Demonette2 - Une base de données dérivationnelle du français à grande échelle : premiers résultats

Fiammetta Namer, Lucie Barque, Olivier Bonami, Pauline Haas, Nabil Hathout, Delphine Tribout

Demonette2 - Une base de données dérivationnelles du français à grande échelle : premiers résultats

Fiammetta Namer[1] Lucie Barque[2] Olivier Bonami[2] Pauline Haas[3] Nabil Hathout[4]
Delphine Tribout[5]
(1) UMR 7118 ATILF, Nancy, (2) UMR 7110 LLF, Paris
(3) UMR 8094 Lattice, Paris, (4) UMR 5263 CLLE-ERSS, Toulouse
(5) UMR 8163 STL, Lille
fiammetta.namer@univ-lorraine.fr, {lucie.barque;pauline.haas}@univ-paris13.fr, olivier.bonami@linguist.univ-paris-diderot.fr, Nabil.Hathout@univ-tlse2.fr, delphine.tribout@univ-lille3.fr

RÉSUMÉ

Cet article présente la conception et le développement de Demonette2, une base de données dérivationnelle à grande échelle du français, développée dans le cadre du projet ANR Démonext (ANR-17-CE23-0005). L'article décrit les objectifs du projet, la structure de la base et expose les premiers résultats du projet, en mettant l'accent sur un enjeu crucial : la question du codage sémantique des entrées et des relations.

ABSTRACT

Demonette2 – A large scale derivational database for French: first results

This paper presents the design and development of Demonette2, a large-scale derivational database of French, developed as part of the ANR Démonext project (ANR-17-CE23-0005). It describes the objectives of the project, the structure of the database and presents the first results of the project, focusing on the question of the semantic encoding of lexical units and their relationships.

MOTS-CLES : base de données dérivationnelles, codage sémantique, paradigmes dérivationnels, lexique français.

KEYWORDS: derivational database, semantic encoding, derivational paradigms, French lexicon.

1 Introduction

Démonette2 est une base de données morphologiques (BDM) qui décrit les propriétés dérivationnelles des mots du français de manière extensive et systématique. Alimentée par des ressources lexicales existantes de nature variées, cette base constitue une combinaison inédite d'informations répondant à des besoins multiples, comme la confirmation empirique et l'élaboration d'hypothèses en morphologie, le développement d'outils en traitement automatique des langues, l'enseignement du vocabulaire et le traitement des troubles du langage développementaux ou acquis. Développée dans le cadre du projet Démonext[1] (2018-2021), cette base décrira à terme un réseau dérivationnel totalisant au moins 366 000 entrées, comportant des relations morphologiques directes, indirectes, ascendantes et descendantes et une représentation du sens construit ; les lexèmes seront munis d'annotations morphologiques, de caractérisations sémantiques, de représentations phonologiques, de fréquences d'emploi dans différents corpus,

[1] Démonext bénéficie du soutien de l'ANR 17-CE23-0005, et réunit 4 UMR : ATILF, STL, CLLE-ERSS et LLF.

d'indications de l'âge d'acquisition, etc. Le résultat sera accessible grâce une plateforme offrant un accès adapté à différents publics. La BDM sera distribuée sous licence libre via l'EQUIPEX Ortolang (https://www.ortolang.fr/) et la plateforme REDAC (http://redac.univ-tlse2.fr/).

