Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

VerNom : une base de paires morphologiques acquise sur très gros corpus

Alice Missud[1, 2] Pascal Amsili[2] Florence Villoing[1]
(1) Modyco (UMR 7114 CNRS/Paris Nanterre)
(2) Lattice (UMR 8094 CNRS/Paris 3/ENS)
missud.a@parisnanterre.fr, pascal.amsili@gmx.fr,
villoing@parisnanterre.fr

RÉSUMÉ

Alors qu'une part active de la recherche en morphologie dérivationnelle s'intéresse à la compétition qui oppose les suffixations construisant des noms d'événement à partir de verbes (*-age, -ment, -ion, -ure, -ance, -ade, -aison*), l'accès à des données en large quantité devient nécessaire pour l'application de méthodes quantitatives. Dans l'optique de réunir des paires de verbes et de noms morphologiquement reliés dans le cadre de ces suffixations rivales, nous présentons VerNom, une base morphologique comprenant 25 857 paires verbe-nom, construite automatiquement à partir d'un corpus massif issu du web.

ABSTRACT

VerNom : a French derivational database acquired on a massive corpus

While a significant part of the litterature in word formation revolves around the competition between suffixations that construct event nouns from verbs in French (*-age, -ment, -ion, -ure, -ance, -ade, -aison*), accessing massive data is necessary for applying quantitative methods. With the purpose of gathering pairs of verbs and nouns that are morphologically related in the case of this competition, we present VerNom, a lexical database consisting of 25 875 verb-noun pairs acquired automatically from a massive web corpus.

MOTS-CLÉS : morphologie dérivationnelle, compétition morphologique, nominalisation, base lexicale.

KEYWORDS: word formation, morphological rivalry, nominalization, lexical database.

1 Introduction

Nous présentons VerNom, une ressource morphologique comprenant un ensemble de paires verbe-nom (et leurs fréquences d'apparition) collectées dans un large corpus issu du web. Les paires ont été appariées morphologiquement de façon automatique, et elles recouvrent les principales dérivations construisant des noms d'événement à partir de verbes : les suffixations en *-ion, -age, -ment, -ure, -ance, -ade* et *-aison*.

Les données utilisées proviennent de frCOW (Schäfer & Bildhauer 2012, Schäfer 2015), un corpus de 9 milliards de mots issu du web francophone. La constitution de cette ressource a pour objectif de donner accès à une grande quantité de données "authentiques" du français écrit, pour permettre

une analyse quantitative et qualitative de la compétition entre nominalisations rivales du français (construisant des noms déverbaux événementiels), qui constitue une part active de la recherche actuelle en morphologie dérivationnelle (Lüdtke 1978 ; Martin 2010 ; Ferret *et al.* 2010 ; Uth 2010 ; Ferret & Villoing 2012 ; Uth 2016 ; Fradin 2014, 2019 ; Dal *et al.* 2018 : Wauquier *et al.* 2018, 2019).

VerNom s'inscrit dans un contexte marqué par une relative carence, pour le français, de ressources purement dérivationnelles permettant des recherches avec des méthodes de modélisation quantitative. En effet, alors que les travaux sur la compétition morphologique tendent de plus en plus à s'appuyer sur ce type de méthode (par exemple, Wauquier *et al.* 2018, Bonami & Thuilier 2019), on recense principalement deux ressources exploitables pour le français, le corpus Nomage (Balvet *et al.* 2011) et la base lexicale Démonette (Hathout & Namer 2014, Namer *et al.* 2019), qui s'appuie elle-même sur Verbaction (Hathout & Tanguy 2002). Bien que ces ressources soient extrêmement précieuses, en particulier parce qu'elles fournissent des annotations fines tant syntaxiquement que sémantiquement, voire morphologiquement et phonologiquement pour la seconde (qui en outre construit tout un réseau dérivationnel), elles ne répondent pas entièrement aux besoins d'une analyse quantitative à gros grain. En effet, la minutie des informations renseignées par ces ressources les conduit à traiter un jeu de données assez réduit (736 noms événementiels et 679 verbes pour Nomage et 8848 paires V/N d'action pour Démonette). Par ailleurs, leurs lexiques n'enregistrent que peu de néologismes, indispensables au calcul de la productivité des schémas morphologiques. Cette lacune tient au fait qu'ils sont issus soit du journal *Le Monde* (cf. Nomage, basé sur le French Treebanck d'Abeillé *et al.* 2003), soit de dictionnaires (cf. Démonette qui compte 6765 paires V/N d'action tirées de TLFnome). L'ensemble de ces contraintes nous a conduit à élaborer une nouvelle ressource qui réponde à ces objectifs spécifiques.

