivationnelle unique, malgré les nuances qui ont pu être mises en évidence dans les sections 1.4. sur la notion de réseau secondaire et 3.1.1. sur la présentation de Lexeur. L'idée est en effet de voir si les voisins distributionnels que partageraient un nom d'agent et un verbe ou un verbe et un nom d'action pourraient nous orienter quant à leur sémantisme, et expliquer des indices de proximité plus ou moins importants. Pour cela, on parcourt chaque ligne de Lexeur, et comme précédemment, on segmente chaque ligne, pour obtenir les différents lexèmes composant la famille dérivationnelle. On associe alors à chaque lexème isolé les voisins distributionnels (s'ils en ont) fournis par Word2Vec via la matrice créée à partir du corpus voulu. Une fois tous les lexèmes de la famille dérivationnelle complétés de leurs potentiels voisins, on les compare et on affiche par ordre lexicographique chaque voisin distributionnel issu de Word2Vec ainsi que les mots issus de Lexeur auxquels il a été associé. Une fois la ligne de Lexeur ainsi traitée, on passe à la suivante. On obtient donc un fichier texte tabulé à l'image de l'exemple (45b) pour la famille montrée en exemple en (45a).

(45a) *abatteur – abatteuse – abattre – abat – abattement – abatture – abattage - abattis*

(45b)

exercice	*abattement/abattage/*
expropriation	*abattement/abattage/*
fiente	*abat/abattage/*
foin	*abat/abattage/abattis/*

Dans l'exemple (45b), *exercice* et *expropriation* sont donc deux voisins distributionnels (dans la représentation de notre corpus Wiki créée par Word2Vec) que l'on retrouve parmi les 500 voisins les plus proches d'*abattement* et d'*abattage*. De même, *fiente* se retrouve dans les 500 plus proches voisins d'*abat* et d'*abattage*. Le mot *foin*, quant à lui, se retrouve dans les 500 voisins les plus proches de trois lexèmes de la famille présentée (en 45a), *abat*, *abattage* et *abattis*. Le traitement permet de préciser le rang de chacun de ces voisins pour chaque mot cible (à savoir, indiquer quel est le rang de *exercice* parmi les voisins de *abattement* et de *abattage*), mais cela n'a pas été fait ici.

3.3. Démarche

Pour ce mémoire, nous cherchons à montrer que le nom d'action est sémantiquement plus proche du verbe dont il est issu que ne l'est le nom d'agent issu du même verbe. Pour cela, nous allons chercher à montrer que les noms d'action sont distributionnellement plus proches des verbes que ne le sont les noms d'agent, puisque l'on considère que la distribution peut être indice de compréhension du sémantisme des mots, comme vu dans la section 2.

Nous allons dans un premier temps nous pencher sur les triplets enrichis obtenus et décrits dans la section 3.2.2. Nous allons tout d'abord observer ces triplets de façon générale, pour voir si une tendance générale peut se dégager des indices de proximité dont sont enrichies nos données. Nous allons ensuite essayer de décrire certains de ces triplets pour espérer pouvoir tirer quelques observations générales. Pour ce faire, nous allons essayer de décrire selon les caractéristiques mises au jour dans la section 1. les noms d'agent et les noms d'action. Enfin, nous observerons les noms d'agent et noms d'action au sein de leur famille dérivationnelle complète, en nous penchant non plus seulement sur l'indice de proximité mais aussi sur les voisins distributionnels. Les résultats issus de la matrice n'étant pas très transparents, nous allons nous aider d'un concordancier pour étudier en contexte certains mots.

L'utilisation de tests mis en avant dans la section 2.3. se fait sur la base de l'intuition. Lorsqu'un mot n'est pas connu, nous nous servons de la définition fournie par le TLFi pour nourrir notre intuition. Nous avons recours au concordancier AntConc lorsque nous souhaitons connaître les contextes d'apparition d'un mot.

4. Mise à l'épreuve de l'hypothèse

Les prochaines sections relatent le travail d'analyse du lexique Lexeur enrichi des informations distributionnelles fournies par Word2Vec par le biais des matrices créées à partir des corpus Wiki et LM. Le but est d'analyser la proximité sémantique des noms d'action, noms d'agent et verbe d'une même famille dérivationnelle, afin de tester l'hypothèse selon laquelle le nom d'action et le verbe sont sémantiquement plus proches que le nom d'agent et le verbe ne le sont.

Les analyses se portent dans un premier temps sur les données obtenues à partir du corpus Wiki. Le recours au corpus LM10 servira à confirmer ou non les phénomènes observés, comme cela a été signifié dans la section 3.1.2.

4.1. Fréquence et proximité distributionnelle

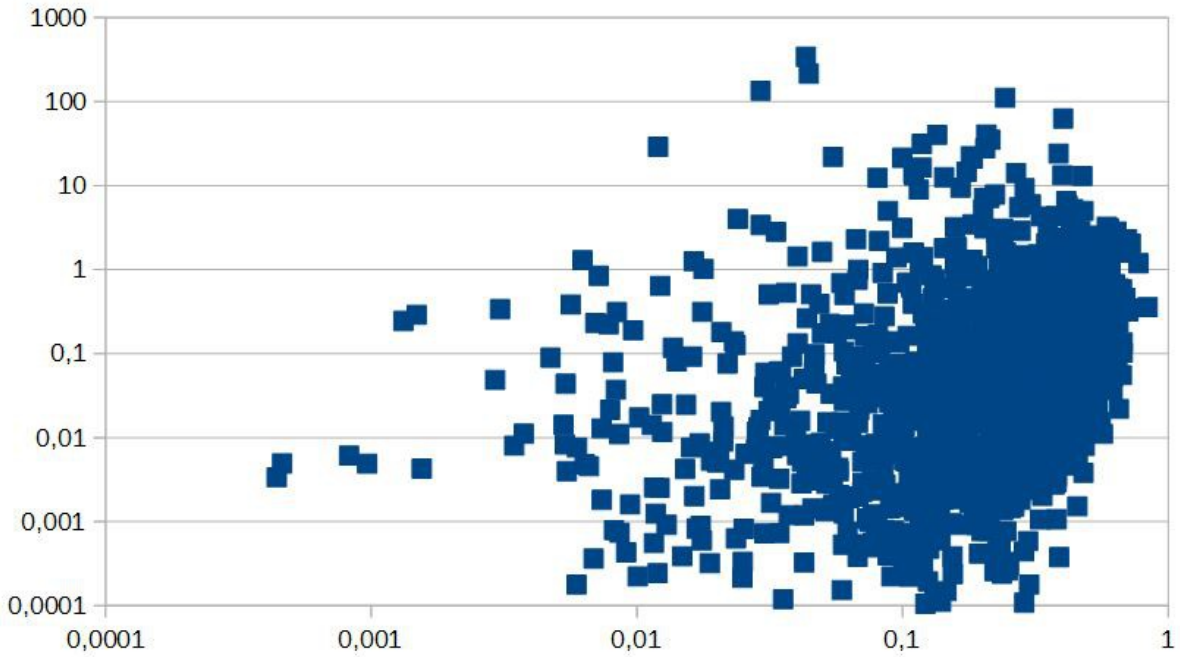
La fréquence est un facteur non négligeable dans la représentation vectorielle dans la matrice du sens des mots. En effet, une observation rapide de triplets avec de faibles indices de proximité comme les exemples illustrés dans le tableau 5 pousse à s'interroger sur la fréquence des mots impliqués.

Agent (fréquence)	Verbe (fréquence)	Action (fréquence)	Indice AgVb	Indice AgAc	Indice VbAc
testeur (160)	tester (4478)	testage (7)	0,346967466067	0,0944171824451	0,284142865058
jureur (50)	jurer (2620)	jurement (7)	0,320609463291	0,00442742433384	0,225733667503
concentrateur (51)	concentrer (8680)	concentration (1061)	0,114683771892	0,00983731922934	0,169598466011

Tableau 5 *– Exemples de triplets accompagnés de leurs indices de proximité et des fréquences brutes des membres des triplets*

Le tableau 5 regroupe trois triplets accompagnés de leurs indices de proximité dans la matrice créée sur la base du corpus Wiki. Pour chaque élément du triplet a également été ajouté la fréquence brute de l'élément dans le corpus Wiki. On observe que pour les trois indices *iAgAc* (indices les plus faibles ici), le nom d'agent ou le nom d'action associé présente une fréquence brute faible. Ainsi, pour le triplet *testeur*/*tester*/*testage*, dont le *iAgAc* est égal à 0,0944171824451, on remarque que le nom d'action *testage* a une fréquence brute de 7. Il en est de même pour le triplet *jureur*/*jurer*/*jurement*. Cela est d'autant plus marquant pour le triplet

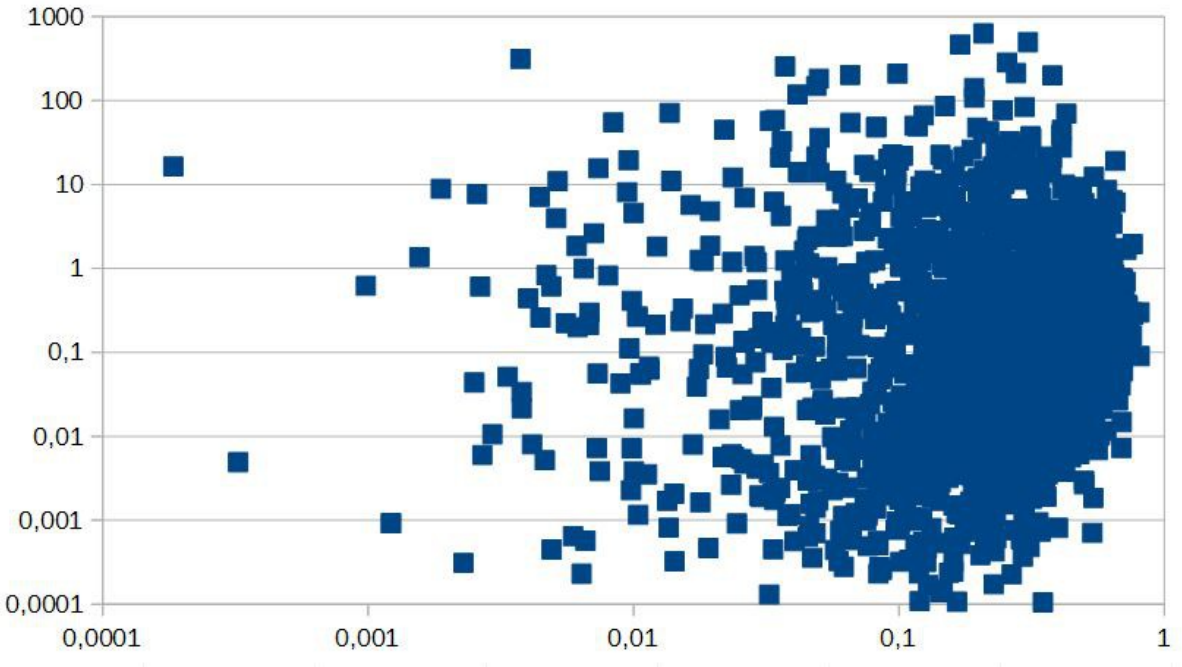
concentrateur/*concentrer*/*concentration* au vu de l'écart entre la fréquence brute du nom d'agent, qui est de 51, et les fréquences brutes du verbe et du nom d'action, respectivement de 8680 et 1061.



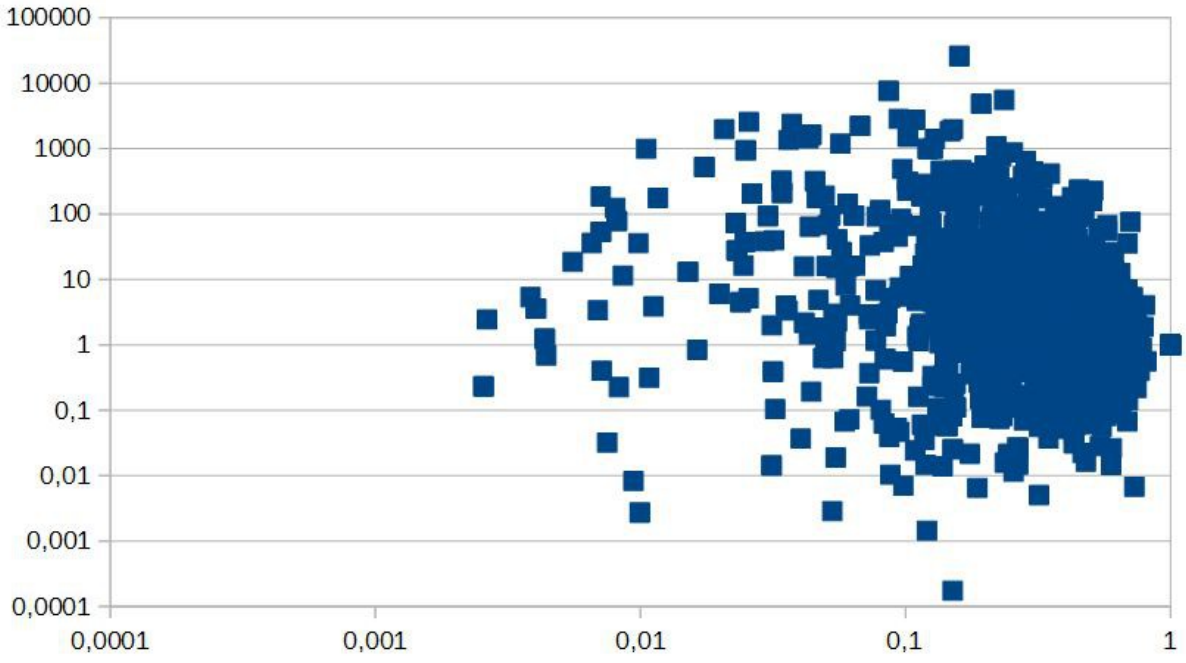
Graphique 1 *– Corrélation entre l'indice de proximité iAgVb et le rapport des fréquences brutes du nom d'agent et du verbe dans le corpus Wiki*

Le graphique 1 montre la corrélation entre l'indice de proximité *iAgVb* entre le nom d'agent et le verbe (sur l'axe des abscisses, en échelle logarithmique) et le rapport des fréquences brutes du nom d'agent et du verbe (sur l'axe des ordonnées, en échelle logarithmique). Si aucune tendance ne se dessine clairement, le coefficient de Pearson est d'une valeur de -0,03762, ce qui indique qu'il n'y a pas de corrélation entre l'indice de proximité *iAgVb* et le rapport des fréquences du nom d'agent et du verbe.

Le graphique 2 représente cette même corrélation mais pour l'indice de proximité *iAgAc* et le rapport des fréquences brutes du nom d'agent et du nom d'action. Or, la tendance est encore moins nette dans le graphique 2 que dans le graphique 1. Le coefficient de Pearson est ici de -0,0727873, ce qui indique qu'il n'y a pas de corrélation selon la formule de Pearson entre l'indice de proximité *iAgAc* et le rapport des fréquences du nom d'agent et du nom d'action. Enfin, la corrélation avec le rapport des fréquences brutes du verbe et du nom d'action est représentée dans le graphique 3. Le coefficient de Pearson est ici de -01118979, ce qui est légèrement plus marqué que pour les deux autres cas, mais cela semble indiquer qu'il n'y a pas de corrélation entre l'indice *iVbAc* et le rapport des fréquences du verbe et du nom d'action.



