Démonette : une base de données dérivationnelle du français

Démonette: a French Derivational Database

Nabil Hathout
CLLE – Université de Toulouse
nabil.hathout@univ-tlse2.fr

Fiammetta Namer
ATILF – Université de Lorraine
fiammetta.namer@univ-lorraine.fr

Ce numéro thématique est consacré à la base de données morphologique dérivationnelle du français Démonette et aux travaux qui ont été réalisés autour de cette base dans le cadre du projet Demonext (ANR-17-CE23-0005) qui a réuni une trentaine de participants (linguistes, spécialistes de traitement automatique des langues, informaticiens, psycholinguistes et orthophonistes) qui travaillent dans cinq unités de recherche : ATILF (Nancy), CLLE (Toulouse et Bordeaux), LLF (Paris), LORIA (Nancy) et STL (Lille). La version 2 de la base Démonette créée dans le cadre du projet est une ressource qui peut répondre à des besoins multiples, comme la confirmation empirique d'hypothèses en morphologie, l'élaboration de nouvelles hypothèses, le développement d'outils pour le traitement automatique des langues, la psycholinguistique, l'enseignement du vocabulaire, le traitement des troubles du langage développementaux ou acquis. Cette seconde version fait l'objet d'une présentation détaillée dans ce numéro.

Les ressources linguistiques sont devenues essentielles à de nombreux travaux en linguistique qui aujourd'hui comportent souvent une dimension quantitative et expérimentale qui ne peut être menée à bien que s'il existe des jeux de données qui contiennent des annotations fiables, riches et systématiques et qui disposent d'une couverture suffisante, pour permettre notamment un traitement statistique de leurs résultats. Démonette répond sur ce plan à un besoin qui est resté longtemps non satisfait. En effet, les bases existantes qui disposent d'une large couverture comme Unimorph (Batsuren et al., 2022) contiennent une description limitée de la construction des lexèmes complexes. À l'inverse, d'autres ressources comme Lexique (New, 2006) contiennent des annotations riches et détaillées, mais ont une couverture réduite pour ce qui concerne les relations dérivationnelles. Démonette est de fait la première réponse effective au constat que Dal et al. (1998) faisaient il y a 25 ans sur le besoin pour le français d'une ressource similaire à la base CELEX (Baayen, 1995).

Lexique, 33 (2023), p. 1-5
e-ISSN : 2804-7397

Démonette a l'ambition d'être une ressource de référence pour la morphologie dérivationnelle du français. Sa qualité provient essentiellement de sa couverture et de la fiabilité des annotations qu'elle contient. Pour atteindre ce double objectif (que l'on pourrait à première vue considérer comme contradictoire), Démonette a dès sa conception été alimentée par des ressources existantes, réalisées par des linguistes, notamment dans le cadre de travaux de thèse. Par ailleurs, elle apporte une visibilité supplémentaire à ces travaux et facilite leur utilisation.

Nous nous sommes attachés dans ce numéro à présenter des utilisations concrètes de la base Démonette. La première utilisation est probablement la recherche d'exemples par l'intermédiaire du site Web de la base. Ce site contient également des fonctionnalités et des outils dédiés (comme les transcriptions phonétiques, la fréquence des mots ou la génération de pseudo-mots) dont certains font l'objet d'articles publiés dans le présent numéro. La base en elle-même a évidemment vocation à servir la **recherche en morphologie** et son enseignement dans le supérieur. Elle a déjà contribué au développement d'une morphologie plus quantitative et expérimentale, comme cela est décrit dans Thuilier et al. (2023) et constitue depuis sa première version une ressource exploitable dans les chaînes de **traitement automatique des langues** (Kyjánek, 2018 ; Kyjánek et al., 2020) ; son utilisation dans des tâches de **remédiation orthophonique** pour la mise au point de matériel d'évaluation et de thérapie ciblée sur le niveau morphologique a déjà donné lieu à des résultats (Da Silva et al., 2019 ; Duboisdindien & Cattini, 2021) ; comme cela est montré dans ce numéro, d'autres applications sont attendues dans **l'enseignement en cycle 2**, où la base peut participer à la diversification des techniques d'enseignement du vocabulaire qui sont à la disposition des professeurs des écoles.

Plusieurs travaux ont aussi été réalisés dans le cadre du projet Demonext en vue de compléter le contenu de la base et d'enrichir ses annotations. C'est le cas de l'étude de Juniarta et al. (2022) dont le but est d'identifier des relations erronées ou manquantes en utilisant l'analyse formelle de concepts.

