Received: March 2023 / Accepted: June 2023
Published on line: December 2023

Typage sémantique des noms dans la ressource morphologique Démonette

Semantic typing of nouns in the Demonette morphological resource

Mathilde Huguin
Université de Lorraine & ATILF (UMR 7118, CNRS & Université de Lorraine)

Lucie Barque
Université Sorbonne Paris Nord & LLF (UMR 7110, CNRS & Université Paris Cité)

Pauline Haas
Université Sorbonne Paris Nord & Lattice (UMR 8094, CNRS, ENS & Paris Sorbonne Nouvelle ; PSL & USPC)

Delphine Tribout
Université de Lille & STL (UMR 8163, CNRS & Université de Lille)

Résumé

Cet article décrit la méthodologie mise en place pour effectuer l'annotation sémantique d'une partie des noms de la base de données morphologiques Démonette-2. Nous y présentons d'abord le jeu d'étiquettes sémantiques sélectionné pour effectuer cette annotation. Ce jeu d'étiquettes est une adaptation révisée des *Unique Beginners* de Wordnet et chaque étiquette est accompagnée d'une définition et de tests linguistiques permettant l'attribution d'une étiquette à un nom. Nous détaillons ensuite les deux méthodes utilisées pour annoter les lexèmes nominaux. La première méthode, automatique, a consisté à apparier les étiquettes présentes dans les bases de données morphologiques alimentant Démonette-2 avec le jeu d'étiquettes révisé. La seconde méthode a consisté à annoter manuellement un sous-ensemble de noms. Nous donnons enfin un bilan quantitatif de notre annotation en présentant notamment la distribution des noms monosémiques / polysémiques et les étiquettes sémantiques les plus fréquentes. Ce premier travail d'annotation sémantique fournit un ensemble de 58 099 noms disposant d'une ou plusieurs étiquettes sémantiques. Cet ensemble de noms offre déjà de multiples possibilités d'analyses, impossibles sans accès à une large base de données annotées sémantiquement et morphologiquement, comme l'étude de procédés concurrents ou encore l'examen de la polyfonctionnalité des affixes.

Mots-clés : annotation sémantique, Démonette-2, supersens, sémantique lexicale

Abstract

This article describes the methodology used to carry out the semantic annotation of part of the nouns in the Démonette-2 morphological database. First, we present the set of semantic labels selected for this annotation. This set of labels is a revised adaptation of Wordnet's *Unique Beginners*, and each label is provided by a definition and linguistic tests enabling a label to be assigned to a noun. We then describe the two methods used to annotate the nouns. The first one is an automatic method that matches the labels present in the morphological databases feeding Démonette-2 with the revised set of labels. The second method involved manually annotating a subset of nouns. Finally, we give a quantitative assessment of our annotation, presenting the distribution of monosemic/polysemic nouns and the most frequent semantic labels. This initial semantic annotation work provides a set of 58,099 nouns with one or more semantic labels. This set of nouns already offers a host of analytical possibilities that would be impossible without access to a large database of semantically and morphologically annotated data, such as the study of competing processes or the examination of the polyfunctionality of affixes.

Keywords: semantic annotation, Démonette-2, supersense, lexical semantics

Financement : Cette recherche a été financée par l'Agence Nationale de la Recherche (France) dans le cadre du projet DEMONEXT [ID : ANR-17-CE23-0005].

1. Introduction

Nous présentons dans cet article l'annotation sémantique effectuée sur une partie des noms de la table des lexèmes de la base de données morphologiques Démonette (désormais Démonette-2), qui compte actuellement 286 790 entrées nominales. Disposer d'une base de données à large couverture, alliant informations morphologiques et sémantiques, est pertinent dans une double mesure : d'une part cela permet la description des liens entre types de formations morphologiques et classes sémantiques ; et d'autre part, cela permet l'observation des divers procédés morphologiques formant des noms relevant d'une classe sémantique donnée afin d'évaluer la concurrence entre affixes, ou encore la mise en évidence des cas de polysémie, notamment pour quantifier les alternances de sens les plus récurrentes. Les objectifs principaux de cet article sont de présenter le plus précisément possible la méthodologie d'annotation adoptée (Section 2) et d'offrir aux utilisateurs de la base une vision d'ensemble des données sémantiques disponibles à ce stade (Section 3).