2 Etat de l'art

L'analyse morphologique constitue l'une des étapes initiales centrales des systèmes de TAL. Les analyseurs, le plus souvent basés sur des méthodes d'apprentissage automatique, réalisent un découpage morphématique des mots et permettent de compenser les limitations des lexiques. On citera Linguistica (Goldsmith, 2001), Morfessor (Creutz, 2003 ; Creutz, Lagus, 2005), l'analyseur de Bernhard (2009) ou plus récemment les modèles de Cotterell *et al*. (2015, 2017). Applicables à n'importe quelle langue, ces systèmes sont plus efficaces pour les langues à morphologie concaténative comme l'anglais, l'allemand et le français. Parallèlement, des analyseurs symboliques ont été développés par des linguistes ; pour un panorama, voir Bernhard *et al*. (2011), ou Namer (2013). Les analyseurs morphologiques peuvent être complétés par des ressources lexicales munies d'annotations dérivationnelles, ayant une couverture lexicale assez importante et un ensemble suffisamment riche et varié de propriétés codées pour être exploitables dans une chaîne de traitement en TAL. Malgré un besoin important pour des ressources de ce type, très peu ont été développées au cours des vingt dernières années – et pratiquement aucune pour les langues romanes, comme le constataient déjà Dal *et al*. (1999). La plus connues de ces ressources est CELEX (Baayen *et al*., 1995), qui décrit pour l'allemand, l'anglais et le néerlandais, les propriétés phonétiques, flexionnelles, morpho-syntaxiques, dérivationnelles et statistiques d'un peu moins de 250 000 mots non fléchis issus de dictionnaires et de corpus littéraires et journalistiques. Sinon, citons CatVar (Habash, Dorr, 2003) qui est un système lexical de 100 000 lexèmes de l'anglais réunis en sous-familles dérivationnelles organisées en graphes ; sur le même principe, DerivBase (Zeller *et al*., 2013) contient 215 000 unités lexicales de l'allemand dont les regroupements en familles dérivationnelles sont motivés sémantiquement ; la version 3.0 de WordNet (Fellbaum *et al*., 2007) est enrichie de relations dérivationnelles annotées sémantiquement entre les verbes et une partie de leurs dérivés nominaux (par exemple, la relation EMPLOY_V/EMPLOYER_N est étiquetée agent). Pour le français, on peut citer deux initiatives récentes visant le développement de réseaux lexicaux à large couverture. Le RL-fr (Lux-Pogadalla, Polguère, 2014) décrit 1 million d'entrées au moyen de relations sémantiques inspirées de (Mel'čuk, 1996). La base JeuxDeMots (Lafourcade, Joubert, 20008), antérieure à RL-fr, est fondée aussi sur le même principe, mais comporte également des relations dérivationnelles. JeuxDeMots obéit à une conception participative de l'enrichissement du réseau, sous forme de jeu en ligne. La fiabilité des termes proposés par les joueurs s'accroît en fonction du nombre de réponses identiques. En 10 ans, la ressource a atteint 270 millions de relations instanciant 151 fonctions lexicales différentes et connectant quelque 3,5 millions de termes. Comme nous le montrons dans la suite, JeuxDeMots et Démonette2 sont en quelque sorte complémentaires puisqu'elles se distinguent en termes de couverture et de mode de développement, dans la mesure où l'évolution de la première dépend de l'imagination des participants, là où la seconde s'appuie sur l'expertise des auteurs des sources d'alimentation de la base pour garantir cohérence et validité théorique à la description morphologique de chaque relation.

La carence de ressources purement dérivationnelles pour le français a motivé le développement, à partir de 2014, du prototype Démonette1 (Hathout, Namer, 2014a, 2014b, 2015, 2016 ; Namer *et al*., 2017). Démonette1 décrit une partie des familles dérivationnelles des verbes, accompagnés de leurs noms d'agent, d'activité et adjectifs de propriété modalisée. Trois objectifs étaient visés : (1)

produire une ressource dont les entrées sont des relations dérivationnelles munies d'annotations riches, notamment sémantiques (Namer, 2002) ; (2) compléter les dérivations base → dérivé par toutes relations motivées qui existent entre membres des familles dérivationnelles, suivant le modèle analogique implémenté dans Morphonette (Hathout, 2009) ; (3) définir une architecture extensible et redondante, qui peut être alimentée par des ressources morphologiques hétérogènes.

Forte de l'expérience acquise avec Démonette1, la BDM Démonette2 construite dans le cadre du projet ANR Démonext se veut à terme une ressource de grande ampleur disposant de descriptions riches des lexèmes, des relations dérivationnelles entre lexèmes et des paradigmes où celles-ci s'insèrent. Démonette2 est par ailleurs compatible avec les principales théories morphologiques actuelles (qu'elles soient morphématiques, lexématiques ou paradigmatiques, cf. respectivement Halle, Marantz, 1993 ; Fradin, 2003 ; Bauer, 1997). Les principes qui la sous-tendent lui confèrent une organisation originale : une entrée de Démonette2 correspond en effet à une relation morphologique dérivationnelle entre deux lexèmes. La BDM est ainsi conforme aux hypothèses théoriques qui considèrent que le lexème est l'unité morphologique fondamentale et que la construction dérivationnelle remplit deux fonctions : (1) créer de nouveaux lexèmes et (2) établir des relations de motivation sémantique et formelle entre les lexèmes présents dans le lexique. Par exemple, l'entrée connectant NATION$_N$ à INTERNATIONAL$_A$ rend compte du fait que l'on peut motiver le second relativement au premier : des *relations internationales* sont des *relations entre plusieurs nations*. Chaque entrée de Démonette2 comporte une description morpho-phonologique et une description morpho-sémantique (1) des lexèmes connectés ainsi que (2) de la relation qui les caractérise. Ces deux descriptions sont indépendantes. Les relations décrites ne sont pas limitées aux motivations classiques base → dérivé ; elles incluent aussi les relations sémantiquement motivées entre membres d'une famille dérivationnelle. Cette configuration permet le regroupement des familles en réseaux formels, sémantiques et dérivationnels, offrant à terme une démonstration à grande échelle de l'organisation paradigmatique du lexique construit.

