Apports de la sémantique distributionnelle pour la morphologie dérivationnelle

Marine Wauquier

▶ **To cite this version:**

Marine Wauquier. Apports de la sémantique distributionnelle pour la morphologie dérivationnelle. Corpus, 2022, 23, 10.4000/corpus.6303 . hal-03947484

HAL Id: hal-03947484
https://hal.science/hal-03947484

Submitted on 19 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CORPUS

Corpus

23 | 2022
Corpus et données en morphologie

Apports de la sémantique distributionnelle pour la morphologie dérivationnelle

Distributional semantic approaches for derivational morphology

Marine Wauquier

OpenEdition Journals

Édition électronique
URL : https://journals.openedition.org/corpus/6303
DOI : 10.4000/corpus.6303
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Référence électronique
Marine Wauquier, « Apports de la sémantique distributionnelle pour la morphologie dérivationnelle », *Corpus* [En ligne], 23 | 2022, mis en ligne le 02 mars 2022, consulté le 05 mars 2022. URL : http://journals.openedition.org/corpus/6303 ; DOI : https://doi.org/10.4000/corpus.6303

Ce document a été généré automatiquement le 5 mars 2022.

© Tous droits réservés

Apports de la sémantique distributionnelle pour la morphologie dérivationnelle

Distributional semantic approaches for derivational morphology

Marine Wauquier

Introduction

1 L'ancrage empirique et quantitatif de la morphologie s'est développé ces dernières années par le biais des données et des méthodes d'exploitation de ces dernières. En effet, les ressources lexicographiques portant sur les données morphologiques se sont multipliées, et regroupent un nombre toujours plus important de lexèmes et de familles morphologiques. Les propriétés décrites par ces ressources sont diverses. Les ressources construites à partir des versions française, anglaise et italienne du Wiktionnaire que sont GLAWI (Hathout et Sajous 2016), ENGLAWI (Sajous *et al.* 2020) et GLAWIT (Caledorne *et al.* 2016) recensent par exemple des propriétés lexicographiques ainsi que phonétiques. D'autres ressources plus spécialisées, comme DerivBase (Zeller *et al.* 2013, Zeller *et al.* 2014) pour l'allemand, CELEX (Baayen *et al.* 1995) pour l'anglais, l'allemand et le néerlandais, ou Verbaction (Hathout *et al.* 2002), Démonette (Hathout et Namer 2016), VerNom (Missud *et al.* 2020) et MORDAN (Koehl 2013) pour le français, entre autres, fournissent quant à elles des informations morphologiques. Ce développement s'est fait en parallèle de l'intégration d'outils statistique, qui permet d'exploiter pleinement ces grandes quantités de données désormais accessibles. L'association des deux permet la confirmation ou la découverte de phénomènes divers, modélisés empiriquement et quantifiés sur la base des données.

2 La sémantique est longtemps restée le parent pauvre de cette évolution empirique de la morphologie. Cela est notamment dû à l'absence de ressources encodant les propriétés sémantiques des dérivés morphologiques, du fait de l'impossibilité d'annoter automatiquement à grande échelle des propriétés sémantiques. La sémantique

distributionnelle se révèle à ce titre une réponse intéressante en cela qu'elle propose une représentation statistique du sens construite à partir des usages en corpus et permet une analyse à grande échelle de grandes quantités de données.

3 Cet article présente un état de l'art de l'utilisation de la sémantique distributionnelle pour la morphologie. Après une première section consacrée à la présentation succincte de la sémantique distributionnelle et à une réflexion sur le choix de l'outil pour la morphologique, nous montrons en section 2 la polyvalence et la puissance de l'outil, permettant parfois à moindre coût une analyse linguistique fine à même d'éclairer des questions variées. Nous consacrons la section 3 à des considérations linguistiques liées à l'utilisation de la sémantique distributionnelle, pour soutenir l'outil informatique.

1. La sémantique distributionnelle

4 Afin de contextualiser l'usage de la sémantique distributionnelle en morphologie, nous revenons brièvement sur les principes sous-tendant cette approche et nous passons ensuite en revue les différents outils à disposition.

1.1. Une approche statistique du sens

5 La sémantique distributionnelle est une approche quantitative et statistique du sens, dans laquelle l'accès à l'information, notamment sémantique, se fait par le biais de la modélisation statistique des usages langagiers (Turney et Pantel 2010). Basée sur l'hypothèse distributionnelle (Harris 1954, Firth 1957), cette approche repose sur l'idée que « the amount of meaning correspond[s] roughly to the amount of difference in their environments » (Harris 1954 : 157). En d'autres termes, plus la distribution de deux mots diffère, plus leur sens tend à différer, et réciproquement, les mots sémantiquement similaires auront tendance à partager un plus grand nombre de contextes.

6 Conséquemment, la similarité sémantique de deux mots peut être évaluée sur la base du partage de contextes. De fait, l'approche distributionnelle établit une corrélation entre la similarité distributionnelle et la similarité sémantique (Sahlgren 2008, Lenci 2018, Boleda 2020), et les contextes des mots permettent de les représenter sous la forme de vecteurs dans un espace où la proximité spatiale traduit la similarité sémantique. Des mots sémantiquement similaires auront donc des vecteurs proches dans l'espace. Cette représentation spatiale de la similarité sémantique se traduit par la spécialisation de l'espace en zones qui sont dans l'ensemble sémantiquement cohérentes et distinctes (mais non exclusives) et qui présentent des propriétés sémantiques particulières (Mickus *et al.* 2020, Wauquier 2020). Si cette segmentation sémantique de l'espace n'est, dans les faits, pas toujours si nette, elle est néanmoins suffisamment sensible pour être exploitée dans le cadre de tâches de TAL ou des études linguistiques.