Graphique 2 *– Corrélation entre l'indice de proximité iAgAc et le rapport des fréquences brutes du nom d'agent et du nom d'action dans le corpus Wiki*



Graphique 3 *– Corrélation entre l'indice de proximité iVbAc et le rapport des fréquences brutes du verbe et du nom d'action dans le corpus Wiki*

4.2. Caractérisation des triplets nom d'agent/verbe/nom d'action

La première étape consiste à exploiter la liste des triplets nom d'agent/verbe/nom d'action obtenus précédemment, accompagnés des indices de proximité séparant tour à tour le nom d'agent et le verbe, le nom d'agent et le nom d'action, et le verbe et le nom d'action à l'image des triplets illustrés dans le tableau 2 de la section 3.2.2.2.

Le choix de manipuler en premier lieu ces données est justifié par le cœur même de ce mémoire : puisque nous cherchons à comparer la proximité sémantique liant le nom d'agent, le verbe et le nom d'action issus d'une même famille dérivationnelle, prendre trois à trois ces éléments semble donc plus pertinent. Cela permet par ailleurs de comparer plusieurs triplets qui seraient issus d'une même famille dérivationnelle, à savoir les cas où il existe une certaine concurrence entre plusieurs noms d'agent ou plusieurs noms d'action dérivés d'un même verbe. Ces triplets se présentent pour rappel sous la forme illustrée dans le tableau 2.

Nom d'agent	Verbe	Nom d'action	Indice AgVb	Indice AgAc	Indice VbAc
directeur	diriger	direction	0,44391475	0,51821282	0,26974376
directeur	diriger	directoire	0,44391475	0,35822141	0,13566181
discriminateur	discriminer	discrimination	0,14538923	0,15078802	0,48874008
diseur	dire	dire	0,07516962	0,07516962	1

Tableau 2 *– Exemple de quatre entrées de la liste de triplets nom d'agent/verbe/nom d'action*

Avant de dégager des tendances générales, nous considérons ces quatre entrées illustrées dans le tableau 2. Nous en avons discuté dans la section 3.2.2.2., certaines valeurs n'ont que peu d'intérêt pour nous. En effet, tout indice égal à 1 correspond à une mauvaise représentation de deux mots distincts. Prenons le cas du triplet *diseur*/*dire*/*dire*. L'indice de proximité *iVbAc* entre le verbe et le nom d'action est ici de 1 car le verbe *dire* et le nom d'action *dire* sont vus par Word2Vec comme un seul et même mot. Or, il existe bien deux lexèmes *dire* distincts. Si les deux nous semblent effectivement proches selon notre intuition de locuteur, notre intuition nous empêche de dire qu'ils sont identiques. Le nom *dire* tend à prendre une certaine connotation juridique, dans des expressions comme *selon les dires des témoins*, ou du moins une connotation solennelle, que le verbe n'intègre absolument pas, bien au contraire, *dire* passant pour un des verbes de discours les plus généraux. Le TLFi nous confirme qu'il existe deux lexèmes, le verbe *dire*, qui désigne selon le TLFi le fait « [d'] énoncer un propos par la parole », et le nom masculin *dire*, qui désigne une « déclaration de témoin […], ce qu'une personne dit, avance, déclare ». De même, si les indices de proximité *iAgVb* et *iAgAc*

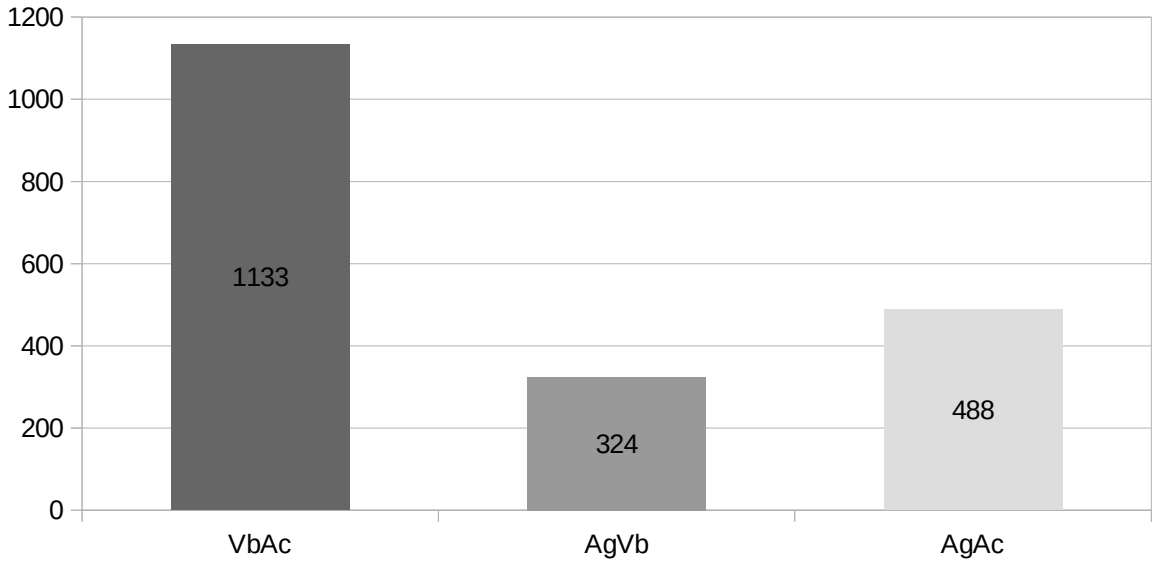
semblent indiquer que le nom d'agent *diseur* est aussi proche du verbe *dire* que du nom d'action *dire*, notre intuition va à l'encontre de ce résultat. De par la plus grande spécificité du nom d'action *dire* par rapport au verbe *dire*, on s'attend à ce que l'indice *iAgAc* soit plus faible que l'indice *iAgVb*, le nom d'agent *diseur* dénotant davantage le sens général dénoté par le verbe *dire* que le sens juridique ou solennel du nom d'action *dire*.

A contrario, si l'on se penche sur le triplet *discriminateur*/*discriminer*/*discrimination* du tableau 2, notre intuition de locuteur est davantage satisfaite. Dans ce triplet, l'indice *iVbAc* est le plus élevé, faisant sur le plan distributionnel du verbe *discriminer* et du nom *discrimination* les deux éléments les plus proches. Qu'il 'agisse d'une question de fréquence, *discriminateur* n'étant pas un mot aussi couramment utilisé et rencontré que le verbe *discriminer* ou le nom d'action *discrimination*, ou de la conséquence de l'hypothèse de Michel Roché concernant la conservation du sens du verbe lors de la dérivation, le nom d'agent nous semble moins proche des deux autres mots du triplet. À l'image de l'analyse du nom *dire*, le nom *discrimination* et le verbe *discriminer* partagent davantage une connotation péjorative que le terme *discriminateur*, qui semble plus général. Les proches valeurs de *iAgVb* et *iAgAc* (respectivement 0,14538923 et 0,15078802) confirment notre idée que *discriminateur* est finalement aussi proche du verbe *discriminer* que du nom *discrimination*.

Nous pourrions manuellement faire des observations semblables à celles faites sur *diseur*/*dire*/*dire* et *discriminateur*/*discriminer*/*discrimination* sur un plus grand nombre de triplets, afin de nous faire une idée plus globale du comportement des triplets et par extension des noms d'agent et noms d'action, mais cela serait fastidieux. Nous avons cherché donc des indices observables à grande échelle.

La première étape de cette observation a été de voir quels étaient en moyenne les deux éléments les plus proches au sein des triplets nom d'agent/verbe/nom d'action. Pour cela, nous avons comparé les trois distances séparant deux à deux les éléments des triplets pour voir laquelle de ces trois distances était est la plus grande (et donc la plus proche de 1). Cela nous permet de déterminer pour chaque triplet quelle était la distance prédominante et donc le couple le plus sémantiquement similaire selon le critère distributionnel : le couple nom d'agent/verbe, le couple nom d'agent/nom d'action ou le couple verbe/nom d'action.

Cette première caractérisation s'est faite à partir de la liste de triplets obtenue dans la section 3.2.2.2. La liste a été parcourue à l'aide d'un programme, et pour chaque triplet rencontré, les trois indices de proximité *iAgVb*, *iAgAc* et *iVbAc* ont été comparés. Lorsque l'indice de proximité *iAgVb* était strictement supérieur aux deux autres indices (et donc potentiellement le plus proche de 1), le triplet s'est vu annoté d'un AgVb. Lorsque l'indice de proximité *iAgAc* était strictement supérieur aux deux autres indices, le triplet a été annoté AgAc. Enfin, lorsque l'indice de proximité *iVbAc* était strictement supérieur aux deux autres, le triplet a été annoté VbAc. Une observation des données a montré qu'aucun triplet ne dérogeait à ces trois inégalités strictes. On obtient ainsi trois groupes rassemblant les triplets labellisés en fonction de l'indice de proximité le plus élevé, et notés AgVb, AgAc et VbAc.

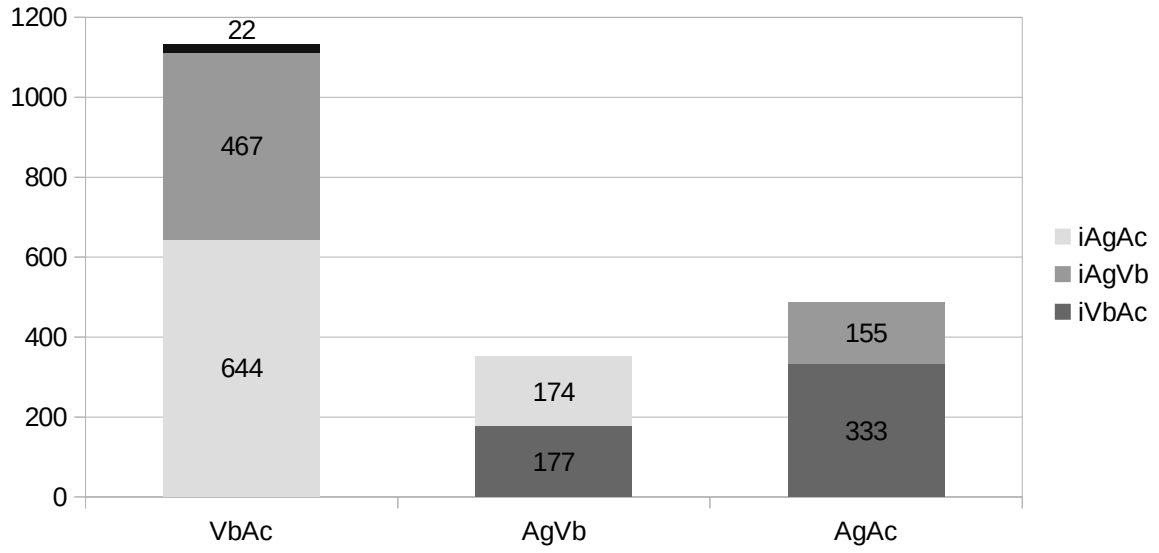


Graphique 4 *– Répartition des triplets nom d'agent/verbe/nom d'action issus de Lexeur en fonction de leur plus grand indice de proximité dans le modèle Word2Vec construit sur le corpus Wiki*

Le graphique 4 nous montre la répartition des triplets en fonction des labels qui leur ont été attribués. Ainsi, on observe que sur les 1945 triplets obtenus à partir de la matrice créée sur la base du corpus Wiki, 1133 d'entre eux sont annotés VbAc, ce qui signifie que ces 1133 triplets (soit 58 % des triplets avant le nettoyage des doublons) ont un indice *iVbAc* strictement supérieur aux deux autres indices. De même, 488 triplets sont annotés AgAc, soit 25 % des triplets qui ont un indice *iAgAc* strictement supérieur aux deux autres indices. Enfin, seuls 324 triplets sur 1945, soit 17 % des triplets, sont annotés AgVb, et ont donc un indice *iAgVb* strictement supérieur aux deux autres indices. Dans plus de la moitié des triplets, le verbe et le nom d'action ont un indice de proximité supérieur à celui entre le nom d'agent et le verbe et le nom d'agent et le nom d'action. Dans plus de 50 % des triplets, le verbe et le nom d'action sont donc, sur le plan distributionnel, les plus proches, ce que nous considérons comme un indice de leur proximité sémantique. *A contrario*, l'indice de proximité *iAgVb* est le plus élevé dans seulement 17 % des triplets. Le nom d'agent et le verbe sont donc les deux éléments du triplet distributionnellement les plus proches dans moins d'un cinquième des triplets. Le verbe a donc tendance à être plus proche du nom d'action que du nom d'agent, ce qui va dans le sens de l'hypothèse que nous avons formulée dans la section 1, à savoir que, du fait de la non activation de l'opération sémantique, le verbe et le nom d'action sont sémantiquement identiques, contrairement au nom d'agent qui subissait une modification sémantique lors du processus de dérivation. Notons que l'indice de proximité *iAgAc* a tendance à être plus souvent supérieur à l'indice *iAgVb* que le contraire, avec plus de 25 % de triplets annotés AgAc, contre 17 % de triplets annotés AgVb. Le nom d'agent a

donc tendance à être plus proche du nom d'action que du verbe, ce qui va dans le sens d'une certaine préférence catégorielle.

Si l'on va plus loin dans cette observation, on peut distinguer au sein de chaque tendance deux comportements distincts en fonction de la deuxième distance la plus grande.

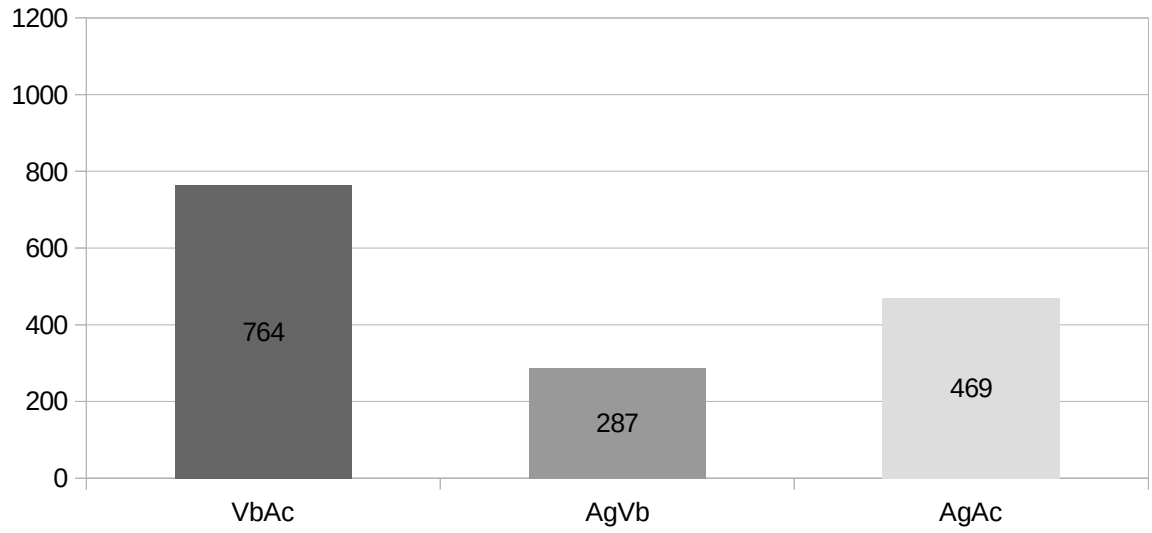


Graphique 5 *– Répartition des triplets nom d'agent/verbe/nom d'action issus de Lexeur au sein des trois tendances mises au jour en fonction du second plus grand indice de proximité dans le modèle Wiki*

Le graphique 5 présente la répartition des triplets au sein de trois tendances mises au jour précédemment, en fonction du deuxième couple le plus proche, donc en se basant sur le deuxième indice de proximité le plus élevé. Ce graphique reprend les labels employés précédemment. Ainsi, parmi les 1125 triplets précédemment annotés VbAc, 644 d'entre eux ont comme deuxième indice de proximité le plus grand *iAgAc* (après *iVbAc*). Pour ces triplets-là, on a donc l'inégalité *iVbAc* > *iAgAc* > *iVbAc*. Toujours au sein des triplets annotés VbAc, 467 d'entre eux répondent quant à eux à l'inégalité *iVbAc*>*iAgVb*>*iAgAc*. Enfin, les 22 triplets restants parmi les triplets annotés VbAc sont des triplets pour lesquels *iAgVb* et *iAgAc* sont égaux. Ils correspondent aussi aux triplets où *iVbAc* est égal à 1, à savoir les triplets comme ceux illustrés dans le tableau 3 où le verbe et le nom sont un même mot-type. Au sein des triplets annotés VbAc, ceux pour lesquels l'indice *iAgVb* est le plus faible des trois sont majoritaires, à près de 58 %. Dans ces triplets, le nom d'agent et le verbe sont les deux éléments les plus distants sur le plan distributionnel. Cela s'observe aussi parmi les triplets annotés AgAc, où seuls 155 triplets ont pour deuxième indice de proximité le plus élevé l'indice *iAgVb*. Dans plus de 68 % des triplets annotés AgAc, le nom d'agent et le verbe sont encore une fois les deux éléments les plus distants distributionnellement. Sur les deux niveaux d'observation étudiés, on remarque donc que le nom d'action et le verbe ont tendance à être les deux éléments les plus proches

au sein d'un triplet, et le nom d'agent et le verbe les deux éléments les plus distants.