3 Structure et sources d'approvisionnement de la BDM

Architecture : La constitution de Démonette2 est doublement hybride : (1) la base est élaborée à partir d'une compilation de ressources existantes et d'annotations nouvelles, manuelles, semi-automatiques et automatiques ; (2) elle combine des ressources de deux natures dont certaines documentent des *unités lexicales* et d'autres des *relations dérivationnelles* entre unités. De fait, Démonette2 se compose essentiellement de deux tables : une pour les lexèmes, et l'autre pour les relations entre lexèmes. Cette architecture limite en partie la redondance dans la base et améliore sa maintenabilité ; elle permet par ailleurs de planifier des campagnes d'annotation parallèles avec une souplesse relative. Ainsi, le travail coûteux d'annotation sémantique des lexèmes peut être mené de front avec un enrichissement du graphe des relations dérivationnelles ou avec une annotation des propriétés formelles de ces relations. La taille de la base contenant la *table des lexèmes* est maximale, de manière à garantir que l'ensemble des relations dérivationnelles décrites dans l'autre base fassent nécessairement références à des lexèmes qui y sont attestés. Pour assurer une couverture optimale, la table des lexèmes comporte l'ensemble des unités lexicales dont les 1 406 857 formes fléchies constituent *Glaff*[2] (obtenu à partir des données de *Wiktionnaire*). Outre la partie du discours, chaque lexème sera à terme décrit par l'ensemble des formes de son paradigme flexionnel transcrits au format SAMPA (cf. §4.1). Sa fréquence dans différentes ressources (Frantext[3], frWaC…) et sa classe sémantique complètent cette description. La base

[2] Glàff : "Gros lexique à tout faire du français", http://redac.univ-tlse2.fr/lexiques/glaff/telechargement.html
[3] BD textuelles littéraires et journalistique hébergée à l'UMR 7118 l'ATILF : https://www.frantext.fr/

contenant la *table des relations* est, elle, alimentée initialement par le contenu de 6 ressources (cf. *infra*). Les informations transposées dans cette table concernent la structure morphologique des deux lexèmes impliqués dans la relation (est-il suffixé, préfixé, construit par conversion, et, le cas échéant, au moyen de quel affixe), et la nature de la relation (directe entre un dérivé et sa base, e.g. BASKETTEUR/BASKET, et indirecte quand elle connecte deux lexèmes de la même famille dérivationnelle comme DECORATEUR/DECORATION, FASCISME/FASCISTE ou AVIATEUR/AVIATION). Comme nous le montrons au §4, les descriptions pertinentes sont validées, et converties (semi-)automatiquement ou manuellement dans le format de la table. Elles sont en outre assorties de nouvelles informations, essentiellement sémantiques. La couverture de la base est enfin enrichie des relations qui complètent les familles dérivationnelles auxquelles appartiennent les relations déjà présentes. La gestion du projet repose sur un circuit de maintenance basé sur *git* qui différencie le rythme de modification des deux bases : la base des relations étant destinée à évoluer rapidement, sa modification est confiée à des éditeurs (humains ou automatiques) dont le travail est destiné à être injecté dès qu'il est terminé dans une version de développement, après contrôles automatiques de sa cohérence et validation par un responsable de tâche, suivant une grille de critères préalablement établis. À l'inverse, les changements dans la table des lexèmes suivent un rythme plus lent, et ne sont injectés dans la base qu'à l'occasion de la publication d'une nouvelle version de référence, sous le contrôle direct de l'administrateur. Il est à noter que la gestion de la base donne lieu à l'implémentation d'un ensemble d'outils (interface de soumission, interfaces de visualisation et d'édition à destination) qui seront valorisés et distribués comme une plateforme de maintenance de ressource morphologique.