7 Cette proximité est quantifiée sur une échelle de 0 (proximité nulle) à 1 (proximité maximale) par un score, cosinus ou euclidien, calculé à partir de l'angle ou de la distance entre les vecteurs. Ce score, à interpréter de façon relative et non absolue, indique de fait l'existence d'un lien sémantique ('relatedness') mais ne donne pas d'indice concernant la nature de ce lien. Il est par ailleurs possible d'identifier les

voisins distributionnels d'un mot donné, c'est-à-dire les mots les plus proches distributionnellement dudit mot.

1.2. Une famille d'outils

8 Depuis ses premières implémentations, la sémantique distributionnelle a connu des évolutions relatives aux méthodes de calcul des représentations vectorielles qui encapsulent l'information contextuelle. De fait, plusieurs implémentations co-existent désormais.

9 Les implémentations originelles, dites *count-based*, se basent sur le décompte en corpus des cooccurrences graphiques. Les métriques et algorithmes utilisés dans le calcul des dimensions peuvent néanmoins varier, tant dans la construction des matrices initiales (Turney et Pantel 2010) que dans la réduction des matrices (Deerwester *et al.* 1990, Landauer et Dumais 1997, Levy *et al.* 2015, Lenci 2018). Bien que toujours utilisés, notamment du fait de leur relative transparence, ces modèles se caractérisent par une lourdeur computationnelle et n'offrent pas de performances plus intéressantes que les implémentations neuronales plus récentes, computationnellement plus accessibles (Baroni *et al.* 2014b). Ces modèles sont néanmoins encore fortement utilisés dans certaines branches, et notamment en psycholinguistique, où les modèles d'analyse sémantique latente, ou LSA (Landauer et Dumais 1997), sont encore majoritaires.

10 Cependant, les modèles *count-based* ont récemment été supplantés par des modèles neuronaux dits prédictifs, qui intègrent des réseaux de neurones visant à prédire les valeurs des dimensions des vecteurs à partir des cooccurrences en corpus. Ces valeurs sont initialisées aléatoirement, puis actualisées à mesure que le système rencontre de nouveaux contextes, toujours dans l'optique de proposer des vecteurs, ou *embeddings*, similaires pour des mots aux contextes semblables. On retrouve notamment parmi ces méthodes l'outil Word2Vec (Mikolov *et al.* 2013) qui a largement prévalu dans les tâches de TAL, de par ses performances et sa facilité d'utilisation. Diverses variations ont été proposées, se basant notamment sur les relations de dépendances (Levy et Goldberg 2014), ou construites à partir de contextes plus limités, à l'image de définitions dictionnairiques, mais sans toutefois connaître la même popularité que la version initiale de Word2Vec.

11 Mais ces méthodes ont elles-mêmes été éclipsées par l'arrivée encore plus récente d'un troisième type d'implémentation, au travers des modèles de langue pré-entraînés dont BERT (Devlin *et al.* 2018) est le représentant. Ces modèles s'avèrent computationnellement plus complexes puisqu'ils reposent sur du *deep learning*, et donc moins accessibles. Ils se caractérisent néanmoins par la construction d'*embeddings* dits contextuels. Ces derniers se distinguent des *embeddings* dits statiques proposés par Word2Vec en cela que la représentation agrège l'information d'une occurrence, dans une phrase donnée, et non d'une forme à l'échelle du corpus entier. Cette approche permet ainsi la modélisation de phénomènes sémantiques complexes comme la polysémie (voir Yenicelik *et al.* 2020 pour une brève revue de l'état de l'art), en permettant la représentation différenciée des différents sens associés à une forme donnée, là où Word2Vec ne propose qu'un *embedding* unique sur la base de l'ensemble des contextes des différents lexèmes.

12 Notons qu'en parallèle de l'opposition entre modèles *count-based* et neuronaux subsiste une autre distinction. Certains outils proposent ainsi une approche compositionnelle,

qui repose sur l'hypothèse que des membres morphologiquement liés sont sémantiquement liés. De fait, la similarité formelle est, dans ces approches, vue et exploitée comme un indice de similarité sémantique. Ainsi, la représentation proposée par les outils intégrant cette approche ne se fait plus au niveau du mot mais au niveau des n-grammes, et le vecteur d'un mot est considéré comme l'association des vecteurs des n-grammes qui le composent. Les méthodes d'association et de découpage des composants sont diverses (Mitchell et Lapata 2010, Lazaridou *et al.* 2013) mais toutes ont pour objectif de proposer une représentation des mots construits plus fine, que le mot soit peu présent voire absent du corpus. De nombreuses études se sont penchées sur la question de la compositionnalité des vecteurs dans une optique morphologique (Luong *et al.* 2013, Botha et Blunsom 2014, Soricut et Och 2015, *inter alia*), et des outils de sémantique distributionnelle compositionnelle à proprement parler ont émergé, tant *count-based* (Baroni *et al.* 2014a) que neuronales, à l'image de fastText (Bojanowski *et al.* 2016) et, à sa façon, BERT. Cette approche montre cependant des limites en cela que le traitement morphologique est relativement basique, puisqu'il amalgame chaîne de caractères et marquage morphologique. Ces modèles tendent ainsi à favoriser un rapprochement formel et non sémantique (Avraham et Goldberg 2017) et à proposer une spatialisation sémantique moins claire que celle des modèles prédictifs (Mickus *et al.* 2020).