Toutes les observations que nous venons de faire s'appliquent dans le cadre du corpus Wiki. Puisque nous avons à notre disposition un autre corpus, le corpus LM, nous avons fait les mêmes manipulations sur la base de la matrice créée à l'aide de ce second corpus afin de voir si les résultats que l'on observait étaient stables ou si des variations notables apparaissaient. Nous avons donc annoté de la même façon que précédemment les 1520 triplets obtenus selon le procédé décrit dans la section 3.2.2.2. mais sur la base du corpus LM10.

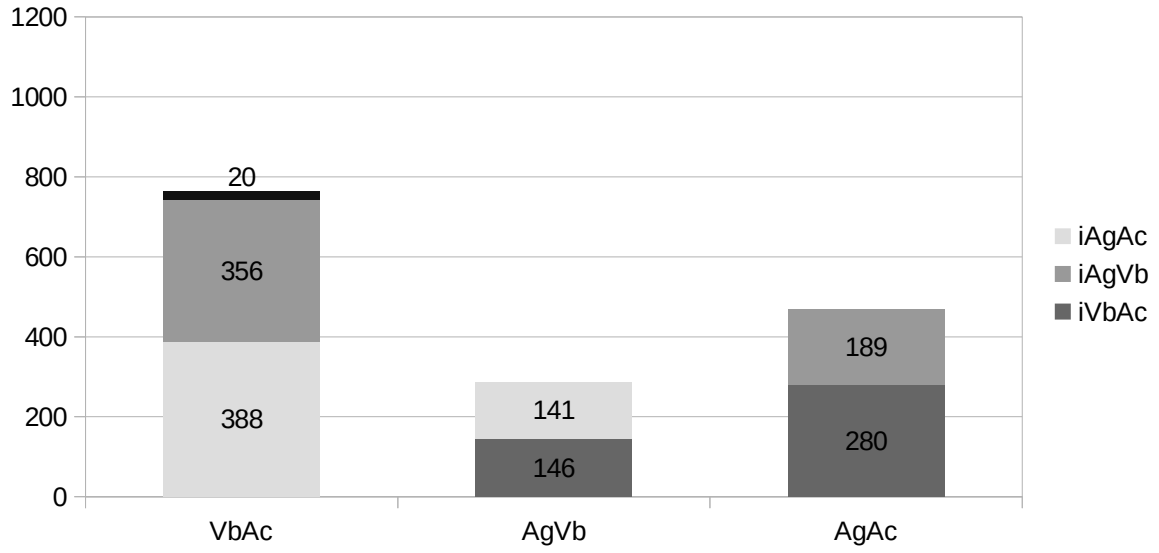


Graphique 6 *– Répartition des triplets nom d'agent/verbe/nom d'action issus de Lexeur en fonction de leur plus grand indice de proximité dans le modèle Word2Vec construit sur le corpus LM10*

Le graphique 6 nous présente donc la répartition des 1520 triplets annotés. On remarque une nouvelle fois qu'une majeure partie des triplets sont annotés VbAc, à raison de 764 triplets, contre 287 triplets annotés AgVb et 469 triplets annotés AgAc. Cela va dans le sens de ce que l'on a observé sur le corpus Wiki. Notons cependant que ces tendances sont moins marquées sur le corpus LM. En effet, seuls 50 % des triplets sont ici annotés VbAc, et ont donc un indice de proximité *iVbAc* le plus élevé, contre 58 % pour le corpus Wiki. *A contrario*, près de 31 % des triplets sont annotés AgAc pour le corpus LM10, contre 25 % dans le corpus Wiki, ce qui signifie que les triplets ont un peu moins tendance à avoir un *iVbAc* élevé, et davantage tendance à avoir un *iAgAc* élevé. De même, on passe de 16 % à près de 19 % de triplets annotés AgVb dans le corpus LM10. Les grandes tendances identifiées grâce au graphique 4 sont globalement identifiables dans le graphique 6, mais dans des proportions moindres.

Nous avons aussi manipulé les triplets de LM10 à la granularité plus fine, pour comparer de

nouveau les tendances observées sur le corpus Wiki à celles que l'on obtient sur le corpus LM10.



***Graphique 7** – Répartition des triplets nom d'agent/verbe/nom d'action issus de Lexeur au sein des trois tendances mises au jour en fonction du second plus grand indice de proximité dans le modèle LM10*

Le graphique 7 nous présente la répartition à un niveau de granularité plus fine des triplets en fonction de l'étiquette et du deuxième indice de proximité le plus élevé. On remarque ainsi que la répartition des triplets de LM10 est similaire à celle des triplets de Wiki représentée dans le graphique 5. Les triplets annotés VbAc ont tendance à avoir comme deuxième indice de proximité le plus élevé le *iAgAc*, à raison de 388 triplets suivant l'inégalité *iVbAc* > *iAgAc* > *iAgVb*, contre 356 triplets répondant à l'inégalité *iVbAc* > *iAgVb* > *iAgAc*. Les 20 triplets restants correspondent au cas où *iAgVb* et *iAgAc* sont égaux, tout comme dans le graphique 5. Il s'agit des triplets où le nom d'action et le verbe sont confondus en un même mot type, avec un *iVbAc* égal à 1, comme illustré dans le tableau 3. De même, les triplets annotés AgAc ont davantage tendance à avoir comme deuxième indice de proximité le plus élevé le *iVbAc*, à raison de 280 triplets répondant à l'inégalité *iAgAc* > *iVbAc* > *iAgVb*, contre 189 pour l'inégalité *iAgAc* > *iAgVb* > *iVbAc*. Enfin, les triplets annotés AgVb se répartissent de façon assez équilibrée entre un *iAgAc* le plus élevé et un *iVbAc* le plus élevé, avec respectivement 141 et 146 triplets. Si les tendances sont donc une fois encore les mêmes que pour le corpus Wiki, les proportions sont moindres. On passe ainsi de 58 % de triplets VbAc répondant à l'inégalité *iVbAc* > *iAgAc* > *iAgVb* pour le corpus Wiki à 51 % pour le corpus LM. Concernant l'inégalité *iAgAc* > *iVbAc* > *iAgVb* chez les triplets AgAc, on passe de 68 % des triplets du corpus Wiki répondant à cette inégalité, contre 60 % pour le corpus LM10.

Les légères différences constatées sont naturellement dues à une représentation vectorielle

différente des mots entre le corpus Wiki et le corpus LM10. On peut cependant s'interroger sur les raisons de cette variation de représentation : est-ce dû à la différence de contenu entre les deux corpus, les différents sujets traités ne permettant pas une même représentation, ou à une évolution de l'usage de certains mots, amenant au rapprochement ou à la distanciation de certains termes sur le plan sémantique et donc vectoriel ?

Malgré ces différences, les différentes observations effectuées pointent toutes dans le même sens, à savoir que le nom d'action et le verbe ont tendance à être les deux éléments les plus proches sur le plan distributionnel au sein d'un triplet, et le nom d'agent et le verbe les deux éléments les plus distants. Si la proximité distributionnelle est réellement un indice de proximité sémantique, cela va donc dans le sens de l'hypothèse émise dans la section 1., à savoir que le contenu sémantique du verbe et du nom d'action est bien plus similaire que le contenu sémantique du verbe et du nom d'agent. Pourtant, un grand nombre de triplets, près de 41 % d'entre eux pour le corpus Wiki et près de 50 % pour le corpus LM10, ne répondent pas à cette hypothèse, les deux éléments les plus proches du triplets étant soit le nom d'action et le nom d'agent, soit le nom d'agent et le verbe. La question est maintenant de savoir si l'on peut tirer de ces tendances des régularités au sein des différents triplets et familles dérivationnelles qui permettraient de généraliser les différents cas de figure.

4.2. Sélection de triplets représentatifs

Nous l'avons vu dans les sections précédentes : les indices de proximité sont de bons outils pour se faire une idée générale, mais ne sont pas toujours pertinents. L'analyse ne peut donc se limiter à l'exploitation de ces distances et se poursuit donc sur l'exploitation des voisins distributionnels dont a été enrichi Lexeur grâce à Word2Vec. Avec plus de 5974 cas à traiter, nous n'avons pas la prétention de pouvoir définir sur le plan distributionnel chaque nom d'agent et nom d'action de Lexeur. Nous préférons analyser de façon précise un nombre plus réduit de cas qui nous semblent représentatifs afin d'ébaucher des généralisations. Différents profils de triplets sont ainsi recherchés : des triplets pour lesquels tous les éléments sont relativement proches entre eux (que l'on qualifiera d'homogène), des triplets avec de grands écarts de proximité, et ce pour des triplets répondant et/ou infirmant l'hypothèse. La question se pose cependant des critères de sélection de ces cas.

4.2.1. En fonction du cumul des indices *iAgVb + iAgAc + iVbAc*

Une première analyse a été faite sur les distances. Nous avons voulu identifier les familles dont les éléments étaient les plus proches en se basant sur le cumul des indices de proximité : on additionne les trois indices au sein de chaque triplet, et l'on compare l'ensemble des indices cumulés obtenus. L'hypothèse sous-tendant cette approche est que plus l'indice cumulé du triplet est élevé, plus on peut supposer que les indices de proximité au sein du triplet sont globalement élevés. *A contrario*,

plus l'indice cumulé sera faible, plus les indices individuels seront globalement faibles, et donc les éléments sémantiquement éloignés.

Pour faire des observations sur la base de l'indice cumulé, les trois indices de chaque triplet ont été additionnés pour obtenir un indice cumulé global. Quelques valeurs de cet indice cumulé sont présentées dans le tableau 6 ci-dessous pour l'ensemble des triplets ainsi que pour les trois tendances que nous avions mis en avant dans la section 4.1. et représentées par les étiquettes AgAc, AgVb et VbAc.

	Global	**AgAc**	**AgVb**	**VbAc**
Minimum	0,01421889	0,03268483	0,07141325	0,01421889
Moyenne	0,929899085	0,87172949	0,81796927	0,77893827
Maximum	2,294929512	2,14393769	2,29492952	1,95536942

Tableau 6 *– Valeurs de l'indice cumulé des triplets nom d'agent/verbe/nom d'action issus du corpus Wiki*

On remarque que les valeurs de l'indice cumulé des triplets varient légèrement en fonction de l'indice de proximité le plus élevé du triplet. En effet, les triplets annotés VbAc ont un indice cumulé moyen plus faible que les autres triplets. L'indice cumulé maximal que l'on peut trouver parmi les triplets annotés VbAc est d'ailleurs plus faible que l'indice cumulé maximal que l'on peut trouver chez les autres triplets. Le triplet ayant l'indice cumulé le plus élevé, et donc potentiellement les indices de proximité les plus élevés, n'est donc pas un triplet annoté VbAc, mais AgVb. *A contrario*, c'est parmi les triplets annotés VbAc qu'on trouve le triplet avec l'indice cumulé faible, et donc les indices de proximité les plus faibles. Les différences sont cependant relativement minimes.

Le tableau 7 regroupe les dix triplets ayant l'indice cumulé *iAgVb* + *iAgAc* + *iVbAc* le plus élevé, ainsi que leurs indices *iAgVb*, *iAgAc* et *iVbAc* respectifs. Les indices les plus élevés de chaque triplet ont été mis en gras. On remarque que les indices de proximité individuels sont relativement élevés (entre 0,5 et 0,8) et relativement homogènes au sein des triplets. Il n'y a généralement pas plus de 0,1 de différence entre deux indices d'un même triplet, sauf dans le cas des triplets *sprinteur*/*sprinter*/*sprint*, *inhalateur*/*inhaler*/*inhalation* et *codeur*/*coder*/*codage* (où cette différence est environ de 0,2). Notons que pour ces deux triplets, la différence touche l'indice *iAgVb*, cet indice devenant le plus élevé dans le premier cas, et le plus faible dans le deuxième cas, faisant du nom d'agent l'élément le plus proche du verbe sur le plan distributionnel pour le triplet *sprinteur*/*sprinter*/*sprint*, et le plus distant du verbe pour le triplet *codeur*/*coder*/*codage*. Notons que s'il fallait labelliser ces triplets selon les étiquettes AgVb, AgAc et VbAc, nous aurions trois triplets AgVb, quatre triplets AgAc et trois triplets VbAc. Ces triplets ne sont donc pas représentatifs des tendances que nous avions pu mettre en avant dans la section 4.1.

Agent	Verbe	Action	Indice AgVb	Indice AgAc	Indice VbAc	iAgVb + iAgAc + iVbAc
inhibiteur	inhiber	inhibition	**0,77320492**	0,760755731	0,760964296	2,29492952
encodeur	encoder	encodage	0,64988274	**0,755388987**	0,738665968	2,14393769
sprinteur	sprinter	sprint	**0,83613872**	0,615093174	0,690190725	2,14142262
inhalateur	inhaler	inhalation	0,67318615	0,634355214	**0,812124936**	2,11966630
égalisateur	égaliser	égalisation	0,673421309	0,650608499	**0,736702884**	2,06073269
régulateur	réguler	régulation	0,629794012	**0,711805619**	0,689144039	2,03074367
torpilleur	torpiller	torpillage	**0,689472606**	0,656037641	0,681531799	2,02704204
remorqueur	remorquer	remorquage	0,669159532	**0,673637644**	0,673480627	2,01627780
climatisateur	climatiser	climatisation	0,618953212	0,686612841	**0,703507534**	2,00907359
codeur	coder	codage	0,544947526	**0,738551365**	0,723302004	2,00680089

Tableau 7 *– Triplets issus du corpus Wiki ayant l'indice cumulé le plus élevé*

L'autre constat que l'on peut faire repose sur la nature des noms d'agent. Sur les dix noms en *-eur* regroupés, seul un est un nom d'agent strict (selon les critères que nous avons évoqués dans la section 1.2.3.2., à savoir *sprinteur*. Quatre autres de ces noms d'agent sont quant à eux des noms d'instrument strict, à savoir *inhbiteur*, *inhalateur*, *égalisateur* et *climatisateur*. Enfin, les cinq autres triplets restants ont une double lecture agentive et instrumentale, en la nature de *encodeur*, *régulateur*, *torpilleur*, *remorqueur* et *codeur*. Par ailleurs, quatre de ces cinq noms à double interprétation appartiennent aux triplets ayant comme indice de proximité le plus élevé *iAgAc*, signifiant que le nom d'agent et le nom d'action sont les deux éléments les plus proches du triplet. De même, trois des quatre noms d'instrument strict appartiennent aux triplets ayant comme indice de proximité le plus élevé *iVbAc*, signifiant que le verbe et le nom d'action sont les deux éléments les plus proches du triplet (et donc que le nom d'instrument est l'élément le plus distant).