Sources : Outre l'apport initial des 96 000 entrées de Demonette1.3, le contenu de Démonette2 sera dans un premier temps obtenu par migration de ressources dérivationnelles existantes, développées et validées par des morphologues (cf. Tab. 1). Ces ressources sont choisies pour leur disponibilité, leur complémentarité, la richesse des descriptions (annotations morphologiques et, pour la plupart d'entre elles, traits sémantiques et phonologiques). Leur traitement est échelonné en fonction du degré d'immédiateté de leur adaptation dans le format de Demonette2.

Nom (auteur)	Convers (Tribout 2010)	Denom (Strnadová 2014)	Dimoc (Roché 2004, 2008, 2011a,b; Lignon, Roché 2011; Roché, Plénat 2012)	Mordan (Koehl, 2012)	Lexeur (Fabre *et al*, 2004)
Taille	3 500	15 500	60 900	3 900	3 200
Contenu	Convers N/V	Adjectifs dénom.	Noms construits sur N, V et A	Noms désadj.	Noms d'agent en *-eur* et base N / V
Ex.	CRI /CRIER	DUCAL /DUC	ALBIGEOIS /ALBI	BEAUTE/BEAU	BASKETTEUR/BASKET

TABLE 1 : premières sources d'approvisionnement de la base Demonette2

Quand toutes les ressources auront été adaptées et transférées dans la base, celle-ci décrira 183 000 relations dérivationnelles réalisant environ 120 procédés de dérivation, par suffixation (*-ard, -ariat, -at, -âtre, -el, -aie, -iser, -erie, -esque, -esse, -eur, -eux, -iste* …), conversion et préfixation (*a-, anti-, bi-, co-, contre-, dé-, é-, extra-, hyper-, hypo-, in-, infra-, inter-* ...). Le contenu de Démonette2 ne se résume pas aux seules relations héritées de ces ressources. L'ambition du projet est de constituer (semi-)automatiquement les familles dérivationnelles des lexèmes codés au cours de l'étape de migration et décrire, comme autant de nouvelles entrées, les relations entre les membres de ces familles. Plusieurs approches sont envisagées : l'extension des régularités

[4] Échantillon de la toile, pour le domaine français, totalisant 1,6 milliard d'occurrences, cf. (Baroni *et al.*, 2009)

paradigmatiques encodées lors de la première phase, l'usage de réseaux de neurones (Cotterell *et al.*, 2017) ou l'application de l'analyse des concepts formels (Leeuwenberg *et al.*, 2015).

4 Premiers résultats

4.1 Un échantillon de Démonette2

Chaque entrée de Démonette2 décrit une relation dérivationnelle qui s'établit entre deux lexèmes (**Mot1, Mot2**) de la même famille dérivationnelle où Mot1 est considéré comme morphologiquement *motivé* par Mot2. En général, la motivation de Mot2 par Mot1 est également possible, ce qui fait que l'entrée symétrique (Mot2, Mot1) est aussi présente dans la base. Nous adoptons une conception élargie de la notion de famille dérivationnelle, où divers degrés de variation radicale sont possibles entre Mot1 et Mot2. La relation formelle peut être totalement transparente, comme avec (DANSE, DANSEUR) où le radical commun /dãs/ est immédiatement identifiable ; elle peut impliquer une allomorphie régulière qui n'empêche pas la reconnaissance du radical, comme avec la variation /ø/-/oz/ observée dans (NERVEUX, NERVOSITE) ; enfin, l'un des membres peut être construit sur le radical savant de l'autre, ce qui opacifie la relation formelle entre les deux, comme avec (SAINT-ETIENNE, STEPHANOIS). Chaque entrée est identifiée par la graphie et la catégorie grammaticale de Mot1 et Mot2 (cols 1 et 2 du Tab. 2), et inclut la structure morphologique des deux lexèmes (cols 3 à 6), les propriétés (cols 7 et 8) de leur relation dérivationnelle, et les caractéristiques des éventuelles transformations morphophonologiques qui y sont liées (cols 9 et 10). Le codage de ces alternances est (semi-)automatique : à chaque Moti est associée la transcription phonologique présente dans Glaff de chacun des membres de son paradigme flexionnel ; les transcriptions sont complétées et uniformisées au moyen de règles apprises à partir des codages phonétiques contenus dans la base "témoin" Flexique[5], dont le contenu est très fiable, mais la couverture réduite ; les variations (cols 9 et 10) résultent de la comparaison des transcriptions. Ainsi, les paires (Mot1, Mot2) sont regroupables selon leurs similarités morpho-phonologiques, et leurs familles dérivationnelles superposables en paradigmes formels : eg. (REALISER, REALISATEUR, REALISATION) et (ADMIRER, ADMIRATEUR, ADMIRATION) sont dans le même paradigme formel (X, Xatœr, Xasjõ). Une autre dimension paradigmatique, fondamentale pour expliquer l'organisation du lexique, est induite par les propriétés sémantiques des familles dérivationnelles. C'est pourquoi chaque entrée de Démonette est aussi caractérisée par un autre ensemble de traits, qui décrit son comportement sémantique.