13 Le choix de l'outil va ainsi principalement dépendre de l'objectif de l'étude. Au choix de l'outil s'ajoute par ailleurs le paramétrage de celui-ci. L'utilisation de Word2Vec implique par exemple le choix d'une architecture, du nombre de dimensions, et de divers algorithmes qui influencent à différentes étapes le calcul des *embeddings*. Nous n'explorons pas ces aspects dans cet article, mais si ces paramètres ont une influence sur les représentations sémantiques construites (Baroni *et al.* 2014B, Bernier-Colborne et Drouin 2016, *inter alia*), leurs choix relèvent principalement de l'optimisation des performances et non d'une remise en question des représentations.

2. Composer avec la sémantique distributionnelle

14 Largement utilisée en TAL, la sémantique distributionnelle se fait progressivement une place dans la recherche en linguistique et plus spécifiquement en morphologie, en offrant de nouvelles perspectives d'analyse sémantique, tant théoriques que méthodologiques.

2.1. Vers une morphologie distributionnelle

15 Si les apports de la sémantique distributionnelle sont principalement sémantiques, son champ d'application est néanmoins multiple.

16 Divers travaux se sont penchés sur l'étude des affixes. Il s'agit notamment de proposer une représentation distributionnelle (sous la forme d'un vecteur, d'une matrice ou d'une fonction) des schémas morphologiques, afin de construire ou reconstruire les représentations vectorielles des mots morphologiquement construits (Baroni et Zamparelli 2010, Lazaridou *et al.* 2013, Marelli et Baroni 2015). La réflexion autour des affixes et de leurs propriétés sémantiques se retrouve aussi dans le cadre du traitement de la concurrence affixale, et plus largement de la comparaison de schémas morphologiques de formation de lexèmes. Si la concurrence affixale est déjà largement

étudiée dans une approche quantitative (Lindsay 2012, Bonami et Thuilier 2019, Naccarato 2019, Bobkova et Montermini 2020, Missud et Villoing 2020, *inter alia*), les propriétés sémantiques n'étaient pas prises en compte de façon systématique. Les études exploitant la sémantique distributionnelle se développent désormais et proposent ainsi la comparaison des propriétés distributionnelles (et donc sémantiques) des lexèmes construits. Ces études cherchent notamment à faire émerger des différences plus au moins importantes entre schémas afin d'établir le degré de rivalité de paires de schémas deux à deux (Missud 2019, Guzmán Naranjo et Bonami 2020, Huyghe et Wauquier 2020, Wauquier *et al.* 2020a, 2020b, Huyghe et Wauquier 2021a, 2021b, Varvara *et al.* 2021). Ces approches permettent des analyses sémantiques à des degrés variables permettant notamment l'étude de la valence émotionnelle et axiologique associée à certains schémas morphologiques (Lapesa *et al.* 2017, Wauquier et Bonami 2021).

17 Les dispositifs distributionnels ne se limitent pas à l'étude des affixes mais soutiennent aussi l'analyse sémantique des mots construits eux-mêmes. Cotterel et Schütze (2018) proposent ainsi d'explorer la compositionnalité, tant morphologique que sémantique, des lexèmes construits. Leur modèle permet en outre de comparer dans un contexte distributionnel la productivité de divers affixes. Notons qu'un type particulier de modèles dits CDSM, pour *compositional distributional semantic models* a émergé, offrant une représentation distributionnelle construite sur l'idée de compositionnalité (Mitchell et Lapata 2010, Melymuka *et al.* 2017, entre autres). Lapesa *et al.* (2018) se basent sur les représentations vectorielles pour désambiguiser en contexte des nominalisations néologiques en *-ment* en anglais. Le caractère événementiel ou non événementiel du néologisme est ainsi prédit sur la base de son vecteur. Plus largement, les représentations vectorielles sont utilisées pour opérer des classifications sémantiques de dérivés morphologiques. Verhoeven *et al.* (2012) évaluent à ce titre la similarité de composés binominaux en néerlandais et en afrikaans à partir de la similarité des vecteurs de leurs composants. Huyghe et Wauquier (2020) évaluent quant à eux la similarité de noms candidats à l'agentivité à des représentations prototypiques de catégories sémantiques de noms d'humain et d'agent calculées à partir de vecteurs de membres de ces classes.

18 Outre les propriétés sémantiques du dérivé, le lien sémantique entre les membres d'une famille morphologique donnée bénéficie aussi, de façon variée, de l'approche distributionnelle. Wauquier (2020) propose ainsi de quantifier la similarité sémantique entre un verbe et ses dérivés agentifs et processifs pour tester l'hypothèse d'une plus grande distanciation sémantique du nom d'agent dérivé. Padó *et al.* (2015) et Kisselew *et al.* (2016) proposent quant à eux d'utiliser la proximité distributionnelle comme un indice pour définir l'orientation de procédés dérivationnels comme la conversion, c'est-à-dire identifier la base et le dérivé. Ces travaux reposent sur l'hypothèse que les noms dérivés tendent à être moins fréquents et sémantiquement plus spécifiques que leur base. La spécificité est à ce titre calculée à partir de propriétés distributionnelles issues des vecteurs des bases et des dérivés. Lombard *et al.* (2021) se proposent quant à eux d'étudier la démotivation morphosémantique, c'est-à-dire la perte d'un lien sémantique entre une base et son dérivé, en quantifiant le degré de démotivation de paires en synchronie au travers de la proximité distributionnelle. Plus largement, toutes ces études reposent sur l'évaluation de la transparence sémantique d'un mot construit vis-à-vis de sa base (Gagné *et al.* 2016, Varvara *et al.* 2021, *inter alia*).