L'une des spécificités de Démonette est d'être une base de relations dérivationnelles décrites « à plat » que l'on peut considérer comme relativement œcuménique dans la mesure où ses descriptions ne reposent sur aucune hypothèse théorique particulière. Elles sont notamment compatibles avec le cadre théorique de la morphologie paradigmatique (Bauer, 1997 ; Stump, 2017 ; Bonami & Strnadová, 2019 ; Hathout & Namer 2019, 2022) dans lesquelles les familles dérivationnelles ont une place centrale.

Les articles et contributions de ce numéro thématique se répartissent en trois parties. La première partie contient un article de référence « Démonette-2, a derivational database for

French with broad lexical coverage and fine-grained morphological descriptions » de **F. Namer et al.** qui présente la base Démonette, son évolution depuis la première version, son contenu actuel et son interface d'interrogation. L'article est cosigné par l'ensemble des membres du projet Demonext (à quelques exceptions près). Il est complété par la présentation de deux travaux qui contribueront à terme à l'amélioration des descriptions présentes dans la base. Tout d'abord, celui de **M. Huguin et al.** « Typage sémantique des noms dans la ressource morphologique Démonette » est consacré à l'annotation sémantique de 286 790 noms de la base. Ce codage utilise les étiquettes du projet FrSemCor (Barque et al., 2020) et un guide d'annotation adapté du même projet. Ces annotations sémantiques seront prochainement intégrées à la table des lexèmes de Démonette. Le troisième article de la première partie, de **B. Calderone et al.** « Phonolette: a grapheme-to-phoneme converter for French » décrit un phonologiseur du français capable de produire des transcriptions phonémiques à partir de formes fléchies graphémiques. Le système est implanté sous la forme d'un réseau de neurones seq2seq entraîné sur les lexiques Flexique (Bonami et al., 2014) et GLàFF (Sajous et al., 2013).

La deuxième partie illustre quelques-unes des utilisations de Démonette les plus récentes dans le domaine de l'enseignement (universitaire ou scolaire) et pour la remédiation orthophonique (adultes aphasiques ou enfants en situation de troubles développementaux du langage). Elle comporte cinq articles. Le premier, « Generation of exercises for derivational morphology using the Démonette database » de **N. Hathout et al.,** présente un système de génération d'exercices de morphologie qui permet de produire des énoncés en grand nombre en faisant varier les exemples qu'ils contiennent. Ces exemples sont extraits des tables de Démonette et de ressources additionnelles. Le deuxième : « Explorer des corpus oraux à l'aide de la base de données Démonette-2 : usage de mots construits dans des interactions adulte(s)-enfant(s) » de **S. Caët et al.,** porte d'une part sur l'effet de la fréquence des mots construit dans les productions spontanées d'adultes et d'enfants et d'autre part sur la manière dont les adultes étayent la compréhension des mots construits par les enfants. Dans le troisième article, de **F. Brin et F. Namer** : « Mesurer la similarité morphologique entre mot produit et mot attendu chez les adultes avec aphasie : étude pilote », les autrices (i) mettent au point une méthode de mesure du niveau de similarité morphologique entre le mot attendu dans une tâche de dénomination et la réponse produite par le patient aphasique, et (ii) vérifient s'il serait possible d'en déduire des profils de patients. Le quatrième article, « Améliorer les compétences lexicales dans le cadre d'un Trouble Développemental du Langage avec la base Démonette-2 » de **G. Duboisdindien et al.**, a pour objectif, à travers une situation clinique scénarisée, d'accompagner les orthophonistes dans leur souhait de développer une intervention en morphologie dérivationnelle visant à améliorer les compétences lexicales d'un patient de neuf ans présentant un Trouble Développemental du Langage. Enfin, le cinquième article, « Programme de recherche participative DEMONEXT : partenariat et co-construction des

savoirs entre chercheurs et orthophonistes » de **G. Duboisdindien et G. Dal**, présente quatre axes de travail relevant de la recherche participative et déployés pendant le projet Demonext, qui ont permis de coconstruire du savoir entre des chercheurs institutionnels et des orthophonistes, d'ajuster l'ergonomie de la base de données morphologiques, cœur du projet, et de valoriser le potentiel de la morphologie dérivationnelle comme moyen d'étayage clinique pour les usagers (orthophonistes et patients, apprenants).