Mot1	**Mot2**	**Const.1**	**exp1**	**Const.2**	**exp2**	**Orientation**	**Complexité**	**Série morpho-phonologique**	**Alternance**
BOIRE$_V$	BUVEUSE$_{Nf}$	--	--	suf	euse	ascend	simple	X/Xøz	=
ACTEUR$_{Nm}$	ACTION$_{Nm}$	suf	eur	suf	ion	indirect	simple	Xtœr/Xsjõ	t/s
BOIRE$_V$	IMBUVABLE$_A$	--	--	pre	in	ascend	complexe	X/ɛ̃Xabl	
FOIE$_N$	HEPATIQUE$_A$	--	--	suf	ique	ascend	simple	fwa/epatik	NONE

TABLE 2 : Entrées simplifiées de Demonette2 (extrait)

[5] http://www.llf.cnrs.fr/flexique-fr.php, (Bonami *et al.*, 2014)

4.2 Codages sémantiques

Les décisions à prendre en termes de codage des propriétés (morpho-)sémantiques sont cruciales, car elles conditionnent la structure et l'homogénéité du contenu de la base. Pour chaque entrée, l'information sémantique à définir est le produit de trois annotations qui se complètent, et que l'on ne trouve à notre connaissance dans aucune autre BDM : la classe ontologique de Mot1 et Mot2, la catégorie sémantique de la relation morphologique entre Mot1 et Mot2, et la motivation réciproque de Mot1 et Mot2, sous forme d'une paraphrase inspirées du modèle des *frame definitions* de Framenet (Fillmore *et al.*, 1998).

Codage des unités du lexique : Actuellement, par défaut, les verbes s'interprètent comme des *situations* et les adjectifs comme des *propriétés*. Pour le codage des noms, un jeu d'étiquettes a été adapté des *WordNet Unique Beginners,* désormais *UB* (Miller *et al.* 1990, Fellbaum 1998), ce qui permet de couvrir l'ensemble du spectre lexical et d'offrir un degré de généralité qui convient *a priori* à la description morphologique. La liste des *UB* (en gras), complétée par l'étiquette sous-spécifiée *Top* s'organise comme indiqué Fig.1. Nous avons établi des tableaux de correspondance entre les *UB* (ou leurs super-types, soulignés dans la Fig.1) et les types sémantiques proposés dans les ressources morphologiques intégrées dans Démonette2. L'*UB* Person correspond ainsi aux types "AGF" (agent féminin) et "AGM" (agent masculin) de Demonette1, aux types "Ah" (humain) et "Ahg" (gentilé) de Dimoc, etc. Sur la base de cet appariement, la plupart des noms issus des cinq ressources ont pu être munis d'une ou de plusieurs étiquettes normalisées.

Classe sous-spécifiée : *Top*

1 Situation [Situation stative [**Feeling, State, Attribute**]
Situation dynamique [**Act, Event**]]

2 Entité [**Objet** [Non Animé [**Objet Naturel/Artefact**]]
[Animé [**Animal**, **Person**]]

Classes relationnelles : **Group, Part**

FIGURE 1 : Hiérarchie partielle des étiquettes ontologiques d'après *WordNet Unique Beginners*

Codage des relations morpho-sémantiques : La description des relations sémantiques associées aux règles morphologiques s'inspire du modèle des fonctions lexicales (Mel'čuk, 1996). Pour l'heure, seules les relations directes entre un dérivé et sa base ont été décrites, et dépendent du type ontologique attribué à Mot1 et Mot2. La caractérisation de la relation prend en compte la catégorie grammaticale de la base et de son dérivé et leur classe sémantique générale (*Situation* ou *Entité*) et se compose d'un type général (synonymie, résultatif, causatif, etc.), d'un schéma sémantique abstrait prenant en compte l'orientation de la relation entre Mot1 et Mot2, et du procédé dérivationnel impliqué. On distingue pour l'instant 20 types de relations sémantiques, cf. Tab 3.