19 Plus à la marge, il est intéressant de noter que certains travaux ont aussi utilisé la sémantique distributionnelle pour étudier l'intersection entre morphologie flexionnelle et dérivationnelle. Varvara *et al.* (2021) et Mickus *et al.* (2019) interrogent ainsi la notion de binarité et d'exclusion mutuelle entre flexion et dérivation par l'analyse de phénomènes à l'intersection – au travers respectivement de l'étude de la concurrence entre schémas morphologiques, et de l'analyse du genre des noms et des adjectifs. Bonami et Paperno (2018) questionnent quant à eux la régularité sémantique supposée distinctive de la flexion et de la dérivation en comparant sur le plan distributionnel et de façon systématique des bases et des mots construits ou fléchis.

2.2. Des méthodes à construire

20 Les représentations distributionnelles offrent, par ailleurs, une grande flexibilité quant à leur exploitation. Cette exploitation repose sur des méthodes à la complexité variable, qui permettent une analyse de mots morphologiquement construits à différents niveaux de granularité.

21 Une première façon, et sans doute la plus simple, d'exploiter les représentations distributionnelles pour l'analyse linguistique de mots morphologiquement construits est de reposer sur la mesure de la proximité distributionnelle. Cette méthode ne requiert pas la manipulation des vecteurs, simplement l'interrogation de modèles distributionnels, qui sont de plus en plus accessibles en ligne. Outre sa simplicité de mise en œuvre, cette approche permet de quantifier et de comparer la similarité sémantique de mots construits entre eux ou vis-à-vis de leur base. Les études sur la transparence sémantique entre bases et dérivés telles que proposées par Wauquier (2020), Lombard *et al.* (2021) ou Varvara *et al.* (2021) reposent notamment sur cette approche. Il est au même titre tout aussi aisé de récupérer les voisins distributionnels d'un mot, sur la base des scores de proximité. Cette approche offre, en plus de la quantification de la similarité d'un mot cible à un ensemble de mots voisins, une analyse plus qualitative puisqu'elle fait émerger les mots qui présentent un profil distributionnel similaire en corpus. Les études de Wauquier (2020), Wauquier *et al.* (2020a, 2020b) et Huyghe et Wauquier (2020, 2021a, 2021b) reposent notamment sur cette approche. Notons que certains travaux exploitent des mesures ne correspondant pas à des scores de proximité, mais calculées à partir des vecteurs. C'est le cas notamment de Varvara *et al.* (2021) qui en plus du score cosinus, calculent un score d'inclusion obtenus à partir de la différence des valeurs entre deux vecteurs.

22 Mais il est aussi possible d'aller plus loin en manipulant les vecteurs. En effet, les vecteurs peuvent être combinés pour construire des abstractions qui sont ensuite utilisés comme des vecteurs classiques. Cette approche est notamment utilisée par Huyghe et Wauquier (2020, 2021a, 2021b), Wauquier (2020) et Wauquier *et al.* (2020a, 2020b) qui, pour mettre en évidence des propriétés sémantiques partagées par un ensemble de mots, construisent un vecteur moyen, dit barycentre, à partir des vecteurs desdits mots. À l'inverse, Bonami et Paperno (2018) et Mickus *et al.* (2019) proposent de construire des vecteurs de différence, en soustrayant le vecteur d'une base au vecteur d'un dérivé ou d'un mot fléchi pour ne conserver que le vecteur lié à la transformation. Ces vecteurs de différence sont alors comparés, entre eux ou vis-à-vis d'un vecteur de différence moyen, pour évaluer la régularité de la transformation sur la base de la distance de ces vecteurs de différence par rapport au vecteur de référence.

23 Il est aussi possible d'intégrer les représentations distributionnelles dans des tâches de *machine learning*. Huyghe et Wauquier (2021b) proposent ainsi d'opérer un *clustering* de noms d'agent du français construits par des schémas dérivationnels distincts pour en étudier les propriétés morphosémantiques, et ce à partir des vecteurs des noms d'agent. Chez Guzmán Naranjo et Bonami (2020), le clustering est réalisé sur la base des vecteurs de différence et non des vecteurs de mots. Wauquier et Bonami (2021) proposent quant à eux de prédire le suffixe d'un mot construit à partir de son vecteur.

3. Considérations linguistiques

24 Toute puissante soit-elle, la sémantique distributionnelle ne saurait échapper à une expertise linguistique rigoureuse, dont l'impact sur les analyses inférées à partir des représentations distributionnelles n'est pas négligeable. Le choix du corpus et des données morphologiques n'est, à ce titre, pas anodin.

3.1. Le corpus

25 Les représentations distributionnelles sont intrinsèquement dépendantes du corpus sur lequel elles sont entraînées. Il s'agit de fait du premier levier linguistique sur lequel on peut agir, et ce par plusieurs biais.