Avant d'entrer dans le vif du sujet, il nous semble important de faire un point terminologique pour faciliter la lecture de la présentation qui va suivre. On désignera ici par *entrée nominale* (d'une ressource) l'association d'un lemme et d'une catégorie grammaticale. Par exemple, le lemme *mandataire* a deux entrées nominales dans Démonette-2 : *mandataire-nf* et *mandataire-nm*. Pour ce qui est de la description sémantique, une entrée nominale est

associée à autant d'*étiquettes sémantiques* que le nom décrit dans cette entrée a de sens. L'entrée nominale *mandataire-nf* est ainsi associée à une seule étiquette puisque le nom est monosémique (il désigne uniquement une personne). Une entrée comme *gomme-nf*, en revanche, est associée à plusieurs étiquettes puisque le nom peut désigner aussi bien une substance qu'un objet fait de cette substance. Notons que lorsqu'une entrée se voit associer plusieurs étiquettes, dans le cas des noms à sens multiples, l'ordre des étiquettes suit un classement alphabétique, et non un rapport de dominance entre les sens décrits.

2. Méthodologie d'annotation

Dans cette section, nous commençons par présenter le jeu d'étiquettes sémantiques choisi pour décrire la sémantique des entrées nominales de Démonette-2 (Section 2.1). Nous détaillons ensuite la méthode utilisée pour importer les informations sémantiques déjà disponibles dans les bases de données morphologiques alimentant Démonette-2, désormais BDM-sources (Section 2.2). Enfin, nous précisons la méthode utilisée pour l'annotation manuelle d'un second sous-ensemble d'entrées nominales de Démonette-2 (Section 2.3).

2.1. Jeu d'étiquettes

Le jeu d'étiquettes choisi pour l'annotation des noms dans Démonette-2 est composé d'étiquettes représentant des classes sémantiques d'un grain assez large, tels que `Act` pour les noms d'action ou encore `Person` pour les noms d'humain. Ces étiquettes, souvent appelées *supersenses* dans la littérature (Ciaramita & Johnson, 2003), sont issues des *Unique Beginners* de Wordnet (Miller et al., 1990 ; Fellbaum 1998). Elles ont été adaptées pour le français dans le cadre d'une annotation d'occurrences nominales en corpus (corpus *FrSemCor,* Barque et al., 2020) et ont été reprises ici dans l'optique d'une annotation lexicale, hors contexte.

Chaque étiquette[1] correspond à une classe sémantique, décrite au moyen d'une définition et associée à un ou plusieurs tests linguistiques indicatifs, comme illustré en (1) ci-dessous.

(1) **Classe** : `Event`
Définition : situation dynamique sans agent. Inclut les événements naturels (*avalanche*), les autres événements fortuits non agentifs (*rupture*), les noms dénotant un processus naturel de changement d'état (*coagulation*).
Test indicatif : Dét N {a eu lieu / s'est produit} {à tel moment / à tel endroit}

Notre jeu d'étiquettes est constitué de 43 étiquettes sémantiques hiérarchisées, correspondant à 22 classes sémantiques simples (voir Figure 1 ci-dessous) et 21 classes sémantiques complexes. Les étiquettes complexes sont construites à partir des étiquettes simples combinées

[1] Pour une description détaillée des étiquettes, voir le guide d'annotation du projet de Huguin et al. (2022).

à l'aide des deux opérateurs « x » et « + ». L'opérateur « x » indique une distribution de sens et permet d'annoter les noms collectifs comme *famille* (`GroupxPerson`) ou *meute* (`GroupxAnimal`). L'opérateur « + » indique une conjonction de sens et permet d'annoter les noms à facettes comme *livre* (`Artifact+Cognition`) ou *exposé* (`Act+Cognition`). Les noms à facettes sont des noms présentant des interprétations distinctes mais compatibles en contexte (Cruse 1986, 1995 ; Godard & Jayez, 1996 ; Asher & Pustejovsky, 2006 ; entre autres). Le nom *livre,* par exemple, peut dénoter un objet physique (*un livre abîmé*) et un contenu informationnel (*un livre intéressant*) et ces deux interprétations peuvent être mobilisées conjointement en contexte, comme l'indique le test de la coprédication (*un livre abîmé mais intéressant*).

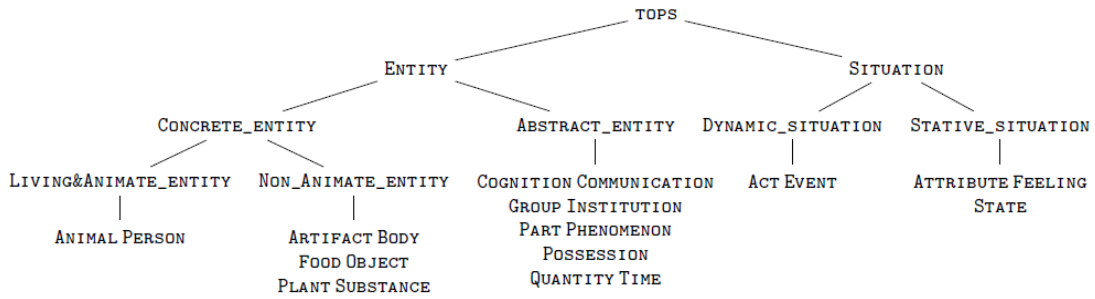

Figure 1. Hiérarchie des étiquettes simples

Comme nous le verrons dans la description des données qui va suivre, l'utilisation des étiquettes intermédiaires de la hiérarchie (ex. `Entity`) est rare dans le cadre de l'annotation, puisqu'elle est réservée au cas des noms sous-spécifiés sémantiquement (voir Section 3.1 l'exemple du nom *concurrent*).