Qualif. Gén. de la rel. sém.	**Type sém. de Mot1**	**Type sém. de Mot2**	**Orien-tation**	**Type sém. de la rel.**	**Schéma sém. abstrait**	**Exemple (Mot1,Mot2)**
Entité-Entité	group x pers	person	descen	collectif	ensemble de N	(EQUIPIER, EQUIPE)
Situation-Entité	person	sit.dyn	descen	agent	ce(lui) qui V	(CHANTEUR, CHANTER)
	sit.stat	person	ascen	experiencer	ressentir ce que ressent N	(ADMIRER,ADMIRATEUR)

TABLE 3 : Echantillon de relations morpho-sémantiques et de schémas sémantiques abstraits

Comme avec les séries morpho-phonologiques, les schémas sémantiques abstraits combinés aux types ontologiques participent à la structuration du lexique en permettant de constituer les entrées, et, de là, les familles dérivationnelles, en paradigmes sémantiques : (ENSEIGNER, ENSEIGNANT, ENSEIGNEMENT) et (CHANTER, CHANTEUR, CHANT) se retrouvent dans le même paradigme sémantique centré sur une *situation dynamique*, contrairement à (ADMIRER, ADMIRATEUR, ADMIRATION), où le verbe est statif.

Paraphrases glosant les relations dérivationnelles : Le troisième niveau d'annotation sémantique d'une entrée (Mot1, Mot2) est une paraphrase faisant intervenir Mot1 et Mot2 qui exprime la motivation réciproque de chaque lexème relativement au sens de l'autre, cf. Tab. 4. Cet énoncé définitoire est généralisé sous la forme d'une glose où Mot_i est remplacé par son type sémantique.

(Mot1,Mot2)	**Paraphrase concrète**	**Paraphrase abstraite**
enseigner, enseignant	Un $enseignant_1$ $enseigne_2$	Un N*person* V*sit.dyn*
hôpital, hospitaliser	On $hospitalise_2$ qqc dans un $hôpital_1$	On *Vsit.dyn* dans *Nartefact*

TABLE 4 : Paraphrase de motivation sémantique de Mot1 et Mot2

A ce jour, une table des correspondances assure l'appariement des types sémantiques codés dans les ressources de la Tab. 1 avec les *Unique Beginners* de WordNet. 30% de la base Denom et 50% de Convers sont convertis au format Demonette2 suivant les principes présentés, et complétés de nouvelles relations. Enfin, la plateforme de dépôt est opérationnelle.

5 Conclusion

Au-delà des acteurs du TAL, destinataires naturels de la BDM, le contenu final du projet Démonext permettra aussi de répondre aux demandes émanant des chercheurs, des universitaires, des enseignants du primaire et des orthophonistes. En effet, la *recherche* récente en morphologie adopte des méthodes de modélisation quantitative. Celles-ci ont connu de grands succès, en particulier dans le domaine de la flexion, mais elles butent dans le domaine de la dérivation sur l'absence de ressources à large couverture contenant des informations morphologiques riches. Dans *l'enseignement supérieur* en morphologie, la base Demonette2 sera utilisée pour extraire des données morphologiques et les inclure dans des questions à réponses intégrées dans le cadre de MOOC. Enfin, la BDM permettra aux *enseignants* du *primaire* et aux cliniciens *orthophonistes* (qui s'intéressent depuis une 30aine d'années à la conscience morphologique, cf. Casalis *et al*. 2003) de construire et utiliser des outils d'évaluation, d'entraînement ou de remédiation ciblés sur la morphologie, et d'approfondir les connaissances sur l'acquisition de la morphologie dérivationnelle et les troubles développementaux qui y sont liés et pour le moment peu connus comparativement aux troubles en morphologie flexionnelle (Maillart 2003). L'orthophonie pourra également mettre à profit Démonette2 dans le traitement des adultes aphasiques présentant des troubles acquis du langage (Semenza et al. 1990, Pillon et al. 1991).