26 La question du choix du corpus se pose vis-à-vis de deux paramètres principaux que sont sa nature et sa taille. Le registre et la thématique du corpus ont, en effet, un impact direct sur les contextes linguistiques sur lesquels les représentations sont construites. Ainsi, Wauquier (2018) montre que le contenu sémantique encapsulé par le vecteur moyen de noms d'agent et d'instrument en *-eur* calculé dans un corpus encyclopédique et dans un corpus du web diffère sensiblement, les voisins de ce vecteur moyen se distinguant par leur degré de généricité et par leurs propriétés ontologiques. Le choix du type de corpus dépend cependant de considérations quantitatives objectives, à savoir la taille du corpus. Si les modèles d'analyse sémantique latente peuvent se contenter de corpus modestes, les modèles prédictifs et les modèles de langues pré-entraînés nécessitent de grandes quantités de données (plusieurs centaines de milliers, voire plusieurs milliards de mots). Cette contrainte limite de fait le nombre de corpus potentiels, ou les langues qu'il est possible d'étudier.

27 Outre la question de la nature du corpus, se pose aussi la question de son traitement. S'il est en effet possible d'entraîner un modèle à partir d'un modèle brut, il peut être souhaitable de le lemmatiser au préalable, voire de l'étiqueter morphosyntaxiquement. Cela permet ainsi de construire la représentation d'un lexème sur la base de l'ensemble de ses occurrences, fléchies ou non. L'étiquetage morphosyntaxique peut quant à lui permettre au contraire de distinguer des formes qui seraient syntaxiquement ambiguës (à l'image de cas de transposition ou de conversion). Le choix de lemmatiser et/ou d'étiqueter implique cependant de dépendre de l'annotation morphosyntaxique et de la lemmatisation nécessairement automatiques puisque réalisées sur de grands corpus. Cela peut donc impliquer du bruit, dont il s'agira d'évaluer l'impact, négligeable ou non, en fonction de l'étude envisagée. À titre d'exemple, l'intérêt de la lemmatisation est ainsi fortement discuté, notamment en regard d'études sur le genre (Gonen *et al.* 2019, Wauquier 2020). Si elle permet la neutralisation du genre grammatical en

contexte, elle entraîne souvent un effacement des noms lexicalement féminins, certains noms féminins étant remplacés par leur équivalent masculin.

28 Notons qu'à la marge du choix du corpus se pose la problématique des biais appris par les modèles distributionnels. De nombreux travaux ont ainsi montré que les représentations distributionnelles étaient sensibles aux biais de genre ou de race, intégrant dans leur représentation des préjugés socio-culturels tels que le caractère prototypiquement féminin ou masculin de certains métiers (Bolukbasi *et al.* 2016, Caliskan *et al.* 2017). Bien qu'ils y soient moins sensibles, les *embeddings* contextuels ne sont cependant pas exempts de ces biais (Basta *et al.* 2019), ce qui se traduit par des performances altérées dans les tâches basées sur des représentations distributionnelles. Par exemple, la performance des systèmes de résolution de coréférence sera moindre lorsqu'il s'agira de prédire une entité féminine qu'une entité masculine (Zhao *et al.* 2019). C'est un phénomène issu des corpus qu'il est nécessaire de prendre en compte lors d'analyses sémantiques fines, qu'il soit indésirable ou non.

3.2. Les données morphologiques

29 De par leur implémentation, il est aisé de projeter de grandes listes de mots dans un modèle distributionnel. Mais la sélection des données morphologiques pose elle aussi des questions méthodologiques. Si certaines problématiques sont spécifiques à une méthode donnée (voir Wauquier 2020 pour la gestion de la taille et de la cohérence sémantique des amorces des vecteurs moyens), nous abordons ici deux critères principaux que sont la polysémie et la fréquence.

30 Comme évoqué en section 1.2., les modèles *count-based* et prédictifs ne permettent pas la gestion de la polysémie et de l'homonymie, et une représentation unique est proposée pour une forme donnée. Le vecteur de la forme *vol* est calculé sur la base de l'ensemble des occurrences de *vol*, mélangeant les occurrences et donc les informations distributionnelles du lexème VOL_1 (synonyme de *larcin*) et du lexème VOL_2 (synonyme de *déplacement aérien*). Une solution consiste à choisir un modèle contextuel, qui permet de passer outre le problème de la polysémie, du moins partiellement (Yenicelik *et al.* 2020), mais cela implique d'autres contraintes techniques (en termes de corpus et de computation notamment). Une autre option consiste à sélectionner des mots strictement monosémiques (à l'image de Huyghe et Wauquier 2020, 2021a, 2021b), avec néanmoins les conséquences que cela implique en termes de taille d'échantillon. Enfin, il est possible de conserver les formes polysémiques, mais il faut alors accepter que les résultats en soient parfois affectés, à l'image de la proximité entre *port* et *porter* dans (Wauquier 2020). L'étude de la démotivation morphosémantique, et de la perte de lien sémantique entre un verbe et son dérivé nominal, à l'aide de méthodes expérimentales et distributionnelles par Lombard *et al.* (2021) souffre ainsi de la polysémie de verbes comme *abattre* et *raser* qui affecte leur proximité aux noms *abattoir* et *rasoir*, et qui mène à un désaccord entre les différentes mesures utilisées.