2.2. Première phase : importation des données issues des BDM-sources

Une partie des noms issus des BDM-sources alimentant la base Démonette-2 présentaient déjà un étiquetage sémantique de type ontologique, mais non uniformisé puisqu' issus de différents projets. Le Tableau 1 donne la répartition des données proposant des informations sémantiques dans les BDM-sources. La dernière colonne indique le total des lemmes, entrées et sens sans les doublons. Certaines ressources peuvent en effet partager une même entrée lexicale et une même description. Par exemple, le nom féminin *abstraction* est présent dans Démonette-1.2 et dans DiMoC et a reçu la même étiquette dans ces deux ressources. Il est par ailleurs possible que deux ressources contiennent une même entrée lexicale mais proposent des descriptions complémentaires pour cette entrée. C'est le cas du nom *articulation*, qui est décrit dans les deux mêmes BDM-sources mais le sens actionnel du nom (*l'articulation d'un son*) est décrit dans la première tandis que le sens anatomique (*l'articulation de la jambe*) est

décrit dans la seconde. Les noms *abstraction* et *articulation* ont donc chacun une seule entrée nominale comptabilisée dans le total sans doublon et le sens `Act` du nom *abstraction* n'y est comptabilisé qu'une fois.

BDM-sources	Démonette-1.2	DiMoC	Mordan	Convers	Denom	Total	**Total sans doublons**
Nb de lemmes	14 522	28 506	1 896	567	303	45 794	**44 403**
Nb d'entrées	14 529	32 418	1 896	567	333	49 743	**48 328**
Nb de sens	14 529	33 601	2 148	573	333	51 184	**50 328**

Tableau 1. Effectifs des lemmes / entrées / sens sémantiquement informés dans les BDM-sources

Notre premier objectif a été de traiter ces données existantes en transposant les informations sémantiques contenues dans les BDM-sources en étiquettes Démonette-2. Pour ce faire, un appariement a été effectué manuellement entre chaque type sémantique proposé dans les ressources et notre propre jeu d'étiquettes, donnant lieu à une grille de conversion. Ainsi, les noms typés AGF (agent féminin) ou AGM (agent masculin) dans Démonette-1.2 se sont vu attribuer automatiquement l'étiquette `Person` de notre jeu d'étiquettes. Le résultat de l'appariement des 50 328 sens présents dans les BDM-sources est présenté dans le Tableau 2. Les types de sens y sont classés par ordre de fréquence décroissante.

Étiquettes présentes dans les BDM-sources	**Étiquettes de Démonette-2**	**Nb de sens**
Démonette-1.2 : @AGF ; @AGM / **DiMoC :** ha ; ahg ; age / **Denom :** Nc person ; Np	`Person`	17 479
Démonette-1.2 : @ACT / **DiMoC** : na ; ne / **Converts** : action	`Act`	11 872
DiMoC : ob ; obb ; obm ; obv ; oi ; oiv ; ol / ol1 / ol2 ; olb / **Mordan :** objet concret ; lieu / **Denom** : Nc lieu, Nc objet / **Converts** : artefact	`Artifact`	6 771
DiMoC : nr ; pn ; pr ; prt / **Denom** : Nc matière ; Nc matière-végétale / **Converts** : objet	`Substance`	2 788
DiMoC : ns / **Mordan** : fonction ; taux ; mesure ; quantification ; relation ; relation spatiale ; méronymie ; autre N de propriété ; activité ; occurrence ; couleur ; propriété feinte ; attitude ; propriété chromatique	`Attribute`	1 860
DiMoC : on / **Denom** : Nc nature / **Converts** : objet naturel	`Object`	1 769
DiMoC : an / **Converts** : animé / **Denom** : Nc animal	`Animal`	1 602
DiMoC : op / **Denom** : Nc végétal	`Plant`	1 515
Denom : Nc discipline / **DiMoC :** oa, oac / **Converts** : technique	`Cognition`	1 086