Remerciements

Merci aux relecteurs anonymes pour leurs remarques qui ont permis d'améliorer sensiblement ce texte, ainsi qu'à nos ingénieurs partenaires du projet : Alexander Delaporte, Achille Falaise, Loïc Liégeois, et Alexandre Roulois, qui ont conçu et développé la plateforme de dépôt de Démonette2.

Références

Baayen R.H., Piepenbrock R., Gulikers L. (1995). *The CELEX Lexical Database (CD-ROM)*. Philadelphia, PA : Linguistic Data Consortium, University of Pennsylvania.

Baroni M., Bernardini S., Ferraresi A., Zanchetta E. (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation* 43(3), 209-226.

Bauer L. (1997). Derivational Paradigms. *Yearbook of Morphology 1996*. Booij G., van Marle J. eds. Dordrecht : Kluwer, 243-256.

Bernhard D. (2009). Unsupervised Morphological Segmentation Based on Segment Predictability and Word Segments Alignment. Actes de *Morphochallenge 2006*, 19-23.

Bernhard D., Cartoni B., Tribout D. (2011). A Task-Based Evaluation of French Morphological Resources and Tools. *Linguistic Issues in Language Technology* 5(2), 1-41.

Bonami O., Caron G., Plancq C. (2014). Construction d'un lexique flexionnel phonétisé libre du français. Actes du *4ème Congrès Mondial de Linguistique Française*, 2583-2596.

Casalis S., Mathiot E., Bécavin A.-S., Colé P. (2003). Conscience mophologique chez les lecteurs tout venant et en difficultés. *Silexicales* 3, 57-66.

Cotterell R., Schütze H. (2015). Morphological word-embeddings. Actes de *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1287-1292.

Cotterell R., Vylomova E., Khayrallah H., Kirov C., Yarowsky D. (2017). 'Paradigm Completion for Derivational Morphology'. Actes de *The 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Danemark, Association for Computational Linguistics*, 1-7.

Creutz M. (2003). 'Unsupervised Segmentation of Words Using Prior Distributions of Morph Length and Frequency'. Actes de *The 41th annual meeting of the ACL*, 280-287.

Creutz M., Lagus K. (2005). 'Inducing the Morphological Lexicon of a Natural Language from Unannotated Text'. Actes de *AKRR'05*, 106-113.

Dal G., Hathout N., Namer F. (1999). Construire un lexique dérivationnel: théorie et réalisations. Actes de *TALN-1999*, 115-124.

Fabre C., Floricic F., Hathout N. (2004). Collecte outillée pour l'analyse des emplois discordants des déverbaux en -eur. Communication présentée à *Journées d'étude sur la place des méthodes quantitatives dans le travail du linguiste*.

Fellbaum C. (ed.) (1998), *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
Goldsmith J. (2001). Unsupervised Learning of Morphology of a Natural Language. *Computational Linguistics* 27(2), 153-198.

Fellbaum C., Oherson A., Clark P.E. (2007). Putting semantics into Wordnet's "Morphosemantic" links. *Human Language Technology Challenges in the Information Society, 3d Language and Technology Conference*. Vetulani Z., Uszkoreit H. eds. Berlin : Springer Verlag, 350-358.

Fillmore C., Baker C., Lowe J. (1998). 'The Berkeley FrameNet Project'. Actes de *COLING-ACL*, 86-90.
Fradin B. (2003). *Nouvelles approches en morphologie*. Paris: Presses Universitaires de France.

Habash N., Dorr B. (2003). A Categorial Variation Database for English. Actes de *The North American Association for Computational Linguistics*, 96-102.

Halle M., Marantz A. (1993). Distributed Morphology and the Pieces of Inflection. *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*. Hale K., Keyser S.J. eds. Cambridge, MA : MIT Press: 111-176.

Hathout N. (2009). Acquisition of morphological families and derivational series from a machine readable dictionary. *Selected Proceedings of the 6th Décembrettes: Morphology in Bordeaux*. Montermini F., Boyé G., Tseng J. eds. Cambridge, MA : Cascadilla Proceedings Project, 166-180.

Hathout N., Namer, F (2014a). La base lexicale Démonette : entre sémantique constructionnelle et morphologie dérivationnelle. Actes de *TALN*, 208-220.

Hathout N., Namer F. (2014b). Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology* 11(5): 125-168.