À la suite de Hathout *et al.* (2009), qui ont exploité le web pour collecter une liste extensive de paires verbe-nom d'action (notamment les suffixations en *-age, -ment* et *-ion*) en procédant à l'aide de règles, nous proposons de réactualiser les méthodes en nous basant sur un corpus plus massif et plus récent (2016) afin de récolter une plus grande quantité de données, plus diverses, regroupant toutes les suffixations déverbales construisant des noms d'événement, et qui permette en outre d'explorer de nombreux néologismes.

Dans cet article, nous décrivons la ressource en détail, les méthodes d'extraction des noms et des verbes et leur appariement, ainsi que l'évaluation des paires collectées. Enfin, nous présentons une brève étude de la productivité des schémas morphologiques basée sur la ressource.

2 Méthodologie

2.1 Extraction des données

La base a été constituée à partir de frCOW16 (Schäfer & Bildhauer 2012, Schäfer 2015), un corpus massif du français issu du web datant de 2016 et comprenant 9 milliards de mots. Les corpus COW sont catégorisés et lemmatisés, généralement avec TreeTagger, de telle sorte que tous les tokens reçoivent une catégorie (il n'y a pas de POS-tag 'unknown') mais pas forcément toujours un lemme. Nous avons fait le choix d'exploiter directement ces informations de catégorie et lemmatisation, sans en contrôler au préalable la qualité. Puisque nous cherchions à apparier des verbes et des noms reliés morphologiquement, nous avons extrait toutes les formes catégorisées 'verbe' ou 'nom', que ces formes aient été lemmatisées ou non (les formes non lemmatisées nous intéressent particulièrement

puisqu'elles peuvent correspondre à des néologismes qui échappent aux lexiques construits à partir de dictionnaires). La table 1 présente des exemples d'entrées du corpus frCOW comprenant des formes lemmatisées et non lemmatisées. Ces deux groupes de données ont été traités séparément dans les phases suivantes.

forme	catégorie	lemme
mangera	VER	manger
voiture	NOM	voiture
pourcenter	VER	(unknown)
pipolisation	NOM	(unknown)

TABLE 1 – Exemples d'entrées de frCOW

2.2 Nettoyage des formes

Nous avons dû procéder à toutes les étapes à l'élimination de lemmes/formes contenant des caractères spéciaux, de la ponctuation ou des séquences impossibles.

Formes lemmatisées Pour les verbes lemmatisés, les formes associées à des lemmes se terminant en *-er*, *-ir* ou *-re* ont été récupérées. Pour les noms lemmatisés, les formes extraites ont un lemme se terminant par l'une des 7 suffixations constructrices de noms d'événement à savoir *-age, -ment, -ion, -ure, -ance, -ade* et *-aison*. La suffixation en *-erie*, susceptible de construire des noms d'événement également, a cependant été mise de côté en raison de la difficulté de différencier automatiquement son homonyme plus productif dérivant des noms de lieux à partir de noms. Au total, 25 209 verbes et 23 200 noms lemmatisés ont été extraits.