DiMoC : nq / nq2 ; nu / **Modan** : appartenance identitaire / **Denom** : Nc maladie / **Converts** : état	`State`	1 116
DiMoC : ahc / **Mordan** : ensemble	`GroupxPerson`	518
DiMoC : opc / **Converts** : lieu	`GroupxPlant`	444
DiMoC : oal	`Communication`	339
DiMoC : oam	`Possession`	271
DiMoC : nt / **Converts** : phénomène naturel	`Event`	237
DiMoC : anc ; ans	`GroupxAnimal`	223
DiMoC : oat / **Mordan** : période	`Time`	217
DimoC : obc ; oic	`GroupxArtifact`	152
Denom : Nc partie du corps / **Converts** : partie du corps	`Body`	47
Converts : sentiment / **Denom** : Nc abstrait	`Feeling`	13
Denom : Nc nourriture	`Food`	8
Converts : unité de mesure	`Quantity`	1
Total		**50 328**

Tableau 2. Répartition des étiquettes suite à l'importation automatique des données sémantiques des BDM-sources

La transposition des données issues des BDM-sources est intéressante car moins coûteuse qu'une annotation *ex nihilo*. Cet appariement nous a permis d'annoter environ un sixième des entrées nominales de Démonette-2 (48 328 / 286 790). Il est toutefois important de souligner que ce premier ensemble de données, dont nous proposerons une évaluation qualitative dans la Section 2.4, devra par la suite faire l'objet d'une révision manuelle pour corriger les erreurs d'appariement et augmenter la couverture des sens décrits. En effet, la polysémie éventuelle des noms traités dans ces ressources n'a pas été décrite en tant que telle. Si certains noms se sont vu attribuer plusieurs étiquettes provenant d'une ou plusieurs BDM-sources (1 857 noms, parmi lesquels *gomme*, dont les deux sens `Substance` et `Artefact` sont encodés dans DiMoC), la polysémie de la plupart des noms n'a pas été décrite de manière extensive. Par exemple, le nom *bec* est étiqueté `Artefact` dans la ressource DiMoC (*le bec de la théière)*, or, ce nom est polysémique et doit aussi recevoir l'étiquette `Body` (*le bec d'un oiseau*). De même, la polyfonctionnalité éventuelle des affixes n'a pas été systématiquement prise en compte. Par exemple, dans Démonette-1.2, les suffixes agentifs produisent majoritairement des noms d'humain (ex. *agricultrice*, *danseur*), mais également des noms d'instrument (ex. *tondeuse*, *tracteur*) or c'est uniquement la classe majoritaire de ces suffixes agentifs dans Démonette-1.2 (`Person`) qui a été attribuée lors de l'appariement automatique. Par ailleurs, il faut garder à l'esprit que les noms provenant des BDM-sources ne reflètent pas le lexique général. En effet, les BDM-sources ont été créées dans le cadre d'études morphologiques spécifiques, portant généralement sur quelques procédés de formation morphologique, ce qui n'est pas sans conséquence sur les proportions des classes sémantiques répertoriées. Ainsi, l'étiquette

`GroupxPlant` semble surreprésentée dans cet échantillon, parce que le suffixe produisant des noms d'ensemble de plantes (*-aie*) a fait l'objet d'une étude morphologique spécifique dans l'une des BDM-sources (DiMoC, cf. Roché, 2011). En somme, les informations sémantiques figurant dans les BDM-sources, bien qu'initialement produites de manière manuelle par les auteurs de ces bases, sont issues de méthodes et de principes d'annotation hétérogènes et doivent faire l'objet d'une homogénéisation que l'appariement automatique avec le jeu d'étiquettes de Démonette-2 n'accomplit que partiellement (voir la Section 2.4 pour plus de détails).

2.3. Seconde phase : annotation manuelle d'un nouvel ensemble de noms

La seconde étape d'annotation a visé d'une part à effectuer une évaluation des données importées automatiquement depuis les BDM-sources, et d'autre part à augmenter la couverture de la description sémantique dans la base Démonette-2. Pour ce faire, nous avons sélectionné un nouvel échantillon de 16 902 entrées nominales, réparties comme suit :

- 7 131 entrées nominales issues des BDM-sources ont fait l'objet d'une nouvelle annotation manuelle, sans accès aux informations importées automatiquement.
- 9 771 nouvelles entrées nominales de Démonette-2 ont été sélectionnées pour augmenter la couverture de la description sémantique. Ce second sous-ensemble inclut notamment 2 674 noms du corpus *FrSemcor* (Barque et al., 2020) figurant dans la table des lexèmes de Démonette-2, dont les descriptions ont été révisées et complétées pour décrire extensivement la polysémie éventuelle de ces noms. Il inclut également 4 883 noms[2] de la base *Echantinom* (Bonami & Tribout, 2021). Les 2 214 entrées nominales restantes ont été sélectionnées de manière aléatoire.

La méthode d'annotation manuelle suit la procédure décrite en détail dans le guide d'annotation sémantique des noms de Démonette-2, auquel nous renvoyons le lecteur (cf. note 1). L'attribution d'une ou plusieurs étiquettes sémantiques à chacune des 16 902 entrées nominales, selon les emplois du nom décrit, s'est ainsi appuyée sur : (i) la description du nom dans un dictionnaire de référence (e.g. *TLFi*) ; (ii) les propriétés distributionnelles du nom mises en évidence par les tests, voir exemple (1) ; (iii) la vérification de certains emplois dans des corpus de référence (e.g. *FrTenTen2020*, voir Jakubicek et al., 2013).