31 La fréquence est un autre critère de sélection non négligeable dès lors que l'on utilise des modèles distributionnels. En effet, la fréquence d'un mot a un impact direct sur sa représentation vectorielle (Tanguy *et al.* 2015), en limitant le nombre de contextes et en contribuant à définir la longueur de son vecteur. Dans les faits, plus la fréquence d'un mot est faible, moins sa représentation vectorielle sera robuste. De nombreux travaux décident à ce titre de fixer un seuil de fréquence minimum élevé (Sahlgren 2006) ou de

créer des groupes de fréquence visant à obtenir des conditions expérimentales comparables. Cela peut rendre la comparaison de bases et de dérivés problématiques, si l'on part de l'hypothèse qu'un dérivé est moins fréquent que sa base, et cela rend l'étude de la néologie complexe voire impossible, à moins d'utiliser des méthodes compositionnelles permettant de (re)construire une représentation vectorielle, mais qui ne traduit pas des usages réels en corpus.

Conclusion

32 Nous avons dressé dans cet article un état des lieux succinct de l'utilisation de la sémantique distributionnelle pour la morphologie dérivationnelle. Nous avons notamment exposé les fondements et implémentations de cette approche et avons présenté la diversité des thématiques et méthodes qu'elle offre. Nous avons ainsi essayé de montrer que la sémantique distributionnelle était un outil permettant de soutenir une analyse linguistique variée et riche. Ce faisant, nous avons aussi souligné l'importance de considérations linguistiques comme garde-fou dans l'utilisation de cet outil, qui dépend donc largement des *inputs* linguistiques pour permettre la production de connaissances nouvelles.

33 Comme cela est brièvement esquissé au travers de cet article, la sémantique distributionnelle est un champ qui évolue très vite, de nouveaux outils chassant régulièrement des outils que la recherche en linguistique a à peine le temps d'effleurer, encore moins de s'approprier. Les études qui se développent suggèrent néanmoins que cette approche est porteuse, à même de contribuer à la révolution empirique de la morphologie, à condition qu'une approche raisonnée et rigoureuse soit adoptée.

BIBLIOGRAPHIE

Avraham O. & Goldberg Y. (2017). « The interplay of semantics and morphology in word embeddings », *arXiv preprint arXiv : 1704.01938*.

Baayen R. H., Piepenbrock R. & Gulikers L. (1995). « CELEX2 LDC96L14 », Web Download. Philadelphia : Linguistic Data Consortium.

Baroni M., Bernardi R. & Zamparelli R. (2014a). « Frege in space : A program of compositional distributional semantics », *LiLT (Linguistic Issues in Language Technology)* 9.

Baroni M., Dinu G. & Kruszewski G. (2014b). « Don't count, predict ! a systematic comparison of context-counting vs. context-predicting semantic vectors », *in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 238-247.

Baroni M. & Zamparelli R. (2010). « Nouns are vectors, adjectives are matrices : Representing adjective-noun constructions in semantic space », *in Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Cambridge : Association for Computational Linguistics, 1183-1193.

Basta, C., Costa-Jussà M. R. & Casas N. (2019). « Evaluating the underlying gender bias in contextualized word embeddings », *arXiv preprint arXiv : 1904.08783*.

Bernier-Colborne G. & Drouin P. (2016). « Évaluation des modèles sémantiques distributionnels : le cas de la dérivation syntaxique », *in Proceedings of the 23rd French Conference on Natural Language Processing (TALN)*, 125-138.

Bobkova N. & Montermini F. (2020). « Suffix rivalry in russian : what low frequency words tell us », *in* J. Audring, N. Koutsoukos et C. Manouilidou (éd.) *Rules, patterns, schemas and analogy, MMM12 Online Proceedings,* volume 12, 1-17.

Boleda G. (2020). « Distributional semantics and linguistic theory », *Annual Review of Linguistics* 6 : 213-234.

Bojanowski P., Grave E., Joulin A. & Mikolov T. (2016). « Enriching word vectors with subword information », *arXiv preprint arXiv* : 1607.04606.

Bolukbasi, T., Chang K.-W., Zou J. Y., Saligrama V. & Kalai A. T. (2016). « Man is to computer programmer as woman is to home-maker ? debiasing word embeddings », *in Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS)*. Barcelone, 4349-4357.

Bonami O., & Paperno D. (2018). « Inflection vs. derivation in a distributional vector space », *Lingue e linguaggio 17*(2) : 173-196.

Bonami O. & Thuilier J. (2019). « A statistical approach to rivalry in lexeme formation : French -iser and -ifier », *Word Structure* 12(1) : 4-41.

Botha J. & Blunsom P. (2014). « Compositional morphology for word representations and language modelling », *in International Conference on Machine Learning*, 1899-1907.

Calderone, B., Sajous F. & Hathout N. (2016). « GLAW-IT : A free large Italian dictionary encoded in a fine-grained XML format », *in Proceedings of the 49th Annual Meeting of the Societas Linguistica Europaea (SLE 2016).* Naples, 43-45.

Caliskan A., Bryson J. J. & Narayanan A. (2017). « Semantics derived automatically from language corpora contain human-like biases », *Science 356* (6334) : 183-186.

Cotterell R. & Schütze H. (2018). « Joint Semantic Synthesis and Morphological Analysis of the Derived Word », *Transactions of the Association for Computational Linguistics* 6 : 33-48.

Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K. & Harshman R. (1990). « Indexing by latent semantic analysis », *Journal of the American Society for Information Science 41*(6) : 391-407.