Formes non lemmatisées La non-lemmatisation de certaines formes dans frCOW peut dépendre de plusieurs problèmes : les formes peuvent être mal catégorisées, et donc ne pas correspondre avec un lemme (par exemple : *expressément*, catégorisé 'nom' plutôt que 'adverbe'), elles peuvent également correspondre à des variantes orthographiques qui les distancient d'un lemme (par exemple : *develloppé* plutôt que *développé*), ou encore constituer des néologismes non répertoriés (comme *pipolisation*). Nous avons tenté d'une part de rassembler les variantes orthographiques avec les bons lemmes, et d'autre part de lemmatiser les formes susceptibles de correspondre à des néologismes.

Seules les formes étiquetées 'verbe' avec une finale en *-er* ont été extraites parmi les verbes non lemmatisés, car la construction de verbes irréguliers du deuxième et troisième groupe est peu probable parmi les néologismes. Pour les noms, seules les formes en *-ion, -age, -ment, -ure, -ance, -ade* et *-(a)ison* ont été récoltées. Afin d'obtenir les fréquences des formes fléchies de ces entrées, nous avons utilisé des expressions régulières pour retrouver dans frCOW des formes fléchies pour chaque verbe et chaque nom. Les fréquences des formes fléchies trouvées ont été ajoutées aux lemmes. Ceci a permis, par exemple pour *wikifier*, de passer d'une fréquence de 11 à 18. Pour réunir les variantes orthographiques d'un même lemme, nous avons d'abord tenté d'ignorer les diacritiques, sujets à de nombreuses erreurs d'orthographe en français, afin de regrouper ensemble les formes non lemmatisées avec les bons lemmes. Nous avons supprimé tous les diacritiques des deux lexiques, et avons regardé si parmi les formes non préalablement lemmatisées se trouvait un équivalent identique chez les formes

lemmatisées. Ceci a permis d'ajouter des fréquences aux noms et verbes lemmatisés comme *étudier* par exemple, dont la fréquence a été augmentée de 5 occurrences. Au total, 96 770 verbes et 130 341 noms non préalablement lemmatisés ont été récoltés. L'ensemble des verbes et des noms décrits a ensuite servi à l'appariement morphologique. L'influence des différentes étapes de nettoyage sur les proportions des lexiques des formes non lemmatisées est détaillée dans la table 2.

Mots-forme non lemmatisés	VERBES	NOMS
Se terminent en *-er*	140 036	-
Se terminent en *-ion, -age, -ment, -ure, -ance, -ade, -aison*	-	225 177
Retrait des caractères spéciaux	97 275	132 498
Regroupement des lemmes doublons	96 937	132 300
Regroupement des lemmes mal orthographiés (diacritiques)	96 770	130 341

TABLE 2 – Nettoyage des formes non lemmatisées

2.3 Appariement morphologique

De nombreux travaux se sont intéressés à la question de l'appariement morphologique automatique, en particulier dans le domaine de la recherche d'information. Les méthodes consistent principalement à constituer des ensembles de formes partageant les mêmes racines, en réunissant à la fois les formes fléchies, les lexèmes simples et les lexèmes construits (dérivation ou composition). Par exemple, il s'agira de regrouper *chanter* avec *chant*, *chanteur* et leurs formes fléchies. Etant donné la prise en compte d'allomorphes et de multiples schémas de construction morphologique (préfixation, suffixation, conversion, composition), les travaux proposent l'utilisation de règles symboliques (Gaussier 1999, pour la dérivation en français), de mesures de similarité sémantique entre lexèmes (Schone & Jurafsky 2000), ou encore d'algorithmes de clusterisation non-supervisés (Singh & Gupta 2019) qui ne nécessitent pas de connaissances linguistiques préalables. En ce qui concerne la constitution de ressources dérivationnelles, les approches à base de règles induites à partir de connaissances ont montré leur intérêt pour l'allemand (Zeller *et al.* 2013).

Dans notre cas, nos objectifs sont moins ambitieux ; nous ciblons des suffixes de nominalisation spécifiques et cherchons à extraire des paires plutôt que des ensembles. En ce sens, et parce que nous cherchons à réunir un très large ensemble de paires avec le moins de bruit possible, pour l'appariement morphologique, nous optons pour une approche à base de règles.