Afin d'évaluer le caractère opératoire de cette méthode, nous avons effectué une première phase d'annotation en double aveugle sur 144 noms correspondant à l'intersection des cinq BDM-sources utilisées. Cette phase nous a permis d'identifier les difficultés inhérentes à

[2] Sur les 5 000 entrées nominales de la base *Echantinom*, seules 4 883 sont effectivement enregistrées dans Démonette-2.

l'annotation lexicale (accord brut de 0,59 et kappa de Cohen 0,5)[3], en particulier les questions du découpage polysémique et de la lexicalisation des sens, et d'affiner le guide d'annotation avant de procéder à l'annotation manuelle des 16 902 noms.

2.4. Évaluation de l'appariement automatique

Comme indiqué précédemment, un sous-ensemble de 7 131 noms a été doublement annoté de manière indépendante : par appariement automatique de données natives des BDM-sources, d'une part, par annotation manuelle en suivant la méthode qui vient d'être décrite, d'autre part. Cette double annotation nous permet d'obtenir une évaluation des données, en calculant la précision et le rappel des données obtenues par appariement automatique par rapport à l'annotation manuelle, considérée ici comme valeur de référence.

Le calcul de la précision a été obtenu, pour chaque entrée lexicale, en divisant le nombre de sens corrects (c'est-à-dire en accord avec la référence) par le nombre total de sens parmi les données importées des BDM-sources. Le rappel, calculé quant à lui pour évaluer la couverture des sens décrits dans les cas de noms ambigus, correspond au nombre de sens corrects divisé cette fois par le nombre total de sens lexicaux dans la référence. Compte tenu de la présence dans nos données d'un nombre important de sens complexes (i.e. représentés par des étiquettes construites à l'aide d'un opérateur, voir Section 2.1), deux types d'accord ont été pris en compte.

- Accord strict : la comparaison s'effectue sur les étiquettes prises dans leur intégralité.
- Accord partiel : une description est considérée comme correcte si, en cas de sens complexe, au moins l'une des étiquettes comparées est incluse dans l'autre, ou inversement.

Prenons l'exemple du nom *accusation*, qui a reçu l'étiquette `Act` dans les données importées des BDM-sources et les deux étiquettes `Act+Cognition` (***cette grave accusation a été portée devant témoin***) et `Institution` (***l'accusation est représentée par l'avocat général***) dans l'annotation manuelle, considérée ici comme la référence. D'un point de vue strict, il n'y a pas d'accord entre les deux annotations mais si l'on accepte les recouvrements partiels entre constituant d'une étiquette complexe, il y a accord sur le sens actionnel du nom (entre `Act` et `Act+Cognition`). Le Tableau 3 donne les résultats de l'accord observé sur l'ensemble de notre échantillon, pour les deux types d'accord considérés.

[3] Les 144 noms ont donné lieu à l'étiquetage de 297 sens (polysémie moyenne dans ce sous-ensemble : 2,05). On compte 194 / 328 accords. Deux types de désaccords sont à distinguer :
(i) difficultés d'identification d'un sens donné : les deux annotatrices ont ciblé un même sens du nom mais n'ont pas donné la même étiquette ;
(ii) difficultés liées au découpage de la polysémie : une des annotatrices a décrit un sens du nom que l'autre n'a pas décrit.

	Accord strict (%)	**Accord partiel (%)**
Précision	71,66	76,87
Rappel	53,61	57,50

Tableau 3. Évaluation effectuée sur un sous-ensemble (env. 15 %) des données importées automatiquement des BDM-sources

Une analyse des données révèle deux grands types de désaccords. Dans le premier cas, un même sens est visé mais l'étiquette choisie pour le décrire est plus ou moins générale. Il s'agit donc d'un problème de précision, mais qui se trouve, dans la majorité des cas, être davantage un problème de degré de granularité qu'une erreur proprement dite. Ainsi, le sens anatomique du nom *tendon* est décrit par l'étiquette `Object` dans les données importées et par l'étiquette `Body` dans la référence. Le second type de désaccord concerne la couverture des sens. Le nom masculin *aigle*, par exemple, est décrit par les sens `Animal` et `Artifact` dans les données importées et par les sens `Animal` et `Person` dans la référence. Les deux désaccords apparents concernent ici des sens attestés du nom, qui peut en effet désigner une personne et un artefact (*pupitre d'une église*). Ces deux sens étant plus rares, la question se pose toutefois de leur inclusion ou non dans une description sémantique idéale du nom.