Devlin J., Chang M.-W., Lee K. & Toutanova K. (2018). « Bert : Pretraining of deep bidirectional transformers for language understanding », *arXiv preprint arXiv* : 1810.04805.

Firth J. R. (1957). « A synopsis of linguistic theory, 1930-1955 », *in* J.R. Firth (éd.) *Studies in Linguistic Analysis*. Oxford : Basil Blackwell, 1-32.

Gagné C. L., Spalding T. L., & Nisbet K. A. (2016). « Processing English compounds : Investigating semantic transparency », *SKASE Journal of Theoretical Linguistics 13*(2).

Gonen H., Kementchedjhieva Y. & Goldberg Y. (2019). « How does grammatical gender affect noun representations in gender-marking languages ? », *arXiv preprint arXiv :*1910.14161.

Guzmán Naranjo M. & Bonami O. (2020). « Distributional assessment of derivational semantics », Communication orale présentée à SLE 2020.

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

Hathout N. et Namer F. (2016). « Giving lexical resources a second life : Démonette, a multi-sourced morpho-semantic network for French », *in* N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk et S. Piperidis (éd.) *10th International Conference on Language Resources and Evaluation (LREC 2016).* Portorož, 1084-1091.

Hathout N., Namer F. & Dal G. (2002). « An Experimental Constructio nal Database : the MorTAL Project », *in* P. Boucher (éd.) *Many morphologies*. Sommerville : Cascadilla Press, 178-209.

Hathout, N. et Sajous, F. (2016). « Wiktionnaire's wikicode glawified : a workable french machine-readable dictionary », *in* N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk et S. Piperidis (éd.) *10th International Conference on Language Resources and Evaluation (LREC 2016).* Portorož, 1369-1376.

Huyghe R. & Wauquier M. (2020). « What's in an agent ? a distributional semantics approach to agent nouns in French », *Morphology* 30 : 185-218.

Huyghe R. & Wauquier M. (2021a). « Une étude distributionnelle des noms d'agent en *-ant*, *-eur*, *-ien*, *-ier* et *-iste* », *Verbum*, XLIII, 1 : 13-40.

Huyghe R. & Wauquier M. (2021b). « Distributional semantics insights on agentive suffix rivalry in French », *Word Structure* 14(3) : 354-391.

Kisselew M., Rimell L., Palmer A. & Padó S. (2016). « Predicting the direction of derivation in english conversion », *in Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Berlin, 93-98.

Koehl A. (2013). « Une base de données des noms désadjectivaux du français : le modele mordan », *in Proceedings of Corpus et Outils en Linguistique, langues et parole.* Strasbourg, 3-5.

Landauer T. K. & Dumais S. T. (1997). « A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge », *Psychological review* 104(2) : 211-240.

Lapesa G., Padó S., Pross T. & Roßdeutscher A. (2017). « Are doggies really nicer than dogs ? The impact of morphological derivation on emotional valence in German », *in IWCS 2017—12th International Conference on Computational Semantics—Short papers.*

Lapesa G., Kawaletz L., Plag I., Andreou M., Kisselew M. & Padó S. (2018). « Disambiguation of newly derived nominalizations in context : A distributional semantics approach », *Word Structure* 11(3) : 277-312.

Lazaridou A., Marelli M., Zamparelli R. & Baroni M. (2013). « Compositional-ly derived representations of morphologically complex words in distributional semantics », *in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. Sofia, 1517-1526.

Lenci A. (2018). « Distributional models of word meaning », *Annual Review of Linguistics* 4 : 151-171.

Levy O. & Goldberg Y. (2014). « Dependency-based word embeddings », *in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*. Baltimore, 302-308.

Levy O., Goldberg Y. & Dagan I. (2015). « Improving distributional similarity with lessons learned from word embeddings », *Transactions of the Association for Computational Linguistics* 3 : 211-225.

Lindsay M. (2012). « Rival suffixes : synonymy, competition, and the emergence of productivity », *in Mediterranean Morphology Meetings, volume 8*, 192-203.

Lombard A., Wauquier M., Fabre C., Hathout N., Ho-Dac M. & Huyghe R. (2021). « Evaluating morphosemantic demotivation through experimental and distributional methods », Communication orale présentée à la troisième édition de l'International Symposium of Morphology (ISMo), en ligne, septembre 2021.

Luong T., Socher R. & Manning C. D. (2013). « Better word representations with recursive neural networks for morphology », *in Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Sofia, 104-113.

Marelli M. & Baroni M. (2015). « Affixation in semantic space : Modeling morpheme meanings with compositional distributional semantics », *Psychological review* 122(3) : 485-515.

Melymuka M., Lapesa G., Kisselew M. & Padó S. (2017). « Modeling Derivational Morphology in Ukrainian », *in IWCS 2017—12th International Conference on Computational Semantics—Short papers.*

Mickus T., Bonami O. & Paperno D. (2019). « Distributional effects of gender contrasts across categories », *in Proceedings of the Society for Computation in Linguistics, volume 2*, 174-184.

Mickus T., Paperno D., Constant M. & van Deemter K. (2020). « What do you mean, BERT? », *in Proceedings of the Society for Computation in Linguistics 2020*, 235-245.

Mikolov T., Chen K., Corrado G. & Dean J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv : 1301.3781*.

Missud A. (2019). *Modélisation quantitative de la rivalité entre la suffixation en -age et la conversion de verbe à nom*. Mémoire de master, Université Paris Nanterre, Paris.