Troncation des formes Afin d'apparier les noms et les verbes, nous avons tronqué les formes de leurs suffixes (*-age, -ment, -ion, -ure, -ance, -ade* et *-(a)ison*) ou de leur finale verbale (*-er*, *-ir* ou *-re*) dans le but de faire correspondre leurs radicaux. Nous avons également généré des radicaux susceptibles d'être sujets à allomorphie. Toutes les troncations possibles ont été gardées pour un même lemme, à condition que la longueur de la forme tronquée ne soit pas inférieure à 2 caractères.

Appariement Les formes tronquées verbales et nominales ont été appariées sur la base d'une identité immédiate. De nombreux noms en *-ion* correspondant à des verbes à la deuxième personne du pluriel de l'imparfait (*entendion* plutôt que *entendions*) ont été évincés en cherchant pour chaque forme en *-ion* si un équivalent avec un *-s* existait parmi les formes fléchies du verbe base dans le

Glàff (Sajous *et al.* 2013). Pour les formes en *-ment* qui se trouvaient être des adverbes faussement étiquetés 'nom', nous nous sommes servis de l'ensemble des formes étiquetées 'adverbe' dans frCOW. Si la fréquence de la forme étiquetée 'adverbe' était plus élevée que la fréquence de la forme étiquetée 'nom', nous avons supprimé la paire comprenant le nom en *-ment* de la base. Afin de récupérer des paires incluant des radicaux allomorphiques et supplétifs que nos méthodes ne parvenaient pas à collecter, nous avons exploité la base lexicale Démonette (Hathout & Namer, 2014). Seules les paires qui ne figuraient pas déjà dans notre base ont été conservées. L'ajout de ces données a permis d'enrichir la base de 1 380 nouvelles paires (avec leurs fréquences dans frCOW quand les verbes et noms s'y trouvent). Enfin, en raison de nombreuses paires appariées à la suite d'erreurs orthographiques, nous avons retiré toutes les paires comprenant des noms non préalablement lemmatisés lorsque des équivalents lemmatisés avec une distance de Levenshtein de 1 étaient présents dans la base. Le détail des proportions de paires obtenues selon les étapes est décrit dans la table 3.

	Nombre de paires
Identité immédiate	40 940
Nettoyage des lemmes mal orthographiés	36 519
Retrait des noms en *-ment* mal étiquetés	33 478
Ajout des paires de Démonette	34 858
Distances de Levenshtein (1)	27 857

TABLE 3 – Etapes pour l'appariement verbe-nom

3 Description de la ressource

Au total, la base est constituée de 25 857 paires verbe-nom dont les dérivés nominaux sont issus des suffixations en *-ion, -age, -ment, -ance, -ure, -ade* et *-aison*. Pour chaque paire sont données les fréquences du verbe et du nom dans frCOW, leur provenance (préalablement lemmatisé 'lemmatized' ou non 'nolemma' dans frCOW ou bien issu de Démonette) ainsi que le suffixe à l'origine de la dérivation. La table 4 présente 2 entrées : la paire comprenant le dérivé le plus fréquent, et une paire comprenant un verbe et un nom n'apparaissant chacun qu'une fois dans le corpus.

verbe	freq_verbe	origine_verbe	nom	freq_nom	suffixe	origine_nom
former	662730	lemmatized	formation	3966933	ion	lemmatized
kikouloler	1	nolemma	kikoulolage	1	age	nolemma

TABLE 4 – Exemples d'entrées de la base

Distribution Le détail des proportions de paires par suffixe est donné dans la table 5. La base réunit une majorité de dérivés en *-ion, -age* et *-ment*, ceux-ci étant à eux-seuls à l'origine de 85% des paires. Les suffixations en *-ance* et *-ure*, bien moins nombreuses, représentent un peu plus de 10% des données, tandis que seuls 4,3% des paires concernent les suffixations en *-ade* et *-aison*.