Cette analyse des désaccords vient confirmer que l'annotation sémantique des sens lexicaux d'un nom est une tâche difficile, qui requiert de décider quels sens doivent être pris en compte. Le choix de prendre les données issues de l'annotation manuelle comme référence peut également être discuté. Toutefois, la polysémie y est décrite de manière plus extensive que dans les données importées des BDM-sources (10 643 sens dans la référence contre 7 966 sens dans les données issues des BDM-sources). Quoi qu'il en soit, ces données doublement annotées sont conservées dans Démonette-2, avec indication de la source de l'annotation, permettant une révision future et une unification des descriptions le cas échéant.

3. Annotation manuelle : bilan quantitatif

Nous nous proposons à présent de faire un bilan quantitatif des données sémantiques produites dans le cadre de l'annotation manuelle. Nous laissons de côté les données appariées automatiquement depuis les BDM-sources car, comme on vient de le voir, celles-ci sont (i) globalement moins complètes du point de vue de la description des noms à sens multiples, (ii) plus sujettes à erreur puisque traduites automatiquement, et enfin (iii) moins représentatives du lexique général puisque provenant de ressources morphologiques spécifiques. Après un bilan sémantique général où nous présentons la répartition des noms monosémiques / polysémiques et examinons les étiquettes les plus fréquentes (Section 3.1), nous donnons quelques chiffres sur les procédés morphologiques en jeu et les suffixes présents dans les noms annotés manuellement (Section 3.2).

3.1. Bilan sémantique

Le Tableau 4 donne la répartition des types d'entrées et d'étiquettes pour les 16 902 noms de la base ayant bénéficié d'une annotation manuelle. On remarque que les noms monosémiques sont beaucoup plus nombreux que les noms ambigus, qui désignent ici les noms auxquels ont été attribuées plusieurs étiquettes représentant des sens distincts liés par polysémie (ex. *copie*) ou par homonymie (ex. *avocat*). Cette caractéristique s'explique tout d'abord par le degré de granularité sémantique utilisé pour l'annotation. Un nom n'ayant reçu qu'une seule étiquette est « monosémique » au regard du grain de l'annotation effectuée mais pourrait être polysémique dans le cadre d'une annotation plus fine. Par exemple, le nom *bureau* est polysémique : il peut désigner une pièce (*un bureau avec fenêtre*) ou un meuble (*un bureau avec deux tiroirs*). Comme ces deux sens correspondent à la même étiquette dans Démonette-2 (`Artifact`), la polysémie est invisible dans la base. La prédominance de la monosémie s'explique également par la nature de nos données. Par exemple, les noms de plantation et les gentilés, présents en grand nombre dans la base DiMoC, sont rarement polysémiques.

Nb d'entrées		
N monosémiques	13 094	ex. *acacia* (`Plant`), *abricoteraie* (`GroupxPlant`)
N ambigus	3 808	ex. *copie* (`Act;Artifact+Cognition`), *avocat* (`Food;Person`)
Total	**16 902**	
Nb d'étiquettes		
Simples	21 403	ex. `Plant, Act, Artifact`
Complexes	796	ex. `GroupxPlant, Artifact+Cognition`
Total	**22 191**	
Taux d'ambiguïté		
Pour l'ensemble des N	1,31	
Pour les N à sens multiples	2,39	

Tableau 4. Propriétés sémantiques des entrées dans l'échantillon annoté manuellement

Le Tableau 5 présente la distribution des 20 types de sens lexicaux les plus fréquents dans l'échantillon annoté manuellement. Les sens lexicaux de type `Person` représentent à eux seuls 24 % des données, ce qui s'explique en partie par le sous-ensemble des gentilés inclus dans l'échantillon[4]. Globalement, les noms d'entité concrète (`Person`, `Artifact`, `Institution`) y sont majoritaires.

[4] Cette proportion de sens nominaux de type `Person` dans Démonette-2 fait écho à celle des noms de gentilé dans le Wiktionnaire. En effet, une étude des définitions des entrées nominales de cette ressource montre que 18 % d'entre elles (54 822 / 306 530) commencent par « habitant(e) de… » (Angleraud, 2023).