Si l'on décrit les noms d'action regroupés dans le tableau 7 selon les critères évoqués dans la section 1.3.2. et sous la forme de l'exemple (34), on obtient les descriptions illustrées en (46).

(46) *inhibition* : [+occurrentiel][-duratif][-culminant][+fortuit]

encodage : [+occurrentiel][+duratif][+culminant][-fortuit]

sprint : [+occurrentiel][+duratif][+culminant][-fortuit]

inhalation : [+occurrentiel][+duratif][+culminant][-fortuit]

égalisation : [+occurrentiel][-duratif][-culminant][+fortuit]

régulation : [-occurrentiel][+duratif]

torpillage : [+occurrentiel][-duratif][-culminant][-fortuit]

remorquage : [+occurrentiel][+duratif][+culminant][-fortuit]

climatisation : [-occurrentiel][-duratif]

codage : [+occurrentiel][+duratif][+culminant][-fortuit]

En plus de cette représentation, notons que les noms *sprint*, *climatisation* et *codage* possèdent une deuxième interprétation : *sprint* peut être interprété comme une activité, *climatisation* peut être interprété comme un instrument, et *codage* comme un objet résultatif.

Ce tableau nous invite à émettre deux premières hypothèses. Tout d'abord, les triplets ici regroupés et leurs indices de proximité laissent à penser qu'au sein d'un triplet, la lecture instrumentale d'un nom rapproche ce dernier, sur le plan distributionnel, du verbe. *A contrario*, la double interprétation semble rapprocher les deux noms d'un même triplet entre eux. On peut s'interroger si c'est la prise en compte du voisinage propre aux deux interprétations qui provoque cette distanciation du nom d'agent et du nom d'action du verbe.

Le processus d'analyse que nous venons de décrire a été répété sur les dix triplets ayant le plus faible indice cumulé.

Le tableau 8 regroupe les dix triplets issus du corpus Wiki ayant l'indice cumulé le plus faible. À l'image du tableau 7, l'indice de proximité individuel le plus élevé de chaque triplet est affiché en gras. Contrairement à ce que l'on avait pour le tableau 7, on constate que les indices de proximité au sein de chaque triplet varient fortement (on qualifiera ces triplets d'hétérogènes). On a par exemple un écart de l'ordre de presque 10e2 entre *iAgAc* et *iVbAc* dans le cas des triplets *acteur*/*agir*/*action* et *lieur*/*lier*/*liaison.* S'il fallait là encore labelliser les triplets réunis, on compterait deux triplets annotés AgVb, quatre triplets annotés AgAc et quatre triplets VbAc.

Il est intéressant de constater qu'on a plusieurs triplets issus d'une même entrée Lexeur : ainsi, les triplets *acteur*/*agir*/*action* et *actrice*/*agir*/*action* sont issus d'une même famille dérivationnelle telle que présentée dans Lexeur et illustrée en (4), et les triplets *ouvreur*/*ouvrer*/*ouvrage* et *ouvreuse*/*ouvrer*/*ouvrage* sont quant à eux issus d'une autre famille dérivationnelle unique dans Lexeur, elle aussi illustrée dans l'exemple (47).

(47) *acteur – actrice – agir – action – acte*

ouvreur – ouvreuse – ouvrer – ouvrage

Agent	Verbe	Action	Indice AgVb	Indice AgAc	Indice VbAc	iAgVb + iAgAc + iVbAc
acteur	agir	action	0,00623914	**0,10541378**	0,00435295	0,11600587
lieur	lier	liaison	**0,10107971**	0,00978796	0,00403341	0,11490108
pointeuse	pointer	pointement	0,00595078	0,04553416	**0,04554851**	0,09703346
ouvreur	ouvrer	ouvrage	0,04023231	**0,04063323**	0,00994657	0,09081127
ouvreuse	ouvrir	ouvrage	0,00819511	0,00121825	**0,07729682**	0,08671021
batteuse	battre	battement	**0,05752456**	0,00679881	0,00708987	0,07141325
accepteur	accepter	acception	0,00082501	**0,06116155**	0,00655215	0,06858718
dameuse	damer	damage	0,00345381	0,02104953	**0,03133288**	0,05583621
actrice	agir	action	0,00629143	**0,02173103**	0,00435295	0,03268482
ouvreuse	ouvrer	ouvrage	0,00305406	0,00121825	**0,00994657**	0,01421889

Tableau 8 *– Triplets issus du corpus Wiki ayant l'indice cumulé le plus faible*

Deux constats peuvent être faits à partir de ces deux triplets. Notons tout d'abord la différence de comportement des triplets composés à partir du verbe *ouvrer*. En effet, si les *iVbAc* de ces deux triplets sont identiques, et ce logiquement puisque ces deux triplets partagent un même verbe et un même nom d'action, leurs *iAgVb* et *iAgAc* varient, là encore de façon logique, puisque ces deux indices impliquent le nom d'agent qui lui varie d'un triplet à l'autre. Pourtant, les noms d'agent *ouvreuse* et *ouvreur* ne semblent pas avoir le même poids : dans le triplet *ouvreur*/*ouvrer*/*ouvrage*, *iAgVb* et *iAgAc* sont quasiment identiques, le nom d'agent était aussi proche du verbe que du nom d'action sur le plan distributionnel. *A contrario*, le triplet *ouvreuse*/*ouvrer*/*ouvrage* montre une bien plus grande proximité entre le nom d'agent et le verbe qu'entre le nom d'agent et le nom d'action. L'écart est tel que dans le cas du premier triplet, les deux éléments les plus proches sont effectivement le nom d'agent et le nom d'action, quand dans le cas du deuxième triplet, les deux éléments les plus proches sont le verbe et le nom d'action. On peut donc se demander si cela est causé par le changement de suffixe, ce qui signifierait que *-euse* et *-eur* ne sont pas sémantiquement équivalents, idée que nous évoquions dans la section 1.2.2., ou si cela est dû à la différence de fréquence, *ouvreuse* ayant une fréquence brute de 41, contre 174 pour *ouvreur* dans le corpus Wiki.

Étudions maintenant le cas des triplets *acteur*/*agir*/*action* et *actrice*/*agir*/*action* dont la famille dans Lexeur est illustrée en (47). Contrairement aux deux triplets que nous venons d'analyser, l'indice le plus élevé dans les deux triplets est le *iAgAc*, bien que celui-ci diffère d'un triplet à l'autre, étant plus faible dans le triplet *actrice*/*agir*/*action* que *acteur*/*agir*/*action*. Ce sont, nous l'avons évoqué plus tôt, les triplets où l'écart entre les indices est le plus fort. En l'occurrence, les indices nous indiquent que les noms d'agent et d'action sont beaucoup plus éloignés du verbe qu'ils ne le sont l'un de l'autre. Cela met en avant ce que notre intuition nous inspirait, à savoir que la construction de

acteur, *actrice*, ou *action* sur la base de *agir* ne nous paraît pas naturelle. Nous l'avions vu lors de la description de Lexeur dans la section 3.1.1., le lexique contient quelques erreurs quant à la constitution de certaines familles dérivationnelles. Un très faible indice de proximité entre deux éléments pourrait donc potentiellement nous montrer des cas de mauvaise formation. On retrouve ainsi par exemple dans le tableau le triplet *ouvreuse*/*ouvrir*/*ouvrage* dont nous parlions dans la section 3.1.1., mais aussi le cas de *accepteur*/*accepter*/*acception*, la présence de ces noms d'action spécifiques dans ces deux triplets étant questionnable.

Parmi les dix noms déverbaux en -*EUR* regroupés dans le tableau 8, quatre sont des noms d'agent au sens strict, *acteur*, *lieur*, *ouvreur* et *actrice*, deux sont des noms d'instrument, à savoir *batteuse* et *dameuse*, et les quatre autres noms restant ont une double interprétation agentive et instrumentale, à savoir *pointeuse*, *ouvreuse* (que l'on retrouve deux fois), et *accepteur*. Parmi les noms féminins, seul un n'a pas d'acception agentive, à savoir *actrice*. On peut donc se demander si le suffixe -*euse* ne tend pas à créer, en plus d'un nom d'agent féminin équivalent à un nom d'agent masculin, un nom d'instrument. Trois de ces noms d'agent strict, *acteur*, *ouvreur* et *actrice*, ont comme indice de proximité le plus élevé l'indice *iAgAc*, signifiant qu'au sein de ces triplets, le nom d'agent est le plus proche du nom d'action. Cela s'explique notamment par la plus grande proximité que l'on attribue intuitivement à *acteur*/*actrice* et *action* qu'à *acteur*/*actrice* et *agir*. Notons que le mot type *lieur* possède aussi une acception adjectivale, ce qui peut potentiellement fausser les indices de proximité obtenus.

Les noms d'action regroupés dans le tableau 8 sont décrits en (48).

(48) *action* : [+occurrentiel][+duratif][+culminant][-fortuit]

liaison : [-occurrentiel][+duratif]

pointement : [+occurrentiel][+duratif][+culminant][+/-fortuit]

ouvrage : [+occurrentiel][+duratif][+culminant][-fortuit]

battement : [+occurrentiel][+duratif][+culminant][+fortuit]

acception : [-occurrentiel][+duratif]

damage : [+occurrentiel][+duratif][+culminant][-fortuit]

Notons que parmi les noms d'action, on retrouve à deux reprises *action* et à trois reprises *ouvrage*. Par ailleurs, une grande partie de ces noms d'action ont une autre interprétation, celle d'objet. C'est le cas de *pointement*, *ouvrage*, *battement* et *acception*. Le nom *liaison* peut quant à lui être interprété comme un nom d'idéalité. Le nom *pointement* répond quant à lui positivement aux deux tests décrits dans la section 1.3.2. caractérisant la fortuité d'un nom (d'où la présence des deux signes + et -), il peut donc être vu comme décrivant une action préméditée mais jugée fortuite aux yeux d'une

majorité des participants, à l'image de noms comme *braquage* ou *attentat*. Notons que *acception* ne semble pas ici être un bon dérivé processif de *accepter* sur le plan sémantique. Enfin, les noms *pointement* et *damage*, qui ont tous les deux comme caractéristique d'être le nom le plus proche du verbe au sein de leur triplet respectif, partagent la même description dans (48) à savoir *[+occurrentiel] [+duratif][+culminant]*. Le seul point qui diffère est le critère de fortuité, mais les deux tendent vers le caractère volontaire de l'action. On peut donc se demander si le caractère occurrentiel, duratif, culminant et non fortuit d'un nom tend à le rapprocher du verbe dont il est dérivé.

Malgré les quelques remarques que l'on a pu faire, l'étude des triplets aux indices cumulés les plus élevés et les plus faibles nous montre que le cumul des indices n'est pas assez significatif. Il nous faut donc trouver d'autres critères pour mieux représenter nos triplets.

4.2.2. En fonction du rapport *iAgVb / iVbAc*

Puisque nous cherchons à comparer distributionnellement le nom d'agent et le nom d'action vis-à-vis du verbe, le choix a été fait de comparer les indices de proximité *iAgVb* et *iVbAc*. Pour ce faire, nous avons mis en lien ces deux indices par le biais de leur rapport, sous la formule *iAgVb / iVbAc*. Le choix a été fait de mettre *iAgVb* en numérateur et *iVbAc* en dénominateur de façon arbitraire. Plus cette valeur est élevée, plus la valeur de *iAgVb* est supérieure à celle de *iVbAc*, et plus le verbe est proche du nom d'agent au sein du triplet. *A contrario*, plus le rapport a une valeur proche de zéro, plus la valeur de *iVbAc* est supérieure à celle de *iAgVb*, et plus le verbe est distributionnellement proche du nom d'action.

	Global	**AgAc**	**AgVb**	**VbAc**
Minimum	0,00084105	0,00686274	1,00728833	0,00084105
Moyenne	1,23390702	1,42006293	3,68644662	0,45385383
Maximum	72,4999517	54,3281942	72,4999517	0,99965863

***Tableau 9** – Valeurs du rapport iAgVb / iVbAc des triplets issus du corpus Wiki*

Le tableau 9 illustre les valeurs minimale et maximale que prend le rapport *iAgVb / iVbAc*, ainsi que la moyenne de ce rapport, pour l'ensemble des triplets ainsi que pour les triplets réunis sous les différents labels AgAc, AgVb et VbAc. On remarque que les valeurs couvertes par les triplets annotés AgVb et VbAc sont complémentaires, ce qui traduit la description de ce rapport que nous venons de faire, à savoir que si la valeur du rapport *iAgVb / iVbAc* est inférieure ou supérieure à 1, cela signifie que le verbe est respectivement plus proche du nom d'action ou du nom d'agent. Les triplets AgAc couvrent assez largement ces valeurs, la valeur minimale du rapport *iAgVb / iVbAc* pour les

triplets AgAc étant bien en-dessous de 0, et la valeur maximale bien au-dessus de 0. Sur la base de cette mesure, nous avons sélectionné les 10 triplets avec les plus grandes valeurs.

Agent	Verbe	Action	Indice AgVb	Indice AgAc	Indice VbAc	iAgVb / iVbAc
tourneur	tourner	tour	**0,18540528**	0,01672414	0,00255731	72,4999517
tourneuse	tourner	tour	0,13893436	**0,14269545**	0,00255731	54,3281942
menteur	mentir	menterie	**0,41824811**	0,20783527	0,00814789	51,3320855
fonceur	foncer	fonçage	**0,21227154**	0,03671173	0,0055563	38,2037744
sécateur	couper	section	**0,32079655**	0,07145716	0,0107898	29,7314662
lieuse	lier	liaison	**0,11935541**	0,10492363	0,00403342	29,5916473
bidouilleur	bidouiller	bidouille	**0,24311197**	0,13351546	0,00829116	29,3218844
menteuse	mentir	menterie	0,23546329	**0,31774572**	0,00814789	28,8986876
commandeur	commander	commanderie	0,23652225	**0,37879629**	0,00859466	27,5196808
bloqueur	bloquer	bloque	**0,188998732**	0,07531069	0,00708443	26,8175829

Tableau 10 *– Triplets ayant les valeurs de rapport iAgVb / iVbAc les plus élevées*

Le tableau 10 regroupe les dix triplets ayant les valeurs du rapport *iAgVb / iVbAc* les plus élevées. L'indice de proximité le plus élevé au sein de chaque triplet est affiché en gras. Le premier constat que l'on peut faire concerne la valeur des indices de proximité propres à chaque triplet. On remarque en effet qu'à l'image de ce que l'on avait avec la mesure d'indice cumulé, un rapport *iAgVb / iVbAc* élevé ne garantit pas une valeur de *iAgVb* élevée. Il s'agit donc une nouvelle fois d'une mesure relative.

Sans surprise, les indices *iAgVb* sont donc toujours supérieurs aux indices *iVbAc* au sein de chaque triplet. L'indice *iVbAc* n'est par ailleurs jamais l'indice de proximité le plus élevé. Plus étonnant, on constate que pour trois des dix triplets regroupés, l'indice de proximité le plus élevé des trois est l'indice *iAgAc*, signifiant que pour ces trois triplets, le nom d'agent et le nom d'action sont les deux éléments les plus proches. Cela montre que notre mesure nous garantit, dans le cas d'une valeur supérieure à 1, que *iAgVb* est supérieur à *iVbAc*, mais pas que *iAgVb* est lui aussi supérieur à *iAgAc*. On a donc la certitude, lorsque la valeur du rapport est supérieure à 1, d'une plus grande proximité du nom d'agent et du verbe par rapport au verbe et au nom d'action, mais l'on ignore à la vue de cette valeur seule si c'est bien la plus grande proximité au sein du triplet étudié.