Evaluation La base a été évaluée en procédant à 4 tirages aléatoires de 100 paires (table 6, gauche). Pour chaque tirage, les paires ont été annotées "correcte" ou "incorrecte" par un seul annotateur. Deux

	Nombre de paires	Proportion
-ion	10 558	40,8%
-age	6 588	25,4%
-ment	4 865	18,8%
-ance	1 439	5,5%
-ure	1 252	4,8%
-ade	771	2,9%
-aison	384	1,4%
Total	25 857	100%

TABLE 5 – Distribution des paires selon le suffixe

évaluations différentes ont été réalisées sur chaque tirage. Pour la première (*évaluation stricte*), les paires incorrectes pouvaient répondre à des erreurs de catégorisation de la base (nom plutôt que verbe, le plus souvent), à des erreurs d'orthographe sur l'une des deux formes ou sur les deux, des erreurs de langue (italien, créoles à base française, anglais, latin, ancien ou moyen français), ou encore à une correspondance sémantique trop opaque. Pour la deuxième évaluation (*évaluation relâchée*), nous avons considéré que les fautes d'orthographe ne constituaient pas des erreurs d'appariement si les deux formes comportaient les mêmes erreurs (par exemple : *anhiler* → *anhilation*). Les mêmes évaluations ont également été réalisées pour chaque suffixe (table 6, droite, 100 paires par suffixe).

	évaluation stricte	évaluation relâchée
1	75%	88%
2	79%	90%
3	84%	89%
4	78%	81%
moyenne	79%	87%

	évaluation stricte	évaluation relâchée
-ion	75%	89%
-age	86%	96%
-ment	77%	91%
-ance	54%	72%
-ure	62%	68%
-ade	48%	62%
-aison	53%	61%

TABLE 6 – Scores d'exactitude selon le type d'évaluation

Productivité des schémas morphologiques Grâce à la quantité de paires qu'elle regroupe, la base permet en outre de calculer la productivité globale des suffixes, notamment en appliquant la mesure de Baayen (1994). Avec l'idée que les suffixes les plus productifs vont former de nombreux dérivés peu fréquents, cette mesure calcule la productivité d'un suffixe E en calculant le rapport entre le nombre d'hapax suffixés par E et le nombre total d'hapax dérivés dans le même corpus. Plutôt que de donner cette mesure de productivité, nous donnons à la figure 1 les comptages des différents hapax dans notre base, ce qui permet de dessiner les premiers contours de la répartition du lexique par ces schémas en compétition : les suffixations en *-ion, -age* et *-ment* apparaissent comme les plus productives dans nos données, tandis que les suffixations en *-ance, -ure, -ade* et *-aison* se démarquent par leur faible productivité en comparaison.

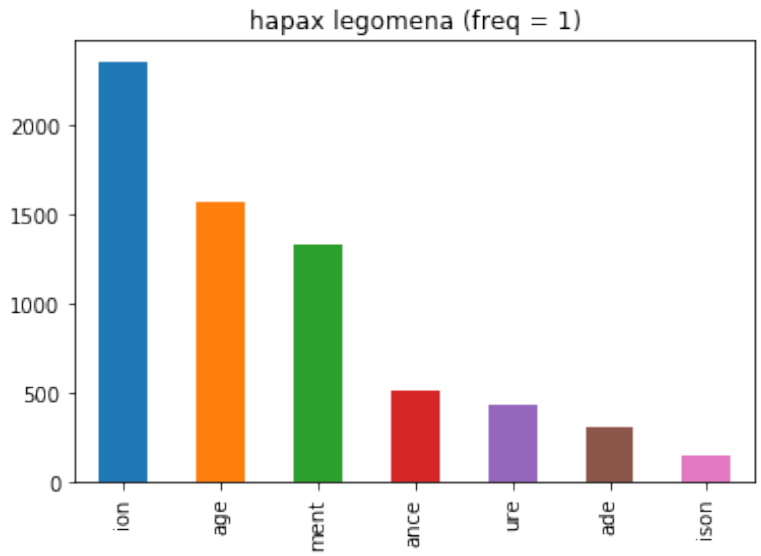

FIGURE 1 – Productivité des suffixes (Baayen (1994), hapax)