	Nb de sens	Fréq. (%)	Exemples
Person	5 376	24	*sénateur*
Artifact	2 388	11	*anti-vol*
Act	2 329	10	*apiculture*
Institution	2 323	10	*Arconville*
Cognition	1 118	5	*arrière-pensée*
Substance	1 038	5	*enzyme*
State	1 008	5	*méningite*
Attribute	705	3	*perspicacité*
Plant	510	2	*peuplier*
Animal	499	2	*cerf*
Food	491	2	*champagne*
Event	460	2	*condensation*
Object	417	2	*continent*
Body	401	2	*humérus*
GroupxPlant	398	2	*jasmineraie*
Communication	376	2	*lemme*
Act+Cognition	352	2	*avertissement*
Quantity	278	1	*are*
Artifact+Cognition	195	1	*bouquin*
Time	160	1	*crétacé*
Autres	1 377	6	
Total	22 191		

Tableau 5. Distribution des 20 types sémantiques les plus fréquents dans l'ensemble des données annotées manuellement

Le groupe désigné par *Autres* dans le Tableau 5 regroupe les effectifs de 54 classes différentes. Il s'agit essentiellement de classes représentées par des étiquettes complexes (ex. GroupxPerson, Event+Phenomenon) ou d'étiquettes situées à un niveau supérieur dans la hiérarchie (cf. Figure 1). Par exemple, le nom *concurrent* a reçu l'étiquette Entity (qui englobe les Concrete_entity et les Abstract_entity) car il peut dénoter un humain, un animal, une plante, une institution, mais aussi un travail intellectuel, etc.

On remarque que trois sens complexes figurent parmi les 20 types de sens les plus fréquents dans l'échantillon. Le premier est GroupxPlant, dont la fréquence s'explique par la présence de données issues de DiMoC dans cet ensemble (voir *supra*). Les deux autres, qui figurent également parmi les 20 classes les plus fréquentes dans le corpus *FrSemCor* (Barque et al., 2020), sont Act+Cognition, attribué principalement aux noms d'acte de parole (ex. *souhait*) et Artifact+Cognition, attribué majoritairement à des noms d'objet pourvu d'un contenu informationnel (ex. *bouquin*). Pour ce qui est spécifiquement des noms à facettes (dont les étiquettes sont construites à l'aide de l'opérateur « + »), 84 % d'entre eux sont couverts par

les trois étiquettes `Act+Cognition`, `Artifact+Cognition`, et enfin `Event+State`, attribuée majoritairement à des noms de maladie (ex. *thrombose*). Les étiquettes complexes faisant intervenir les composantes sémantiques 'une partie de' (`Partx`) ou 'un groupe de' (`Groupx`) sont très majoritairement associées à des entités concrètes. Il s'agit des étiquettes `GroupxPlant` (ex. *prunaie*), `GroupxPerson` (ex. *salariat*), `GroupxArtifact` (ex. *trousseau*) et `PartxArtifact` (ex. *poupe*).

On observe enfin, parmi les noms à sens multiples, des alternances sémantiques récurrentes issues en partie d'extensions de sens régulières (Apresjan, 1974) associées ou non à la construction morphologique. Le Tableau 6 ci-dessous donne la liste des alternances les plus fréquentes avec mention de leur effectif et quelques exemples. Les liens de polysémie qui sous-tendent ces alternances régulières peuvent relever de la métonymie, par exemple l'alternance `Act;Artifact` qui concerne les noms pouvant dénoter une action et un objet correspondant au résultat de l'action (ex. *moulage*), à l'instrument de l'action (ex. *chauffage*), voire aux deux (ex. *bandage*). Ils peuvent également relever de la métaphore, comme on le voit avec l'alternance `Animal;Person`, qui repose sur un lien d'analogie entre l'animal et la personne (ex. *sangsue*, voir Goudet et al., 2018). Ces premières constatations devront faire l'objet d'une analyse plus approfondie, afin d'établir au cas par cas l'origine des alternances observées de manière régulière dans le lexique et plus particulièrement au sein du lexique construit (Salvadori & Huyghe, 2022).

Alternances sémantiques	**Nb de N**	**Exemples**
`Act;Cognition`	253	*journalisme, nécromancie, spéléologie*
`Artifact;Person`	221	*baleinier, violon, tireuse, tambour*
`Act;Artifact`	194	*bandage, chauffage, moulage*
`Animal;Person`	167	*bâtard, lapinet, truie, sangsue, sagouin*
`Act;Event`	100	*balancement, rupture, saccade*
`Act;State`	59	*blanchiment, repos, recueillement*
`Artifact;Body`	53	*coffre, menotte, squelette, trique*
`Attribute;State`	51	*agitation, parité, sécurité, tranquillité*
`Attribute;Cognition`	47	*humour, pédagogie, raison, savoir*

Tableau 6. Alternances polysémiques les plus fréquentes parmi les noms à sens multiples

3.2. Bilan morphologique

Les noms construits annotés sont majoritairement issus de suffixation (70 %, ex. *vomissement*) et de conversion (29 %, ex. *annonce*). Les autres procédés comme la préfixation (ex. *antithèse*) et la composition (ex. *nosographie*) sont marginaux (<1 %). La très faible proportion de préfixés et de composés s'explique en partie par le fait que les BDM-sources décrivent essentiellement des procédés mettant en jeu la suffixation ou la conversion. Cependant, ces chiffres sont peut-être aussi représentatifs de la distribution des différents procédés dans le

lexique général. En effet, Bonami et Tribout (2021) ont étudié un échantillon aléatoire de 5 000 noms, parmi lesquels 2 936 noms ont été repérés comme construits morphologiquement. Au sein de ces noms construits, les deux procédés les plus fréquents sont la suffixation (63 %) et la conversion (19 %), ce qui correspond, en ordre de grandeur, aux résultats observables dans Démonette-2.