Hathout N., Namer F. (2015). *La base lexicale morphologique du français Démonette1.2. Nancy - Toulouse*, https://www.ortolang.fr/#/market/lexicons/demonette et http://redac.univ-tlse2.fr/lexiques/demonette.html.

Hathout N., Namer F. (2016). Giving Lexical Resources a Second Life: Démonette, a Multi-sourced Morpho-semantic Network for French. Actes de *The Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 1084-1091.

Koehl A. (2012). *La construction morphologique des noms désadjectivaux suffixés en français*. Thèse de doctorat, Université de Lorraine.

Lafourcade M., Joubert A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. Actes de *JADT'08 : Journées internationales d'Analyse statistiques des Données Textuelles*, 657-666.

Leeuwenberg A., Buzmakov A., Toussaint Y., Napoli A. (2015). Exploring Pattern Structures of Syntactic Trees for Relation Extraction. Actes de *ICFCA 2015*, 153-168.

Lignon S., Roché M. (2011).Entre histoire et morphophonologie, quelle distribution pour -éen vs -ien ?. *Des Unités Morphologiques au Lexique*. Roché M. ed. Paris : Hermès, 191-250.

Lux-Pogadalla V., Polguère A. (2011). Construction of a French Lexical Network: Methodological Issues. Actes de *First International Workshop on Lexical Resources, WoLeR 2011*, 54-61.

Maillart C. (2003). Les troubles pragmatiques chez les enfants présentant des difficultés langagières. Présentation d'une grille d'évaluation : la Children's Communication Checklist. *Cahiers de la SLBU* 13, 13-32.

Mel'čuk I. (1996). Lexical Functions: A tool for the description of lexical relations in the lexicon. *Lexical Functions in Lexicography and Natural Language Processing*, Wanner L. ed. Amsterdam: John Benjamins, 37–102.

Miller G., Beckwith R., Fellbaum C., Gross D., Miller K. (1990). WordNet: An online lexical database. *International Journal of Lexicography*, 3(4), 235-244.

Namer F. (2002). Acquisition automatique de sens à partir d'opérations morphologiques en français : études de cas. Actes de *TALN-2002*, 235-244.

Namer F. (2013). A Rule-Based Morphosemantic Analyzer for French for a Fine-Grained Semantic Annotation of Texts. Actes de *SFCM 2013*, 93-115.

Namer F., Hathout N., Lignon S. (2017). Adding morpho-phonological features into a French morpho-semantic resource: the Demonette derivational database. Actes de *The First International Workshop on Resources and Tools for Derivational Morphology (DeriMo)*, 49-61.

Pillon A., De Partz M.-P., Raison A.-M., Seron X. (1991). L'orange c'est le fruitier de l'orangine : a case of morphological impairment? *Language and Cognitives Processes* 6(2), 137-167.

Roché M. (2004). Mot construit ? Mot non construit ? Quelques réflexions à partir des derivés en -ier(e). *Verbum* 26(4), 459-480.

Roché M. (2008). Structuration du lexique et principe d'économie: le cas des ethniques. Actes du *1er Congrès Mondial de Linguistique Française*, 1559-1573.

Roché M. (2011a). Quel traitement unifié pour les dérivations en -isme et en -iste. *Des Unités Morphologiques au Lexique*. Roché M. ed. Paris : Hermès, 69-143.

Roché, M (2011b).Pression lexicale et contraintes phonologiques dans la dérivation en -aie du français *Linguistica* 51, 5-22.

Roché M., Plénat M. (2012). Tous les déverbaux en -at sont-ils des conversions du thème 13 ?. Actes du *3ème Congrès Mondial de Linguistique Française*, 1387-1405.

Semenza C., Butterworth B., Panzeri M., Ferreti T. (1990). Word formation : new evidence form aphasia. *Neuropsychologia* 28(5), 499-502.

Strnadová J. (2014). *Les réseaux adjectivaux. Sur la grammaire des adjectifs dénominaux en français*. Thèse de doctorat, Université Paris Diderot / Univerzita Karlova.

Tribout D. (2010). *Les conversions de nom à verbe et de verbe à nom en français*. Thèse de doctorat, Université Paris 7.

Zeller B., Šnajder J., Padó S. (2013). DERIVBASE: Inducing and Evaluating a Derivational Morphology Resource for German. Actes de *The 51st Annual Meeting of the Association for Computational Linguistics*, 1201-1211.