La troisième et dernière partie est consacrée aux réseaux qui composent les familles morphologiques. Ces réseaux constituent l'un des principaux piliers théoriques sur lesquels Démonette a été conçue. Nous proposons dans ce numéro un article de référence sur cette question, intitulé : « Les familles dérivationnelles : comment ça marche ? » de **M. Roché**, dans lequel l'auteur caractérise tout d'abord ce qu'il nomme réseau ACTIVITÉ et réseau ACTION, puis présente un inventaire d'autres types de réseaux dérivationnels. Cet article est précédé de « Repères critiques sur "Les familles dérivationnelles : comment ça marche ?" » de **B. Fradin**, qui le contextualise, et qui présente un ensemble de « clés de lecture » qui permettent au lecteur de l'aborder plus facilement.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database. CD-ROM*. Linguistic Data Consortium, University of Pennsylvania.

Batsuren, K., et al. (2022). UniMorph 4.0: Universal Morphology. *Thirteenth Language Resources and Evaluation Conference* (pp. 840-855). Marseille, European Language Resources Association. https://aclanthology.org/2022.lrec-1.89.pdf

Barque, L., Haas, P., Huyghe, R., Tribout, D., Candito, M., Crabbé, B., & Segonne, V. (2020). FrSemCor: Annotating a French corpus with supersenses. *12th Edition of its Language Resources and Evaluation Conference (LREC)*. Marseille, ELRA. https://aclanthology.org/2020.lrec-1.724.pdf

Bauer, L. (1997). Derivational Paradigms. In G. Booij, & J. van Marle (Eds). *Yearbook of Morphology 1996* (pp. 243-256). Springer.

Bonami, O., Caron, G., & Plancq, C. (2014). Construction d'un lexique flexionnel phonétisé libre du français. *4e Congrès Mondial de Linguistique Française*, Berlin, ILF-CNRS. https://doi.org/10.1051/shsconf/20140801223

Bonami, O., & Strnadová, J. (2019). Paradigm structure and predictability in derivational morphology. *Morphology 29*, 167-197. https://doi.org/10.1007/s11525-018-9322-6

Dal, G., Hathout, N., & Namer, F. (1999). Construire un lexique dérivationnel : théorie et réalisation. *Actes de la 6e conférence sur le Traitement Automatique des Langues Naturelle (TALN'99)*, (pp. 115-124), Cargèse, ATALA.

Da Silva Genest, C., Caët, S., Macchi, L., Tran, M., & Masson, C. (2019). Intérêt orthophonique d'une base de données morphologiques. 1er Congrès inter-universitaire du Collège des Centres de Formation Universitaires en Orthophonie (5-9 avril 2019), Nice.

Duboisdindien, G., & Cattini, J. (2021). La conscience morphologique comme moyen d'intervention : Illustration clinique. *Les Cahiers de l'Association Scientifique et Éthique des Logopèdes Francophones*, *18*(2), 27-34. https://hal.science/hal-03690296/document

Hathout, N., & Namer, F. (2019). Paradigms in word formation: what are we up to? *Morphology*, *29*, 153-165. https://doi.org/10.1007/s11525-019-09344-3

Hathout, N., & Namer, F. (2022). ParaDis: a Family and Paradigm Model. *Morphology*, *32*, 153-195. https://doi.org/10.1007/s11525-021-09390-w

Juniarta, N., Bonami, O., Hathout, N., Namer, F., & Toussaint, Y. (2022). Organizing and Improving a Database of French Word Formation Using Formal Concept Analysis. *LREC 2022*, Marseille (pp. 3969-3976). European Language Resources Association (ELRA). https://aclanthology.org/2022.lrec-1.422.pdf

Kyjánek, L. (2018). *Morphological Resources of Derivational Word-Formation Relations*. Technical Report TR-2018-61. Charles University, ÚFAL. https://ufal.mff.cuni.cz/techrep/tr61.pdf

Kyjánek, L., Žabokrtský, Z., Ševčíková, M., & Vidra, J. (2020). Universal Derivations 1.0, A Growing Collection of Harmonised Word-Formation Resources. *The Prague Bulletin of Mathematical Linguistics*, *115*, 5-30. https://ufal.mff.cuni.cz/pbml/115/art-kyjanek-et-al.pdf

New, B. (2006). Lexique 3 : Une nouvelle base de données lexicales. *TALN 2006*, Louvain.

Sajous, F., Hathout, N., & Calderone, B. (2013). GLàFF, un gros lexique à tout faire du français. *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)* (pp. 285-298). Les Sables-d'Olonne. https://aclanthology.org/F13-1021.pdf

Stump, G. (2017). The Nature and Dimensions of Complexity in Morphology. *Annual Review of Linguistics 3*(1), 65-83.

Thuilier, J., Tribout, D., & Wauquier, M. (2023). Affixal rivalry in French demonyms formation: The role of linguistic and non-linguistic parameters. *Word Structure*, *16*(1), 115-146. https://doi.org/10.3366/word.2023.0223