4 Conclusion

Cette ressource a été constituée dans le but d'explorer quantitativement les raisons de la coexistence de schémas morphologiques rivaux en français dans le cadre de la compétition entre les suffixes constructeurs de noms d'événement à partir de verbes. A notre connaissance, il s'agit de la base lexicale la plus large construite intégralement à partir de données issues du web pour les nominalisations issues du français écrit. Sa qualité pourrait toutefois faire l'objet d'améliorations. Les évaluations de la base ont montré que les erreurs orthographiques constituaient la majorité des erreurs uniquement pour les suffixations les plus fréquentes et les plus productives, à savoir *-ion*, *-age* et *-ment*. En ce sens, le nettoyage des lexiques pourrait être amélioré afin de rassembler au mieux les diverses variantes orthographiques d'une forme sous un même lemme. L'appariement morphologique pourrait en outre bénéficier de méthodes issues de la sémantique distributionnelle (Schone & Jurafsky 2000, Wauquier *et al.* 2018 pour les suffixations en *-age, -ment* et *-ion*). Etant donné les analyses proposées dans la littérature (Tribout 2010, Tribout & Villoing 2014), la base mériterait également d'être enrichie par l'ajout de paires verbe-nom issues de la conversion (*marcher* → *marche*, *découvrir* → *découverte*, *arriver* → *arrivée*), elle aussi en compétition dans la construction de noms d'événement en français. La conversion, non régulière sur le plan formel, pose d'autres enjeux pour l'extraction et l'appariement automatiques, qui constitueront la prochaine étape de ce travail. La ressource est disponible sur le site d'Ortolang sous le nom VerNom.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. In A. ABEILLÉ, Éd., *Treebanks : Building and Using Parsed Corpora*, p. 165–187. Dordrecht : Springer Netherlands. DOI : 10.1007/978-94-010-0201-1_10.

BAAYEN R. H. (1994). Productivity in language production. *Language and Cognitive Processes*, **9**(3), 447–469. DOI : 10.1080/01690969408402127.

BALVET A., BARQUE L., CONDETTE M.-H., HAAS P., HUYGHE R., MARÍN R. & MERLO A. (2011). Nomage : an electronic lexicon of French deverbal nouns based on a semantically annotated corpus. In *WoLeR 2011 at ESSLLI, International Workshop on Lexical Resources*, p. 8–15, Ljubljana, Slovenia. HAL : halshs-01078047.

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

DAL G., HATHOUT N., LIGNON S., NAMER F. & TANGUY L. (2018). Toile versus dictionnaires : Les nominalisations du français en-age et en-ment. In *SHS Web of Conferences*, volume 46, p. 08003 : EDP Sciences.

FERRET K., SOARE E. & VILLOING F. (2010). Rivalry between french–age and–ée : the role of grammatical aspect in nominalization. In M. ALONI, H. BASTIAANSE, T. DE JAGER & K. SCHULTZ, Éds., *Logic, language and meaning, 17th Amsterdam Colloquium, The Netherlands, December 2009, Revised Selected Papers, Lecture Notes in Computer Science (Vol. 6042)*, p. 284–295. Berlin : Springer.

FERRET K. & VILLOING F. (2012). L'aspect grammatical dans les nominalisations en français : les déverbaux en -age et -ée. *Lexique (20)*, p. 73–127.

FRADIN B. (2014). La variante et le double. In F. VILLOING, S. DAVID & L. SARAH, Éds., *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*, p. 109–147. Presses Universitaires de Paris Ouest.

FRADIN B. (2019). Competition in derivation : What can we learn from french doublets in-age and-ment ? In F. RAINER, F. GARDANI, W. U. DRESSLER & H. C. LUSCHÜTZKY, Éds., *Competition in Inflection and Word-Formation. Studies in Morphology, vol 5.*, p. 67–93. Springer, Cham.