Missud A., Amsili P. & Villoing F. (2020). « VerNom : une base de paires morphologiques acquise sur très gros corpus (VerNom : a French derivational database acquired on a massive corpus) », *in Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, 305-313.

Missud A. & Villoing F. (2020). « The morphology of rival *-ion*, *-age* and *-ment* selected verbal bases », *Lexique* 26 : 29-52.

Mitchell J. & Lapata M. (2010). « Composition in distributional models of semantics », *Cognitive science*, 34(8) : 1388-1429.

Naccarato C. (2019). « Agentive (para) synthetic compounds in Russian : a quantitative study of rival constructions », *Morphology* 29(1) : 1-30.

Padó S., Palmer A., Kisselew M. & Šnajder J. (2015). « Measuring semantic content to assess asymmetry in derivation », *in Workshop on Advances in Distributional Semantics*.

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

Sahlgren M. (2008). « The distributional hypothesis », *Italian Journal of Disability Studies* 20 : 33-53.

Sajous F., Calderone B. & Hathout N. (2020). « ENGLAWI : From Human- to Machine-Readable Wiktionary », *in Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, 3016-3026.

Soricut R. & Och F. J. (2015). « Unsupervised morphology induction using word embeddings », *in Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. Denver, 1627-1637.

Tanguy L., Sajous F. & Hathout N. (2015). « Évaluation sur mesure de modèles distributionnels sur un corpus spécialisé : comparaison des approches par contextes syntaxiques et par fenêtres graphiques », *Traitement Automatique des Langues* 56(2) : 103-127.

Turney P. D. & Pantel P. (2010). « From frequency to meaning : Vector space models of semantics », *Journal of artificial intelligence research* 37 : 141-188.

Varvara R., Lapesa G. & Padó S. (2021). « Grounding Semantic Transparency In Context : A Distributional Semantic Study on German Event Nominalizations », *Morphology 31*(4) : 409-446.

Verhoeven B., Daelemans W. & van Huyssteen G. (2012). « Classification of noun-noun compound semantics in Dutch and Afrikaans », *in Proceedings of the 23rd Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*. Pretoria, 121-125.

Wauquier M. (2018). « Analyse des noms agentifs dans les espaces vectoriels distributionnels », *in Actes de la Conférence RJC 2018, conjointe à TALN et CORIA 2018*. Rennes, 27-39.

Wauquier M. (2020). *Confrontation des procédés dérivationnels et des catégories sémantiques dans les espaces distributionnels*. Thèse de doctorat, Université Toulouse Jean Jaurès.

Wauquier M. & Bonami O. (2021). « Social gender and redivational morphology : a distributional study of the gendered import of learned morphology in French », Communication orale présentée à la troisième édition de l'International Symposium of Morphology (ISMo), en ligne, septembre 2021.

Wauquier M., Hathout N. & Fabre C. (2020a). « Contributions of distributional semantics to the semantic study of French morphologically derived agent nouns », *in* J. Audring, N. Koutsoukos et C. Manouilidou (éd.) *Rules, patterns, schemas and analogy, MMM12 Online Proceedings, volume 12*, 111-121.

Wauquier M., Hathout N. & Fabre C. (2020b). « Semantic discrimination of technicality in French nominalizations », *Zeitschrift für Wortbildung / Journal of Word Formation* 4(2) : 100-119.

Yenicelik D., Schmidt F. & Kilcher Y. (2020). « How does BERT capture semantics ? A closer look at polysemous words », *in Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, en ligne, 156-162.

Zeller B., Šnajder J. & Padó S. (2013). « DErivBase : Inducing and Evaluating a Derivational Morphology Resource for German », *in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, 1201-1211.

Zeller B., Padó S. & Šnajder J. (2014). « Towards Semantic Validation of a Derivational Lexicon », *in Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014) : Technical Papers*, 1728-1739.

Zhao J., Wang T., Yatskar M., Cotterell R., Ordonez V. & Chang K. W. (2019). « Gender bias in contextualized word embeddings. *arXiv preprint arXiv : 1904.03310*.

RÉSUMÉS

Cet article dresse un état des lieux de l'utilisation de la sémantique distributionnelle en morphologie. L'approche distributionnelle, qui repose sur une représentation vectorielle du sens des mots, s'intègre dans l'évolution empirique que connaît la morphologie depuis quelques années, en contribuant par une analyse quantitative et basée sur les corpus du sens des mots morphologiquement construits. Nous présentons brièvement cette approche, puis nous donnons un aperçu de la diversité de ses utilisations pour la morphologie dérivationnelle, tant théorique

que méthodologique. Nous soulignons enfin l'importance des choix linguistiques dans l'utilisation de cet outil.

This paper gives an overview of the use of distributional semantics in morphology studies. Based on vectorial representation of meaning, the distributional approach contributes to the empirical evolution morphology is undergoing for the last few years, by allowing a quantitative analysis, based on corpora, of the meaning of morphologically constructed words. We briefly present this approach, then we provide an overview of its diversity of uses, both in theoretical and methodological terms. Finally, we highlight the importance of linguistic choices when using distributional semantics.

INDEX

Mots-clés : sémantique distributionnelle, morphologie computationnelle, dérivation
Keywords : distributional semantics, computational morphology, derivation

AUTEUR

MARINE WAUQUIER

Université de Paris, CNRS, Laboratoire de linguistique formelle