Sur l'ensemble des noms déverbaux en -*EUR* rapprochés dans le tableau 10, notons que l'un d'entre eux est un nom d'instrument strict, à savoir *sécateur*, le nom *lieuse* pouvant avoir quant à lui la double interprétation agentive et instrumentale. On peut donc s'interroger sur la corrélation entre un

rapport *iAgVb* / *iVbAc* élevé et la nature agentive des noms d'agent impliqués. Il serait intéressant de voir si, à l'inverse, un rapport *iAgVb* / *iVbAc* faible impliquerait la nature instrumentale des noms en *-EUR*.

Nous pouvons aussi constater que, à l'image des triplets regroupés dans le tableau 8, cette mesure rapproche plusieurs triplets issus d'une même famille dérivationnelle dans Lexeur. C'est le cas des triplets *tourneur*/*tourner*/*tour* et *tourneuse*/*tourner*/*tour* et *menteur*/*mentir*/*menterie* et *menteuse*/*mentir*/*menterie*, dont les familles dérivationnelles telles que présentées dans Lexeur sont illustrées en (49).

(49) *tourneur – tourneuse – tourner – tourne – tournement – tournage – tour*

menteur – menteuse – mentir – menterie – mensonge

Dans ces deux triplets, seul le nom d'agent varie, par le biais d'une alternance masculin/féminin. Par ailleurs, ces deux triplets font partie des trois triplets où l'indice de proximité le plus élevé est l'indice *iAgAc*, et ce malgré un rapport *iAgVb / iVbAc* élevé. Le cas du triplet *tourneur*/*tourner*/*tour* et du reste de la famille dérivationnelle associée fait l'objet d'une analyse poussée dans la section 4.3.1.

Les noms d'action regroupés dans le tableau 10 sont décrits en (50).

(50) *tour* : [-occurrentiel][duratif]

menterie : Ø

fonçage : [+occurrentiel][+duratif][-culminant][-fortuit]

bidouille : Ø

section : [+occurrentiel][+duratif][+culminant][+/-fortuit]

liaison : [-occurrentiel][+duratif]

commanderie : Ø

bloque : [-occurrentiel][+duratif]

Parmi ces noms, *tour* et *menterie* sont présents à deux reprises. Les noms *menterie*, *bidouille* et *commanderie* ont été marqués d'un Ø car nous n'avons pas trouvé d'interprétation processuelle. Notre intuition nous dit que les noms *menterie* et *bidouille* sont des noms d'idéalité, le nom *commanderie* étant quant à lui un nom d'objet. Par ailleurs, d'autres noms parmi eux ont une lecture objectuelle, à l'image de *tour* et *section*. Le nom *section* affiche les deux signes + et – pour le critère de

fortuité les deux tests liés à la fortuité se sont révélés positifs, à l'image de *la section de l'artère fémorale s'est produite lors du braquage* et *la section du ligament de M.Martin est reportée à la semaine prochaine*.

Comme pour la mesure *iAgVb + iAgAc + iVbAc*, nous analysons de la même manière les triplets ayant les plus faibles rapports *iAgVb / iVbAc*.

Agent	Verbe	Action	Indice AgVb	Indice AgAc	Indice VbAc	iAgVb / iVbAc
batailleur	batailler	bataille	0,00697553	0,19243944	**0,72994111**	0,00955629
régisseur	régir	régie	0,00148256	**0,38676628**	0,21603038	0,00686274
ouvreur	ouvrir	ouvrage	0,000440099	0,04063238	**0,07729682**	0,00570509
trieuse	trier	triage	0,00046216	0,08820458	**0,11452102**	0,00403562
échangeur	échanger	échange	0,00132157	0,21219701	**0,46549781**	0,00283905
tisseuse	tisser	tissage	0,00154896	0,03580797	**0,56714426**	0,00273115
accepteur	accepter	acceptation	0,00082501	0,13877229	**0,51685623**	0,00159621
corruptrice	corrompre	corruption	0,00096507	0,02975486	**0,61789783**	0,00156186
trieuse	trier	tri	0,00046216	0,08820458	**0,35795151**	0,00129113
ouvreur	ouvrir	ouverture	0,000440099	0,23808042	**0,52432479**	0,00084105

Tableau 11 *– Triplets ayant les valeurs de rapport x/z les plus faibles*

Le tableau 11 regroupe les dix triplets ayant les valeurs de rapports *iAgVb / iVbAc* les plus faibles. L'indice de proximité le plus élevé de chaque triplet a été mis en gras. Cela nous permet de constater que la quasi totalité des triplets, à l'exception du triplet *régisseur*/*régir*/*régie*, ont comme indice de proximité le plus élevé *iVbAc*, signifiant que dans ces neuf triplets, le verbe et le nom d'action sont les deux éléments les plus proches. Par ailleurs, si l'on compare les valeurs les plus élevées de ce tableau à celles du tableau 10, on se rend compte qu'elles sont de façon globale largement supérieures à celles du tableau 10. Ainsi, la moyenne des indices de proximité individuels les plus élevés est de 0,2527425052 pour le tableau 10 et de 0,4654907207 pour le tableau 11, la moitié des indices de proximité les plus élevés du tableau 11 étant supérieurs à 0,5. Cela signifie que les éléments les plus proches du tableau 11 sont donc sur le plan distributionnel plus proches que les éléments les plus proches du tableau 10. En l'occurrence, cela se traduit par le fait que les verbes et noms d'action du tableau 11 sont plus proches que ne sont les verbes et noms d'agent du tableau 11, puisque la majorité de ces indices sont respectivement de type *iVbAc* et *iAgVb*.

Pourtant, il faut rester prudent vis à vis de ces chiffres, notamment dans le cas du triplet *ouvreur*/*ouvrir*/*ouvrage*. Comme nous l'avions vu dans la description de Lexeur dans la section 3.1.1., la construction de ce triplet et de la famille dérivationnelle associée illustrée en (41) est questionnable.

Si *ouvreur* et *ouvrir* peuvent être sémantiquement liés, *ouvrage* ne peut être raccordé au réseau sémantique ébauché par les deux autres termes. Il n'existe que dans l'autre réseau sémantique de *ouvreur*, lié au verbe *ouvrer*. Or, si les indices de proximité de ce triplet sont tous faibles, en l'occurrence inférieur à 0,1, il n'en reste pas moins que l'indice de proximité le plus élevé est le *iVbAc*, signifiant que *ouvrir* et *ouvrage* sont les deux éléments les plus proches au sein du triplet *ouvreur*/*ouvrir*/*ouvrage*, ce qui va à l'encontre de notre intuition.

Sur les dix noms en -*EUR* regroupés dans le tableau 11, six sont des noms d'agent strict, à savoir *batailleur*, *régisseur*, *ouvreur* (que l'on retrouve deux fois), *tisseuse* et *corruptrice*. Les quatre autres noms en -*EUR, trieuse* (que l'on retrouve deux fois), *échangeur* et *accepteur*, ont quant à eux une double lecture agentive et instrumentale. Contrairement à ce que l'on avait envisagé dans le cadre de l'analyse liée au tableau 10, à savoir qu'un rapport *iAgVb / iVbAc* élevé impliquait la présence de noms d'agent strict, et qu'un rapport *iAgVb / iVbAc* faible pouvait par conséquent impliquait la présence de noms d'instrument, on retrouve encore une fois de façon majoritaire des noms d'agent, malgré quelques cas d'interprétation instrumentale. Puisqu'ici, les triplets ont tendance à avoir comme proximité la plus grande celle entre le nom d'action et le verbe, on peut donc imaginer que la nature agentive ou instrumentale du nom en -*EUR* n'est pas un facteur de proximité ou de distance.

Par ailleurs, on constate une nouvelle fois la présence de triplets issus d'une même famille dérivationnelle. C'est notamment le cas des triplets *trieuse*/*trier*/*triage* et *trieuse*/*trier*/*tri* et *ouvreur*/*ouvrir*/*ouvrage* et *ouvreur*/*ouvrir*/*ouverture*, issus des familles illustrées en (51).

(51) *trieur – trieuse – trier – tri – triage*

ouvreur – ouvreuse – ouvrir – ouverture – ouvrage

Dans le cas de ces deux familles, on observe que le nom d'action est toujours plus proche du verbe que le nom d'agent ne l'est, comme le montre l'indice de proximité *iVbAc*, toujours largement supérieur à l'indice de proximité *iAgVb* des triplets impliqués. Ainsi, dans le cas de la famille dérivationnelle formée sur le verbe *trier*, *tri* et *triage* sont toujours plus proches distributionnellement de *trier* que *trieuse* ne l'est. Le nom *tri* est cependant plus proche de *trier* que *triage* de *tri*. Cela montre que pour la famille formée sur *trier*, le nom d'agent est toujours l'élément le plus distant, et ce de façon assez marquée, puisque l'on a des indices de l'ordre de 10e-4 pour le nom d'agent et 10-1 pour le nom d'action. Or, notre intuition nous laisse penser que *trieuse* n'est pas si distant de *trier* que cela, *trieuse* étant soit la personne qui trie, soit un outil qui permet de trier. C'est du côté de la fréquence que l'on peut trouver une réponse. En effet, on constate que *trieuse* n'a qu'une fréquence brute de 6 dans le corpus Wiki, contre 2124 pour *tri*, ou 58 pour *triage*, et 1215 pour *trier*. On remarque par ailleurs que *trier* est plus proche de *tri* que de *triage*, malgré le fait que notre intuition nous indique que *tri* est plutôt le résultat de l'action de trier quand *triage* est davantage l'action de trier. Le peu de contexte sur lequel baser la représentation vectorielle de *trieuse* et comparer les contextes

avec *trier* peut expliquer la faible proximité entre les deux éléments. Dans le cas des triplets basés sur *ouvrir*, la question de la constitution de la famille se repose, comme nous l'avons dit un peu plus tôt. En effet, on constate que le nom d'action est dans les deux cas plus proche du verbe que ne l'est le nom d'agent. Or, parmi les deux noms d'action en question, on a *ouverture* et *ouvrage*. Si *ouverture* semble bien lié au réseau sémantique de *ouvrir*, *ouvrage* ne rentre pas dans ce réseau. Notons que si l'indice *iVbAc* de ce triplet reste l'indice le plus élevé du triplet, il reste très faible, surtout par comparaison aux autres indices *iVbAc* du tableau 11.

(52) *bataille* : [+occurrentiel][+duratif][+culminant][-fortuit]

régie : [-occurrentiel][+duratif]

ouvrage : [+occurrentiel][+duratif][+culminant][-fortuit]

triage : [-occurrentiel][+duratif]

échange : [+occurrentiel][+duratif][+culminant][-fortuit]

tissage : [-occurrentiel][+duratif]

acceptation : [-occurrentiel][+duratif]

corruption : [+occurrentiel][+duratif][+culminant][+/-fortuit]

tri : [+occurrentiel][+duratif][-culminant][-fortuit]

ouverture : [+occurrentiel][-duratif][-culminant][-fortuit]

Parmi les noms d'action recensés et décrits en (52), on a notamment des noms d'activité, comme *tissage*, *triage*, *régie*, mais aussi des noms statifs, comme *acceptation*, et des noms d'action au sens strict, comme *tri* ou *ouverture*.

La mesure *iAgVb / iVbAc* mettant aussi en avant les triplets où le nom d'agent et le nom d'action sont à peu près à la même distance du verbe, nous allons rapidement observer certains de ces triplets.

Le tableau 12 regroupe les dix triplets ayant un rapport *iAgVb / iVbAc* le plus proche de 1. L'indice de proximité le plus élevé au sein de chaque triplet est mis en gras. Comme nous l'avons vu dans la présentation de cette mesure, un rapport *iAgVb / iVbAc* proche de 1 signifie que la distance entre le nom d'agent et le verbe est similaire à celle entre le nom d'action et le verbe. Si le rapport est supérieur à 1, le nom d'agent est légèrement plus proche du verbe que ne l'est le nom d'action, et si le rapport est inférieur à 1, c'est le nom d'action qui est le nom le plus proche du verbe au sein du triplet.

Deux cas de répartition des indices s'observent : celui où les indices *iAgAc* et *iVbAc* sont les plus élevés, et celui où l'indice *iAgAc* est le plus élevé. L'indice *iAgAc* peut alors être proche des indices *iAgVb* et *iVbAc,* à l'image du triplet *innovateur*/*innover*/*innovation*, où le nom d'agent, le nom

d'action et le verbe sont tous les trois proches les uns des autres sur le plan distributionnel, ou être distant des indices *iAgVb* et *iVbAc*, à l'image du triplet *éventreur*/*éventrer*/*éventration*, les deux noms étant plus fortement rapprochés. Notons cependant que l'indice de proximité le plus élevé est dans sept des dix triplets l'indice *iAgAc*.

Agent	Verbe	Action	Indice AgVb	Indice AgAc	Indice VbAc	iAgVb / iVbAc
évangélisateur	évangéliser	évangélisation	**0,68954576**	0,557749961	0,68455649	1,00728833
entremetteur	entremettre	entremise	0,20346367	**0,2418891**	0,20264796	1,00402521
rameur	ramer	ramage	0,15740804	**0,21131511**	0,15688553	1,00333051
surfeur	surfer	surf	0,4065985	**0,57709671**	0,40588554	1,00175656
accompagnateur	accompagner	accompagnement	0,35605352	**0,4861325**	0,35589054	1,0004579
chanteur	chanter	chant	0,55758476	0,44866343	**0,55777517**	0,99965862
racleur	racler	raclage	0,57238526	0,33616224	**0,57271572**	0,99942299
innovateur	innover	innovation	0,52578217	**0,56233315**	0,52787249	0,99604011
réassureur	réassurer	réassurance	0,4298796	**0,56991978**	0,43161958	0,99594872
éventreur	éventrer	éventration	0,09474959	**0,389605**	0,09517987	0,99547921

Tableau 12 *– Triplets ayant les valeurs de rapport iAgVb / iVbAc les plus proches de 1*

Parmi les dix noms en *-EUR* regroupés dans le tableau 12, seuls deux d'entre eux ne sont pas des noms d'agent strict, à savoir *rameur* et *racleur*, qui ont une double lecture agentive et instrumentale. Un grand nombre d'entre eux semble avoir une interprétation statuaire ou dispositionnelle, selon les critères évoqués dans la section 1.2.1.2., à l'image d'*évangélisateur*, *surfeur* ou *éventreur*. Concernant les noms d'action, la plupart d'entre eux sont des noms d'action, comme *évangélisation*, *raclage* ou *innovation*. Certains ont aussi une interprétation d'activité, comme *surf* ou *chant*.

4.3. Analyse distributionnelle de triplets

L'ensemble des triplets issus de Wiki a été classé en fonction du rapport *iAgVb/iVbAc* que nous venons de décrire. Nous avons sélectionné plusieurs triplets, dans le but de couvrir plusieurs cas de figures : celui où le nom d'agent est clairement plus proche du verbe que ne l'est le nom d'action, celui où le nom d'action est bien plus proche du verbe que ne l'est le nom d'agent, et un cas intermédiaire. Pour avoir l'assurance que la relation liant l'un ou l'autre nom du triplet au verbe est bien la relation la plus forte du triplet, et ne pas se retrouver avec des cas où la proximité est la plus grande

entre les deux noms, nous allons uniquement sélectionner nos triplets parmi les triplets annotés AgVb et VbAc.

4.3.1. Rapport iAgVb/iVbAc élevé

Nous sélectionnons dans un premier temps deux triplets dont le rapport *iAgVb* / *iVbAc* est élevé, à savoir le cas où le verbe est distributionnellement plus proche du nom d'agent que du nom d'action. Nous étudions alors deux configurations : une première où le rapport est élevé, mais où les indices sont faibles, et une seconde où le rapport est élevé et les indices eux-aussi élevés.