GAUSSIER É. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, p. 24–30.

HATHOUT N. & NAMER F. (2014). Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology*, **11**(5), 125–168.

HATHOUT N., SAJOUS F. & TANGUY L. (2009). Looking for french deverbal nouns in an evolving web (a short history of wac). In *Proceedings of the fifth workshop on Web As Corpus (WAC), San Sebastian, September 7th*, p. 37–44.

HATHOUT N. & TANGUY L. (2002). Webaffix : Discovering Morphological Links on the WWW. In *LREC 2002*, Proceedings of LREC, Las Palmas, Spain. HAL : halshs-01322326.

LÜDTKE J. (1978). *Prädikative Nominalisierungen mit Suffixen im Katalanischen, Spanischen und Französischen*. Tübingen : Niemeyer.

MARTIN F. (2010). The semantics of eventive suffixes in French. In M. RATHERT & A. ALEXIADOU, Éds., *The Semantics of Nominalizations across Languages and Frameworks*, p. 109–141. Berlin : Mouton de Gruyter.

NAMER F., BARQUE L., BONAMI O., HAAS P., HATHOUT N. & TRIBOUT D. (2019). Demonette2 – A large scale derivational database for French : first results. In *TALN*, Toulouse, France. HAL : halshs-02275652.

SAJOUS F., HATHOUT N. & CALDERONE B. (2013). GLÁFF, un Gros Lexique Á tout Faire du Français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 285–298, Les Sables d'Olonne, France.

SCHONE P. & JURAFSKY D. (2000). Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, p. 67–72 : Association for Computational Linguistics.

SCHÄFER R. (2015). Processing and querying large web corpora with the COW14 architecture. In P. BANSKI, H. BIBER, E. BREITENEDER, M. KUPIETZ, H. LÄNGEN & A. WITT, Éds., *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster : UCREL IDS.

SCHÄFER R. & BILDHAUER F. (2012). Building large corpora from the web using a new efficient tool chain. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, M. U. DOÄŸAN, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, p. 486–493, Istanbul, Turkey : European Language Resources Association (ELRA).

SINGH J. & GUPTA V. (2019). A novel unsupervised corpus-based stemming technique using lexicon and corpus statistics. *Knowledge-Based Systems*, **180**, 147–162.

TRIBOUT D. (2010). *Noun to verb and verb to noun conversions in French*. Thèse de doctorat, Université Paris Diderot (Paris 7). HAL : tel-01577528.

TRIBOUT D. & VILLOING F. (2014). La composition VN et la conversion V>N en français : un nouveau cas de concurrence morphologique ? In F. VILLOING, S. DAVID & S. LEROY, Éds., *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux*. Presses Universitaires de Paris Ouest. HAL : halshs-01690227.

UTH M. (2010). The rivalry of the French nominalization suffixes -age and -ment from a diachronic perspective. In M. RATHERT & A. ALEXIADOU, Éds., *The Semantics of Nominalizations across Languages and Frameworks*, p. 215–244. Berlin : Mouton de Gruyter.

UTH M. (2016). The competition of event nominalization procedures of French, in comparison with German. *Zeitschrift für Romanische Philologie*, **132**(1), 58–89.

WAUQUIER M., FABRE C. & HATHOUT N. (2018). Différenciation sémantique de dérivés morphologiques à l'aide de critères distributionnels. In *Congrès Mondial de Linguistique Française (CMLF)*, volume 46 de *6e Congrès Mondial de Linguistique Française*, Mons, Belgium : EDP Sciences. DOI : 10.1051/shsconf/20184608006, HAL : hal-01876027.

WAUQUIER M., HATHOUT N. & FABRE C. (2019). Contributions of distributional semantics to the semantic study of french morphologically derived agent nouns. In *Mediterranean Morphology Meetings*, volume 12, p. 111–121.

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 (Volume 1 : Long Papers)*, p. 1201–1211.