4.3.1.1. tourneur – tourner – tour

Le premier triplet choisi est le triplet *tourneur*/*tourner*/*tour* dont le rapport *iAgVb/iVbAc* a une valeur de 72,4999517, ce qui en fait le triplet avec la valeur de rapport *iAgVb/iVbAc* la plus élevée. Cela signifie donc que c'est le triplet où la proximité entre le nom d'agent et le verbe est proportionnellement la plus grande par rapport à celle entre le nom d'action et le verbe. Ce triplet annoté AgVb est issu de la famille dérivationnelle telle que présente dans Lexeur illustrée en (49), repris ici.

(49) *tourneur – tourneuse – tourner – tourne – tournement – tournage – tour*

Le triplet *tourneur*/*tourner*/*tour* n'est pas le seul qui ait été conservé après le traitement de la famille dérivationnelle par le programme décrit dans la section 3.2.2.2. appliqué à la matrice créée à partir du corpus Wiki. L'ensemble des triplets obtenus à partir de la famille dérivationnelle illustrée en (50) sont regroupés dans le tableau 13, accompagnés de leurs indices de proximité respectifs.

Agent	Verbe	Action	iAgVb	iAgAc	iVbAc	iAgVb / iVbAc
tourneur	tourner	tour	**0,185405289**	0,016724141	0,002557316	72,4999517
tourneur	tourner	tourne	0,185405289	0,198567587	**0,452103672**	0,41009463
tourneur	tourner	tournage	0,185405289	0,199256887	**0,387878935**	0,47799782
tourneuse	tourner	tour	0,138934362	**0,142695445**	0,002557316	54,3281942
tourneuse	tourner	tourne	0,138934362	0,224032471	**0,452103672**	0,30730642
tourneuse	tourner	tournage	0,138934362	0,014223452	**0,387878935**	0,35819001

Tableau 13 *– Triplets issus de la famille dérivationnelle formée sur le verbe* tourner *et leurs indices de*

Le tableau 13 nous renseigne donc sur le fait que les triplets impliquant le nom d'action *tournement* n'ont pas été conservés. Cela s'explique par la très faible présence du nom *tournement* dans le corpus Wiki, enregistrant une fréquence brute de 2.

On remarque tout de suite que les indices de proximité sont relativement faibles, ce qui souligne bien que la mesure utilisée dans la section 4.2.2. est une mesure relative et non absolue. L'autre constat que l'on peut faire sur la base des indices de proximité est que quatre des six triplets réunis dans le tableau 13 ont comme indice de proximité le plus élevé l'indice *iVbAc*, signifiant que pour la majorité des triplets issus de la famille (50), le nom d'action est plus proche du verbe que ne l'est le nom d'agent. Les deux seuls cas où cela est infirmé sont les triplets impliquant le nom d'action *tour*, à savoir *tourneur*/*tourner*/*tour* et *tourneuse*/*tourner*/*tour*. Pour ces deux triplets, l'indice de proximité le plus élevé est respectivement le *iAgVb* et le *iAgAc*. On peut s'interroger quant à cette différence de comportement.

Nous avions envisagé dans la section 4.1. l'impact de la fréquence d'un mot sur la valeur de l'indice de proximité. Ainsi, nous avions constaté que lorsque certains indices de proximité étaient faibles, l'un des deux membres impliqués pouvait avoir une fréquence brute faible. Dans le cas de nos deux triplets, l'indice de proximité le plus faible est le *iVbAc*, impliquant le nom d'action et le verbe. Or, la fréquence brute de *tour* dans le corpus Wiki est de 129403, et celle de *tourner* est de 30026. Par comparaison, la fréquence brute de *tourne* est de 128, celle de *tourneur* est de 1035 et celle de *tourneuse* est de 18. Cela n'explique donc ni la faible proximité entre *tour* et *tourner*, ni la plus grande proximité entre *tourneuse* et *tour*. La fréquence n'a donc pas de rôle spécifique dans notre cas de figure. Nous allons donc regarder du côté des voisins distributionnels pour espérer comprendre ces comportements.

Pour cela, nous exploitons la liste de voisins communs dont nous avons décrit la constitution dans la section 3.2.2.1. Dans le cadre de la famille (50), nous obtenons 100 voisins distributionnels communs, dont quelques exemples sont illustrés en (53). L'ensemble des 100 voisins distributionnels communs est fourni en annexe 1.

(53)

match	*tournage/tour*
enfoirés	*tourne/tournage*
rediffuser	*tourner/tournage*
accrocher	*tourner/tourne*
basculer	*tourner/tourne*
demy	*tourneur/tourne/tournage*
gueltz	*tourneur/tourneuse*

Parmi ces 100 voisins distributionnels, 18 sont communs à *tournage* et au moins un autre membre de la famille dérivationnelle (49). Le nom d'action *tournage* partage des voisins distributionnels avec *tour*, *tourne* et *tourner*. On a comme voisins partagés des mots comme *vidéo-clip*, *filmer*, *rediffuser*, *coproduire* ou *déprogrammer*, tous en lien avec le monde du spectacle, du cinéma ou des médias télévisuels. Le sens du nom *tournage* ébauché par ces voisins est celui lié au monde du cinéma et du média visuel. L'acception liée à l'usinage exploitant l'outil qu'est le tour n'est pas réellement sensible par le biais des voisins distributionnels. En effet, les voisins distributionnels partagés par *tournage* et *tour* réfèrent eux aussi au monde du spectacle, avec *match*, *warm-up* et *warped*. De par les voisins qu'il partage avec *tournage*, une première acception de *tourner* liée au monde du cinéma, tel que *tourner un film*, est ébauchée.

Le verbe *tourner* partage aussi des voisins distributionnels avec un autre des noms d'action de la famille (49), le nom *tourne*. Ces voisins sont au nombre de 79, et sont du type *courber*, *déambuler*, *déraper*, *dévier* ou encore *illuminer*. Une grande partie de ces voisins sont des verbes de mouvement ou de déplacement, à l'image de *cheminer*, *glisser* ou encore *bouger*. Est donc ébauché par le biais de ces voisins un deuxième sens du verbe *tourner*, à savoir la notion de déplacement impliquant une direction. On peut alors s'interroger sur le sens et la nature du mot *tourne* tel que représenté par ces voisins et la matrice. En effet, il n'existe pas à notre connaissance de nom *tourne* dont le sens serait lié à la notion de déplacement ou de direction. Il existe bien deux noms féminins *tourne*, l'un étant le nom régional pour des éléments de maçonnerie, l'autre étant un terme technique pour désigner une réaction chimique, à l'image de la « tourne du lait », ou en lien avec les domaines de la boulangerie et de la presse. Notre intuition nous laisse penser que *tourne* est ici vu dans la matrice créée par Word2Vec comme l'instanciation à la première ou à la troisième personne du présent de l'indicatif ou du subjonctif du verbe *tourner* dont le sens vient d'être identifié. Cela souligne les limites de cette représentation que nous avions évoquée dans la section 3.2.2.2.

Enfin, notons que les noms d'agent *tourneur* et *tourneuse* ne sont impliqués que dans le cas de trois voisins partagés, à savoir *demy*, *grémillon* (tous deux partagés par *tourneur* et *tourne*) et *gueltz*, uniquement partagé par *tourneur* et *tourneuse*. Ces trois mots nous étant inconnus, nous avons utilisé un concordancier, AntConc[4], pour prendre connaissance de leurs contextes d'apparition dans le corpus Wiki. On constate ainsi deux types majeurs d'usage différents, illustrés en (54), pour le mot *demy* : une version ancienne de *demi*, que l'on retrouve dans des articles citant des textes du Moyen Âge ou de la Renaissance, et un nom propre. Ce dernier est fortement lié au premier sens de *tourner* que nous avions identifié, à savoir celui propre au monde du cinéma, puisque c'est le nom de famille d'une lignée de réalisateurs français dont on retrouve les biographies dans Wikipédia.

(54) *Leurs champs ne sont qu'à demy cultivez*

Les Demoiselles de Rochefort, film réalisé par Jacques Demy en coproduction avec la Warner Bros

4 Le concordancier AntConc est disponible à l'adresse http://www.laurenceanthony.net/software.html.

Il en va de même pour *grémillon*, qui s'avère être le nom de famille d'un réalisateur dont la biographie et la filmographie font l'objet de nombreux articles. Notons qu'il existe aussi une entrée Wikipédia pour une romancière portant le nom de Grémillon, ainsi que pour un ruisseau. Mais le ruisseau comme la romancière ne sont liés qu'à un nombre très réduit d'articles, alors que le réalisateur Grémillon voit le nombre d'articles (et donc de contextes pour apprendre le sens du mot-type grémillon) augmenté par chacun des films auxquels il a contribué. De même, le mot *gueltz* s'avère être lui aussi le nom de famille d'un réalisateur, nom que l'on retrouve donc dans tous les articles Wikipédia traitant de ses films.

Le tableau 1 nous montre que le verbe *tourner* est plus proche de *tourneur* que de *tour* et que *tourneuse* est plus proche de *tour* que ne l'est *tourner*. Le nom d'action étant jusque-là dans cette entrée de Lexeur plus proche du verbe que ne l'est le nom d'agent, on peut donc s'interroger sur cette différence. L'analyse de *tourneur* par le biais du concordancier AntConc nous montre que *tourneur* est un nom de métier mais surtout un nom de famille apparaissant dans plusieurs contextes. Ce nom de famille désigne différents réalisateurs, dramaturges, hommes politiques et militaires, mais aussi une femme de lettre et un sculpteur. Néanmoins, on retrouve une nouvelle fois la prédominance du monde du cinéma par le biais des réalisateurs. Cela le rapproche donc du premier sens identifié pour *tourner*. Le mot *tour*, quant à lui, apparaît dans des contextes très variés, comme cela est illustré en (55).

(55) *Le clocher, tour octogone à trois étages*

Peu après, il gagnait le Tour de Suisse

Il est battu lors du second tour des élections municipales

Les contextes de *tour* que nous venons de mettre en avant rendent donc difficile le rapprochement de *tour* avec *tourner* et avec *tourneur*, qui ont quant à eux des sens bien plus clairement définis, notamment dans la matrice. Le terme *tourneuse* n'apparaît quant à lui dans le corpus Wiki que dans le syntagme *La Tourneuse de pages*, qui est le titre d'un film. On peut donc se demander pourquoi *tourneuse* n'est pas rapproché de *tourner*, qui a une acception cinématographique certaine, mais est rapproché de *tour*, au sens plus vague et général. Le seul élément que nous avons à notre disposition est la fréquence brute de *tourneuse*, qui était particulièrement faible, comme nous l'avions vu un peu plus tôt, à savoir de 18. Cette faible fréquence est peut-être la cause d'une représentation peu spécifique du sens de *tourneuse* et donc plus facilement rapprochable de *tour* que de *tourner*.

Dans le cas de cette famille, le nom d'action est le plus souvent le plus proche du verbe, et ce dans le cadre des deux réseaux sémantiques que l'on a pu mettre en évidence, celui du cinéma et celui du déplacement, et ce malgré le rapport *x/z* très élevé qui nous avait incité à partir sur le triplet *tourneur*/*tourner*/*tour*. Notons qu'un troisième réseau, celui de l'artisanat et de l'industrie, semble être

effacé par cette représentation et cette analyse. Cela peut s'expliquer par sa spécificité, qui fait qu'on rencontre moins souvent ces acceptions que les autres : il n'y a en effet que quelques articles spécifiques qui relient les mots-type *tourneur*, *tourner* ou *tour* à la notion d'artisanat. *A contrario*, chaque article sur un film, sur une émission ou sur un acteur permet de lier davantage la famille *tourner* à l'univers du cinéma ou de la télévision, par exemple.

4.3.1.2. fumeur – fumer – fumette

Nous venons d'étudier un triplet ayant un rapport *iAgVb* / *iVbAc* élevé, mais des indices de proximité *iAgVb*, *iAgAc* et *iVbAc* faibles. Nous cherchons à présent à analyser un triplet qui aurait lui aussi un rapport x/z élevé, mais dont les indices *iAgVb*, *iAgAc* et *iVbAc* seraient plus élevés, afin de voir si l'intensité de la proximité a un impact sur la caractérisation des triplets et des relations au sein de la famille dérivationnelle dont ils sont issus. Nous avons donc choisi le triplet *fumeur*/*fumer*/*fumette*, dont le rapport *iAgVb / iVbAc* a une valeur de 5,70606868, ce qui en fait le 59e triplet ayant le rapport le plus élevé. Ce triplet, annoté AgVb, est issu de la famille dérivationnelle telle que présente dans Lexeur et illustrée en (56). Le triplet *fumeur*/*fumer*/*fumette* n'est pas le seul triplet à avoir été conservé suite au processus de constitution et d'enrichissement des triplets.

(56) *fumeur – fumeuse – fumer – fumature – fumage – fumaison – fumerie – fumette*

Agent	Verbe	Action	iAgVb	iAgAc	iVbAc	iAgVb / iVbAc
fumeur	fumer	fumage	**0,58048632**	0,16772477	0,36074867	1,609115596
fumeur	fumer	fumaison	**0,58048632**	0,17535841	0,31233464	1,85853968
fumeur	fumer	fumerie	**0,58048632**	0,31499905	0,36893135	1,57342638
fumeur	fumer	fumette	**0,58048632**	0,14913019	0,1017314	5,70606868
fumeuse	fumer	fumage	0,30235422	0,1021336	**0,36074867**	0,83812982
fumeuse	fumer	fumaison	0,30235422	0,1175344	**0,31233464**	0,96804576
fumeuse	fumer	fumerie	0,30235422	0,2438533	**0,36893135**	0,81954061
fumeuse	fumer	fumette	**0,30235422**	0,06016543	0,1017314	2,97208361

Tableau 14 *- Triplets issus de la famille dérivationnelle formée sur le verbe fumer et leurs indices de proximité*

Le tableau 14 présente l'ensemble des triplets constitués à partir de la famille (56) et conservés suite à l'enrichissement des triplets. Seuls les triplets impliquant le nom d'action *fumature*

n'ont pas été conservés. Cela est dû à l'absence de *fumature* du corpus Wiki, et donc à sa non-représentation dans la matrice.

On remarque tout d'abord que les indices de proximité sont plus élevés au sein des triplets impliquant le nom d'agent en *-eur* qu'au sein des triplets impliquant le nom d'agent en *-euse*. Cela peut être dû à la différence de fréquence, *fumeuse* n'ayant qu'une fréquence brute de 19 dans le corpus Wiki, contre 431 pour *fumeur*. Il s'agira de voir si une nuance de sens accompagne la différence de genre et de fréquence. On constate en tout cas que trois des quatre triplets impliquant *fumeuse* ont comme indice de proximité le plus élevé l'indice *iVbAc*, à l'inverse des triplets impliquant *fumeur* qui ont tous comme indice de proximité le plus élevé l'indice *iAgVb*. Cela laisse entendre que le verbe *fumer* est globalement plus proche de *fumeur* que de n'importe quel nom d'action, mais qu'il est plus distant de *fumeuse* que de n'importe quel nom d'action (à l'exception du triplet *fumeuse*/*fumer*/*fumette*).

Nous allons maintenant analyser les voisins distributionnels que partagent les différents membres de la famille dérivationnelle (56) pour essayer de comprendre ces différences. Ces voisins, au nombre de 414, sont très divers, comme le montrent les quelques exemples illustrés en (57). L'ensemble des voisins distributionnels communs sont fournis en annexe 2.

(57) *salaison* — *fumage/fumaison/fumerie*
formaldéhyde — *fumage/fumaison*
coriande — *fumer/fumage/fumaison/fumerie*
friture — *fumer/fumage*
haschich — *fumeur/fumer/fumerie*
dentifrice — *fumeur/fumer*

Le premier constat que l'on peut faire repose sur les deux tendances assez nettes qui se dégagent en ce qui concernant le verbe et les noms d'action, et ce grâce aux voisins. Tout d'abord, les voisins partagés par les noms d'agent *fumage*, *fumaison* et *fumerie* d'une part et le verbe *fumer* d'autre part semblent tous tendre vers le thème de la cuisine ou des processus chimiques liés à des manipulations culinaires, à l'image de *appertiser*, *dessiccation*, *saumurage* ou *torréfier*. *A contrario*, les voisins partagés par le verbe *fumer* et le nom d'agent *fumeur* tendent vers des notions de goût et d'hygiène, à l'image de *désinfecter*, *cigare*, *cacahuète* ou *cervelle*. Deux sens de *fumer* semblent donc se dessiner : un premier réseau serait constitué de *fumer* et des noms d'action *fumage* et *fumaison*, liés à l'action très spécifique d'un traitement des aliments à base de fumée. On retrouve par ailleurs ce réseau dans une autre famille dérivationnelle de Lexeur, illustrée en (58). Un second réseau serait quant à lui constitué du verbe *fumer*, à l'image de *fumer une cigarette*, et des noms *fumeur* et *fumerie*.

(58) *fumailleur – fumailleuse – fumailler – fumature – fumaison – fumerie – fumette – fumage*

L'indice *iVbAc* des trois triplets impliquant *fumeur* nous montre que *fumeur* a tendance à être un peu plus proche de *fumerie*, nom d'action du second réseau sémantique, que des noms d'action du premier réseau. Le nom *fumeur* reste cependant toujours plus proche du verbe que des noms d'agent. Des mots comme *cigarette*, *vendeur*, *fumeuse*, *droqué*, *revendeur* ou *dealer* parmi les 500 voisins de *fumeur*, fournis en annexe 3, nous montrent que *fumeur* tend à appartenir au second réseau. Ce second réseau semble avoir plus de poids que le premier, au point que ce soit le sens du verbe de ce second réseau qui fasse l'objet d'une comparaison dans le cadre des triplets.

Le second constat que l'on peut faire sur la base de ces voisins distributionnels partagés est que le nom d'agent *fumeuse* ne partage qu'un nombre limité de voisins, à savoir 21, dont 2 uniquement avec *fumer* (à savoir *fumeur* et *théière*), 1 avec *fumette* (*droguiste*), 1 avec *fumage* (*aromathérapie*), et les 17 restants avec *fumerie*. Parmi ces voisins, on retrouve des termes comme *cordonnerie*, *parapharmacie*, *qincaillerie*, *pelleterie*, ou plus étonnamment *xiuhtecuhtli* (qui est le nom d'un dieu de la mythologie aztèque). AntConc nous montre que *fumeuse* est utilisé comme un nom d'agent féminin désignant la femme qui fume, comme dans la phrase *c'est une femme cordiale, fumeuse invétérée*, mais aussi comme un adjectif, dans des phrases comme *c'est une fumeuse élucubration* ou *chez l'homme que chez la femme fumeuse*. Ces éléments ne nous éclairent pas réellement sur le rapprochement effectué. Si l'on regarde directement les 500 voisins distributionnels de *fumeuse*, fournis en annexe 4, on observe que le premier d'entre eux est *fumeur*, avec un indice de proximité de 0,574993. Les voisins suivants tendent à être des noms d'agent statuaires assez variés, comme *vendeur*, *boutiquier*, *tisserande* ou *rentier*, pour lesquels on a du mal à voir le lien sémantique. Le mot *fumeuse* semble néanmoins peu lié aux sens de *fumer* que nous avons pu mettre en avant.

Notons que *fumette*, nom d'action qui semble faire varier les indices de proximité pour les triplets impliquant le nom *fumeuse*, ne partage un voisin distributionnel (*droguiste*) qu'avec un seul autre membre de la famille (56), à savoir avec *fumeuse*. Les voisins distributionnels de *fumette* sont principalement des noms, ce qui donne une représentation sémantique assez éloignée de la notion de tabac ou de substances qu'on lui donne habituellement. Cela peut expliquer que *fumeuse* et *fumette*, sur la base d'un sens assez peu spécifique, soient rapprochés.

Dans le cadre de cette famille dérivationnelle, on constate deux comportements distincts : une plus grande proximité du nom d'agent masculin et du verbe, et une plus grande proximité du nom d'agent féminin et du nom d'action. D'après nos observations, ces rapprochements, qui vont à l'encontre de l'hypothèse que nous émettions, semblent s'expliquer par un manque de contextes suffisants sur lesquels apprendre une représentation sémantique. Par ailleurs, l'existence d'un réseau sémantique secondaire n'aide pas à une représentation juste des proximités, puisqu'à un mot n'est assimilé qu'un seul sens dans la matrice. Il est dès lors délicat de séparer les deux réseaux, ce qui rend les comparaisons moins pertinentes.

4.3.2. Rapport iAgVb/iVbAc faible

Nous sélectionnons maintenant deux triplets dont le rapport *iAgVb* / *iVbAc* est faible, c'est-à-dire des triplets où le verbe est distributionnellement plus proche du nom d'action que du nom d'agent, ce qui correspond au cas général. Nous étudions alors deux configurations : une première où le rapport est faible, mais où les indices sont faibles, et une seconde où le rapport est faible et les indices eux aussi élevés.

4.3.2.1. trieuse – trier – tri

Jusque-là, nous nous étions intéressé à des triplets ayant un rapport *iAgVb* / *iVbAc* supérieur à 1. Nous allons maintenant nous intéresser à des triplets dont le rapport *iAgVb* / *iVbAc* est inférieur à 1. Nous nous penchons notamment sur le cas d'un triplet ayant une des valeurs *iAgVb* / *iVbAc* les plus élevées. Nous aurions pu choisir *ouvreur*/*ouvrir*/*ouverture*, mais ce triplet ayant fait l'objet d'une petite analyse dans la section 4.2.2., nous avons décidé de nous concentrer sur un autre triplet. C'est pour cela que nous avons choisi le triplet *trieuse*/*trier*/*tri*. Ce triplet, annoté VbAc, est issu de la famille dérivationnelle présente dans Lexeur illustrée en (59).

(59) *trieur – trieuse – trier – tri – triage*

Cette famille forme quatre triplets, dont la totalité a été conservée après l'enrichissement effectué.

Agent	Verbe	Action	iAgVb	iAgAc	iVbAc	iAgVb / iVbAc
trieur	trier	tri	0,126837	0,24449927	**0,3579515**	0,35434129
trieur	trier	triage	0,126837	**0,43162057**	0,11452102	1,10754336
trieuse	trier	tri	0,00046216	0,05904974	**0,3579515**	0,00129113
trieuse	trier	triage	0,00046216	0,08820458	**0,11452102**	0,00403562

Tableau 15 *– Triplets issus de la famille dérivationnelle formée sur le verbe* trier *et leurs indices de proximité*

Le tableau 15 regroupe les triplets formés à partir de la famille (59). Deux de ces quatre triplets font partie des triplets ayant un des rapports *iAgVb / iVbAc* les plus faibles, à savoir *trieuse*/*trier*/*tri* et *trieuse*/*trier*/*triage*. Seul un de ces triplets n'a pas l'indice *iVbAc* comme indice de proximité le plus élevé : il s'agit du triplet *trieur*/*trier*/*triage*. Ce triplet a comme indice le plus élevé

l'indice *iAgAc*, et a un rapport *iAgVb / iVbAc* supérieur à 1. Il s'agit donc de comprendre pourquoi *trieur* est plus proche de *triage* que ne l'est *trier*, alors que *trier* est globalement plus proche de l'ensemble des noms d'action pour le reste de la famille dérivationnelle.

Si l'on cherche du côté des fréquences, on constate que *trieur* a une fréquence brute de 32, ce qui est faible, mais *trieuse* a une fréquence brute encore plus faible de 6. Cela n'explique donc pas la différence de comportement de *trieur* face à *triage*, qui a une fréquence de 583, par rapport à *trieuse* et *triage* ou *trieur* et *tri*.

Cette famille partage un total de 83 voisins distributionnels, fournis en annexe 5. Ces voisins relèvent tous d'un vocabulaire relativement technique, comme illustré en (60).

(60)

garage-atelier	*tri/triage*
calibrage	*trier/tri*
granulométrique	*trier/tri*
parser	*trier/tri*
meulage	*trieur/tri*
rotomoulage	*trieur/tri*
frigorifique	*trieur/triage*
dégraissage	*trieur/trier/tri*
tronçonner	*trieur/trier*

Le nom d'agent et d'instrument *trieuse* ne partage que quatre voisins avec d'autres membres de la famille dérivationnelle : *scier*, *essoreuse*, *récurer* et *tronçonner*. Le nom d'agent *trieur* partage un faible nombre de voisins avec *triage* (*bluterie*, *chaudronnerie*, *déchargement*, *frigorigique*, *laminoir*, *manutention* et *trémie*) alors qu'il en partage 56 avec *tri*. Si les voisins communs de *trieur* et *triage* tendent tous vers un champ lexical de la fabrication industrielle, ceux partagés par *trieur* et *tri* laissent aussi une grande place au champ lexical de l'informatique, avec des mots comme *mégaoctets* ou *tableur*. Si l'on observe les voisins distributionnels de *triage*, on constate qu'une grande partie d'entre eux sont des noms topographiques comme *villiers-saint-georges*, *massy-palaiseau* ou *cambrai-ville*. Cela peut expliquer la plus grande distance entre le verbe *trier* et le nom *triage*, même si l'on a du mal à justifier la relativement grande proximité (plus de 0,4) entre *trieur* et *trier*, les voisins distributionnels de *trieur* relevant, eux, davantage du vocabulaire technique de l'industrie. La grande distance du nom d'agent *trieuse* et du verbe, mettant ainsi en avant la plus grande proximité du verbe et du nom d'action, s'explique, elle, par les voisins de *trieuse*, qui sont très variés mais n'appartiennent pas du tout au vocabulaire rencontré dans les voisins de *trier*.

4.3.2.2. corruptrice – corrompre – corruption

À l'image de la comparaison que nous avons faite dans la section 4.3.1. et 4.3.2. entre deux triplets distants pour le cas d'un rapport *iAgVb* / *iVbAc* élevé, nous allons maintenant nous intéresser à un triplet au rapport *iAgVb* / *iVbAc* faible, mais dont les indices de proximité sont élevés. C'est pourquoi nous avons choisi le triplet *corruptrice*/*corrompre*/*corruption*, annoté VbAc, et issu de la famille dérivationnelle de Lexeur illsutrée en (61).

(61) *corrupteur – corruptrice – corrompre – corruption*

Cette famille forme deux triplets, tous deux ayant été conservés lors de l'enrichissement des triplets.

Agent	Verbe	Action	iAgVb	iAgAc	iVbAc	iAgVb / iVbAc
corrupteur	corrompre	corruption	0,52803084	0,37178978	**0,61789783**	0,85456011
corruptrice	corrompre	corruption	0,00096507	0,02975486	**0,61789783**	0,00156186

Tableau 16 – *Triplets issus de la famille dérivationnelle formée sur le verbe* corrompre *et leurs indices de proximité*

Le tableau 16 nous présente les deux triplets issus de la famille (61) ainsi que leurs indices de proximité respectifs et leur rapport iAgVb */ iVbAc*. On constate tout de suite que le nom d'agent masculin et le nom d'agent féminin n'entretiennent pas du tout le même rapport de proximité envers le verbe et le nom d'action. Le nom d'agent féminin *corruptrice* est ainsi distributionnellement plus distant des deux autres éléments que ne l'est le nom d'agent masculin *corrupteur.*

Si *corrupteur* a une fréquence brute faible de 93 dans le corpus Wiki, *corruptrice* est encore moins représentée dans le corpus, à raison d'une fréquence brute de 7. Le nom *corruption* a quant à lui une fréquence brute de 3545 dans le corpus Wiki. Cela se traduit au niveau des voisins distributionnels partagés par le fait que *corruptrice* ne partage aucun voisin distributionnel avec les autres membres de la famille dérivationnelle. Sur un total de 159 voisins distributionnels partagés pour l'ensemble de la famille, fournis dans l'annexe 6, 58 sont partagés par *corrompre* et *corruption*, 25 par *corrupteur* et *corruption*, et 58 par *corrompre* et *corrupteur* uniquement, 18 voisins étant partagés par *corrupteur*, *corrompre* et *corruption* à la fois. Le contenu de ces voisins est globalement similaire sur le plan sémantique, avec des voisins tels qu'illustrés en (62). La seule distinction que l'on peut faire repose sur la catégorie grammaticale des mots, *corrupteur* et *corruption* ou *corrompre* et *corruption*, partageant

davantage de noms ayant tendance à incarner des concepts, que *corrupteur* et *corrompre*, qui partagent davantage d'adjectifs.

(62) *clientélisme corrompre/corruption*
injustice corrompre/corruption
cupidité corrupteur/corrompre/corruption
cynique corrupteur/corrompre
inhumain corrupteur/corrompre
blasphème corrupteur/corruption
préjugé corrupteur/corruption

Si les voisins distributionnels communs nous laissent penser que le sens de ces différents mots est globalement similaire, seule la fréquence joue ici un rôle quant à la proximité des éléments entre eux, comme nous l'avions évoqué dans la section 4.1. En ce qui concerne le nom d'agent féminin *corruptrice*, on retrouve dans ses voisins distributionnels des noms topographiques comme *saint-georges-antignac* ou *pargny-les-bois*, mais aussi des noms *a priori* étrangers, comme *dornröschen* ou *schlachtschiff*. Les contextes observés à l'aide du concordancier AntConc ne nous éclairent cependant pas vis-à-vis de ces voisins distributionnels, le mot-type *corruptrice* étant utilisé comme un nom dans des titres d'œuvres littéraires et télévisuelles, et comme adjectif dans des phrases comme *il s'agit d'une force corruptrice plus indicible*.

Dans le cas de cette famille dérivationnelle, deux hypothèses peuvent expliquer la plus grande proximité distributionnelle du nom d'action *corruption* et du verbe *corrompre* par rapport aux noms d'agent *corrupteur* et *corruptrice* : la fréquence, qui diminue le nombre de contextes et donc les conditions d'apprentissage du sens de ces noms d'agent, et peut-être tout simplement la traduction de l'hypothèse que nous avions formulée quant à la proximité naturelle consécutive au processus de formation du nom d'action et du verbe.

4.3.3. Rapport iAgVb/iVbAc proche de 1

Enfin, nous sélectionnons des triplets pour lesquels le rapport *iAgVb* / *iVbAc* est proche de 1, à savoir que le verbe est quasiment aussi proche du nom d'action que du nom d'agent. Nous choisissons donc un cas où le verbe est plus proche du nom d'agent que du nom d'action, et donc le rapport légèrement plus grand que 1, puis un cas où le verbe est plus proche du nom d'action que du nom d'agent, et donc le rapport légèrement inférieur à 1.

4.3.3.1. évangélisateur – évangéliser – évangélisation

Puisque nous avons étudié les cas extrêmes d'un rapport *iAgVb* / *iVbAc* élevé et d'un rapport *iAgVb* / *iVbAc* faible, nous allons maintenant nous intéresser rapidement aux cas d'un rapport *iAgVb* / *iVbAc* proche de 1. Il s'agit des cas où la proximité du couple nom d'agent/verbe et du couple verbe/nom d'action est quasi similaire. Nous allons essayer de voir ce qui distinguer ces cas de ceux que nous avons étudié précédemment.

Pour cela, nous avons sélectionné le triplet *évangélisateur*/*évangéliser*/*évangélisation*, dont le rapport *iAgVb / iVbAc* est à peine supérieur à 1, avec 1,00728833. Annoté AgVb, ce triplet est issu de la famille dérivationnelle présente dans Lexeur telle qu'illustrée en (63).

(63) *évangélisateur – évangélisatrice – évangéliser – évangélisation*

Agent	Verbe	Action	iAgVb	iAgAc	iVbAc	iAgVb / iVbAc
évangélisateur	évangéliser	évangélisation	**0,68954576**	0,57749961	0,68455649	1,00728833

Tableau 17 *– Triplet issu de la famille dérivationnelle formée sur le verbe* évangéliser *et ses indices de proximité*

Le tableau 17 nous présente l'unique triplet conservé (sur les deux théoriques) suite à l'enrichissement des triplets formés à partir de la famille (63), ainsi que ses indices de proximité. L'absence du triplet *évangélisatrice*/*évangéliser*/*évangélisation* s'explique par la fréquence brute du nom d'agent féminin *évangélisatrice*, qui n'est présent qu'une fois dans le corpus Wiki. Le nom *évangélisateur* a quant à lui une fréquence brute de 255, le verbe *évangéliser* ayant, lui, une fréquence de 553.

On constate que les indices de proximité sont relativement proches, tout particulièrement *iAgVb* et *iVbAc*. Les trois éléments du triplet sont donc sur le plan distributionnel proches. Le nom d'agent et le nom d'action sont ainsi quasiment à la même distance du verbe.

La famille dérivationnelle (6) partage 151 voisins distributionnels, fournis en annexe 7. Plus précisément, 39 voisins distributionnels sont partagés par les trois membres du triplet, à l'image des exemples illustrés en (64).

(64) *apostolat* *évangélisateur/évangélisation*
orient *évangélisateur/évangélisation*
diocèse *évangélisateur/évangéliser/évangélisation*
bienheureux *évangélisateur/évangéliser*
catéchiser *évangéliser/évangélisation*

Si l'on compare les différents voisins partagés, on remarque que l'on retrouve certains mots très proches. Ainsi, on retrouve *orthodoxe* comme voisin commun de *évangélisateur* et *évangélisation*, mais aussi *orthodoxes* comme voisin commun de *évangélisateur*, *évangéliser* et *évangélisation*. De même, nous avons *christianisme* comme voisin commun de *évangélisateur* et *évangélisation*, et nous avons *christianisation* et *christianiser* comme voisins communs de *évangélisateur*, *évangéliser* et *évangélisation*. La seule réelle différence que l'on peut noter au niveau du sémantisme que le lien entre *évangélisateur* et *évangélisation* semble davantage reposer sur des concepts lié à la religion, comme *chritianisme*, *chrétienté*, *donatisme*, *hagiographie* ou *hérésie*, quand le lien entre *évangéliser* et *évangélisation* repose davantage sur l'histoire de la christianisation, avec des termes comme *colonisation*, *instruire*, *rechristianiser* ou *séminariste*. Il y a d'ailleurs davantage de verbes dans ces voisins distributionnels-là.

Cela semble donc aller dans le sens de notre hypothèse, puisque malgré une grande similarité des trois termes, *évangéliser* et *évangélisation* sont davantage rapprochés. Ces deux mots partagent davantage de voisins, et leur lien distributionnel est plus fort.

4.3.3.2. chanteur/chanter/chant

Enfin, nous allons étudier un autre triplet ayant un rapport *iAgVb / iVbAc* proche de 1, mais cette fois-ci étant inférieur à 1. Nous avons donc choisi le triplet *chanteur*/*chanter*/*chant*, annoté VbAc, est issu de la famille dérivationnelle présente dans Lexeur telle qu'illustrée en (65).

(65) *chanteur – chanteuse – chanter – chantage – chant*

Notons que le rattachement de *chantage* à cette famille est questionnable. Si l'on peut définir *chantage* par l'idée faire chanter quelqu'un, il n'y a pas de lien sémantique selon notre intuition entre *chantage*, à savoir le moyen de pression que l'on peut exercer sur quelqu'un, et *chanter*, l'activité musicale exploitant la voix.

Le tableau 18 regroupe les différents triplets conservés après la formation et l'enrichissement des triplets à partir de Lexeur. Tous les triplets ont été conservés. Ce tableau nous permet de constater que la proximité entre *chantage* et les autres membres des triplets où il est impliqué est très faible, ce qui est en accord avec notre intuition. Les noms d'agent *chanteur* et *chanteuse* ne sont d'ailleurs les éléments les plus proches du verbe que lorsque le nom d'action auquel ils sont comparés est le nom *chantage*. On constate une nouvelle fois que le nom d'agent féminin est cependant un peu plus distant du verbe que ne l'est le nom d'agent masculin.

Agent	Verbe	Action	Indice AgVb	Indice AgAc	Indice VbAc	iAgVb / iVbAc
chanteur	chanter	chant	0,55758476	0,44866343	**0,55777517**	9,71316424
chanteur	chanter	chantage	**0,55758476**	0,065644	0,05740506	0,99965862
chanteuse	chanter	chant	0,40937447	0,34476313	**0,55777517**	0,73394172
chanteuse	chanter	chantage	**0,40937447**	0,02510876	0,05740506	7,13133099

Tableau 18 *– Triplets issus de la famille dérivationnelle formée sur le verbe* chanter *et ses indices de proximité*

Encore une fois, la différence de fréquence entre le nom d'agent féminin et le nom d'agent masculin est relativement conséquente. On passe en effet d'une fréquence brute de 35021 pour *chanteur* à une fréquence brute de 308 pour *chanteuse* dans le corpus Wiki. Cette différence de fréquence n'a cependant pas d'impact sur le nombre de voisins distributionnels que les deux noms partagent, illustrés dans l'exemple (66), et fournis dans leur totalité en annexe 8.

(66) *berceuse* *chanter/chant*
medley *chanter/chant*
psalmodier *chanter/chant*
dalida *chanteur/chanteuse/chanter*
arrangeuse *chanteur/chanteuse*
chanteur *chanteuse/chanter*

En effet, sur les 291 voisins distributionnels partagés par les membres de la famille dérivationnelle (66), 148 sont partagés par le nom d'agent féminin et un ou plusieurs autres membres de la famille dérivationnelle, et 212 voisins sont partagés par le nom d'agent masculin et un ou plusieurs autres membres de la famille dérivationnelle. Si l'on veut être plus précis, 126 de ces voisins sont strictement partagés par le nom d'agent masculin et le nom d'agent féminin. Lorsque l'on étudie ces voisins, on n'observe pas de distinction sémantique spécifique en fonction des mots du triplet auxquels ils sont associés, ce que traduit la forte similarité des indices *iAgVb* et *iVbAc*.

Nous avons dans cette section analysé différents triplets. Cette analyse pourrait être perfectionnée. Nous n'avons en effet pas tenu compte de la fréquence des éléments des triplets pour choisir nos triplets. Par ailleurs, nous avons notamment étudié des cas extrêmes de triplets à la constitution douteuse. Il faudrait, pour parfaire cette étude, retirer de la liste des triplets enrichis ceux dont le rapport *iAgVb* / *iVbAc* montre un trop grand écart à la moyenne. Mais l'on peut noter par ces

analyses que le rapport *iAgVb* / *iVbAc* permet d'identifier les cas limites. Se baser sur cette mesure peut potentiellement aider à un nettoyage de Lexeur.

Par ailleurs, l'ensemble des critères de description que nous avions mis en avant dans la section 1. n'a pas été exploité. Si cette description pouvait être automatisée, cela permettrait une étude plus systématique des triplets et familles dérivationnelles, ce qui pourrait faciliter la mise en évidence de phénomène.

Conclusion

Bien qu'à un état embryonnaire, cette étude vise à comparer sémantiquement des dérivés morphologiques mettant en vis-à-vis noms agentifs déverbaux et leur base d'une part, et noms processifs et cette même base d'autre part.

Formulée par Roché (2009), l'hypothèse à l'origine de cette étude est que, du fait de la dérivation morphologique amenant à la formation des noms processifs déverbaux, ces derniers seraient sémantiquement similaires au verbe dont ils seraient issus, contrairement aux noms agentifs déverbaux, pour lesquels une modification sémantique prendrait place. Une des premières étapes dans la démonstration de cette hypothèse a été de voir si les noms d'action, à défaut d'être sémantiquement identiques aux verbes, étaient les éléments sémantiquement les plus proches des verbes au sein de familles dérivationnelles.

Nous l'avons vu dans la section 1.1.4., lors d'une dérivation morphologique, le sens du dérivé est construit à partir du sens de la base, ce qui permet la création d'un réseau sémantique parallèlement à la création de la famille dérivationnelle. Le choix a été fait de tester cette nouvelle hypothèse en comparant sémantiquement le lien entre nom d'action et verbe et celui entre nom agentif en *-EUR* (appellation qui regroupe pour rappel les suffixes *-eur*, *-euse* et *-rice*) et verbe pour limiter dans un premier temps le champ de recherche. Nous avons donc cherché à décrire ce qu'était un nom d'action et un nom d'agent en *-EUR*. Nous avons ainsi dû distinguer les noms agentifs en *-EUR* des noms instrumentaux en *-EUR*, et nous avons dû mettre en avant la notion de réseaux secondaires, ce phénomène de polysémie se rencontrant régulièrement.

De par la nature de la comparaison, qui mettait en lien des noms et des verbes, nous avons fait le choix de baser notre étude sur une approche distributionnelle, dont nous avons vu dans la section 2.2. qu'elle permettait de manipuler librement ces objets et ce, malgré la différence catégorielle. Sur la base de la ressource Lexeur regroupant près de 6 000 familles dérivationnelles, nous avons ainsi pu identifier nos besoins, à savoir un outil, Word2Vec, et des données volumineuses, ce qui nous a conduit à choisir deux corpus, Wikipédia et LM10. Nous avons alors pu enrichir les données fournies par notre ressource Lexeur, dans le but d'analyser distributionnellement les noms d'agent, verbes et noms d'action contenus dans Lexeur.

Cette analyse a permis de mettre en avant plusieurs choses. Nous avons ainsi pu constater que l'analyse distributionnelle des données enrichies issues de Lexeur semble conforter l'hypothèse de base, à savoir que le nom d'action et le verbe étaient en moyenne les éléments les plus proches au sein d'une même famille dérivationnelle. Nous avons aussi pu constater que cela dépendait de différents facteurs, comme la nature du nom d'agent (masculin contre féminin) et la validité de la famille dérivationnelle étudiée. Nous avons tenté d'identifier des tendances en fonction de la nature agentive ou instrumentale du nom en *-EUR*, ainsi que de l'interprétation événementielle, objectuelle ou autre du nom d'action, sans pour autant parvenir à généraliser ces tendances. Nous nous sommes à de nombreuses reprises interrogé sur l'impact du suffixe *-euse* ou *-rice*, de par les différences d'indices de proximité que cela produisait, et d'autre part par les plus nombreux cas d'interprétation instrumentale

que cela semblait causer. Nous avons pu remarquer que l'approche distributionnelle que nous avons employée était globalement pertinente, malgré quelques limites liées d'une part à la quantité et à la propreté des données, et d'autre part aux limites des représentation des réseaux sémantiques secondaires.

Cette étude n'est qu'une première approche de la question, qui se révèle vaste et complexe. De nombreuses pistes sont envisagées pour poursuivre et approfondir l'analyse entamée. Ainsi, il serait intéressant d'approfondir une comparaison entre les noms d'agent en *-EUR* féminins et masculins. De même, insister davantage sur la nature précise des noms agentifs et processifs comparés, selon un étiquetage qu'il faudrait davantage normer voire automatiser, pourrait aider à mettre en avant certains phénomènes. Enfin, la question des réseaux secondaires et de leur représentation dans le cadre d'une approche distributionnelle se pose. Nous avons vu que les profils distributionnels permettaient d'identifier des réseaux sémantiques sous-jacents, mais il serait bon de trouver un moyen d'améliorer cette représentation de sorte que la fusion de plusieurs réseaux secondaires ne vienne pas fausser les chiffres et analyses, par le biais d'une modification de la ressource de base qu'est Lexeur, ou par un travail sur le tri des voisins distributionnels.

Bibliographie

Amiot, D., & Dal, G. (2009). L'autonomie de la morphologie vus du côté de la grammaticalisation. *Mémoires de la Société de Linguistique de Paris*, (Nouvelle série n°17), 33-48

Anscombre, J. C. (2001). À propos des mécanismes sémantiques de formation de certains noms d'agent en français et en espagnol. *Langages* (143), 28-48.

Ascombre, 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*, 11-27.

Benveniste, E. (1975). *Noms d'agent et noms d'action en indo-européen*. Maisonneuve, Paris.

Booij, G. (2009). La morphologie constructionnelle - un aperçu (Traduit par Jacques François). *Mémoires de la Société de Linguistique de Paris*, (Nouvelle série n°17), 13-32.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22-29.

Croft, W. (1991). *Syntactic categories and grammatical relations : the cognitive organization of information*. Chicago : University of Chicago Press.

Dal, G. (2003) Analogie et lexique construit : quelles preuves. *Cahiers de grammaire, 28*, 9-30.

Fabre, C. (2010). *Affinités syntaxiques et sémantiques entre mots : apports mutuels de la linguistique et du TAL* (Doctoral dissertation) Université Toulouse le Mirail, Toulouse.

Fabre, C., Hathout, N., Ho-Dac, L. M., Morlane-Hondère, F., Muller, P., Sajous, F., ... & Van de Cruys, T. (2014). Présentation de l'atelier SemDis 2014: sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés. In *21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014),* 196-205.

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

Firth, J. (1957). A Synopsis of Linguistic Theory 1930-1955. In *Studies in Linguistic Analysis*, Philological Society, Oxford.

Fradin, B. (2009). Morphologie constructionnelle et sémantique. *Mémoires de la Société de Linguistique de Paris*, (Nouvelle série n°17), 89-118.

Gábor, K. (2014). Le systeme WoDiS-WOlf & DIStributions pour la substitution lexicale. In *Sémantique Distributionnelle-Atelier TALN 2014*.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3), 146-162.

Huyghe, R. (2014). La sémantique des noms d'action: quelques repères.*Cahiers de lexicologie*, 105. 181-201.

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

Kerleroux, F. (1999). Sur quelles bases opère l'apocope. *Silexicales 2: la morphologie des dérivés évaluatifs*, 95-106.

Lachance, S. (1988). *La concurrence suffixale en-eur (-euse) et-eux (-euse) en français québécois* (Thèse). Université de Laval, Ottawa.

Lignon, S. (2002). L'adjectif en-ien comme révélateur de phénomènes de concurrence. *Bulletin de linguistique appliquée et générale*, (27), 135-150.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR*.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Plénat, M. (2009). Le conditionnement de l'allomorphie radicale en français. *Mémoires de la Société de Linguistique de Paris*, (Nouvelle série n°17), 119-140.

Rychlý, P., & 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 ACL on interactive poster and demonstration sessions*, 41-44.

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

Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. (Doctoral dissertation) Stockholm University : Stockholm.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, *20*(1), 33-54.

Tuesday, O. S (2011). La suffixation agentive et le blocage affixal: le cas du suffixe ‘’eur’’et ses concurrents morphologiques. *Synergies Afrique Centrale et de l'Ouest* (4), 77-85.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning : Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1), 141-188.

Vossen, P. (1997). EuroWordNet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, 5-7.

Déclaration sur l'honneur de non-plagiat

(à joindre au mémoire à la fin du document)

Je soussigné.e,

Nom, Prénom : Wauquier, Marine
Régulièrement inscrit.e à l'Université de Toulouse II Jean Jaurès
N° étudiant : 21500713

Année universitaire : 2015-2016

certifie que le document joint à la présente déclaration est un travail original, que je n'ai ni recopié ni utilisé des idées ou des formulations tirées d'un ouvrage, article ou mémoire, en version imprimée ou électronique, sans mentionner précisément leur origine et que les citations intégrales sont signalées entre guillemets.

Fait à : Toulouse

Le : 1er juin 2016

Signature :