En ce qui concerne la suffixation, on trouve 112 suffixes différents[5]. Le Tableau 7[6] présente les 10 suffixes les plus fréquents et fournit les annotations sémantiques majoritaires et des exemples associés. Comme l'illustre le tableau, la polyfonctionnalité (Salvadori & Huyghe, 2023) de ces affixes est quasi systématique. Le suffixe *-ion* est celui qui intervient dans le plus de configurations sémantiques différentes. À l'inverse, les noms en *-aie* annotés sont exclusivement des noms dénotant des groupes de plantes.

Suff.	**Fréq. (%)**	**Nb de N**	**Annotations les plus fréquentes**	**Exemples**
-ier	21	820	`Person` (73 %) `Artifact;Person` (8 %) `Animal;Person` (2%) *Autres* (17%)	*vitrier, pilier, barbier*
-ion	11	412	`Act` (28 %) `Act+Cognition` (8 %) `Act;Event` (6 %) `Act;State` (4 %) *Autres* (54 %)	*strangulation, machination, complication, saturation*
-eur	8	319	`Person` (69 %) `Artifact;Person` (11 %) `Attribute` (3 %) *Autres* (16 %)	*tatoueur, torpilleur, minceur*
-ment	6	246	`Act` (35 %) `Act;Artifact` (7 %) `Act;Event` (6 %) *Autres* (51 %)	*tapotement, rangement, grincement*
-aie	6	239	`GroupxPlant` (100 %)	*olivaie*
-iste	5	198	`Person` (99 %) `Artifact;Person` (1 %)	*spécialiste, cycliste*
-aire	5	185	`Person` (81 %) `Artifact` (3 %) *Autres* (16 %)	*vingtenaire, rosaire*
-isme	4	167	`Cognition` (30 %) `Attribute` (13 %) `Attribute;Cognition` (11 %) `Act` (10 %) *Autres* (36 %)	*socialisme, loyalisme, racisme, vandalisme*
-age	3	109	`Act` (44 %) `Act;Artifact` (14 %) `Act;Event` (4 %) `Act+Cognition` (4 %) *Autres* (35 %)	*repiquage, emballage, blocage, témoignage*

[5] La dénomination *suffixe* pourrait être tenue comme abusive. On trouve des exposants identifiés comme des suffixes (*-iste*) mais aussi des constituants néoclassiques (*-logie*) ou des suffixes complexes (*-icien* : *-ique* + *-ien*).

[6] Le tableau 7 ne tient pas compte de l'homonymie des suffixes. Par exemple, la ligne associée au suffixe *-eur* couvre à la fois les cas de constructions de noms d'objet ou de personne déverbaux (ex. *éditeur*) et les cas de noms de propriété désadjectivaux (ex. *hauteur*).

-ité	3	108	`Attribute` (42 %) `Attribute;State` (15%) `State` (14 %) *Autres* (30 %)	*timidité, neutralité, cécité*

Tableau 7. Suffixes les plus fréquents dans l'ensemble des données annotées manuellement

4. Conclusion

Dans cet article, nous avons présenté les méthodes utilisées (appariement automatique et annotation manuelle) pour annoter sémantiquement les entrées nominales de Démonette-2. Le résultat de la campagne d'annotation effectuée dans le cadre du projet couvre environ 20 % de la nomenclature nominale de la table des lexèmes. D'autres campagnes pourront être menées pour étendre cette couverture et pour réviser manuellement les données appariées automatiquement depuis les BDM-sources. L'ensemble des données d'ores et déjà annotées permet toutefois d'envisager différents types d'études relatives aux aspects sémantiques de la dérivation morphologique en français. Ces études pourront venir compléter les études existantes sur la polyfonctionnalité des affixes (e.g. Salvadori & Huyghe, 2023), sur la polysémie des noms construits (e.g. Salvadori & Huyghe, 2022), sur la concurrence entre procédés (e.g. Koehl, 2012 ; Fradin, 2019 ; Bonami & Thuilier, 2019 ; Missud & Villoing, 2020 ; Thuilier et al., 2023) ou encore sur la structure sémantique des familles morphologiques (e.g. Lignon et al. 2014 ; Fradin, 2021 ; Sanacore et al., 2020 ; Hathout & Namer, 2022).

Références

Angleraud, N. (2023). *Classification sémantique des entrées du Wiktionnaire*. Rapport de stage de master. Université Paris Cité.

Apresjan, J. D. (1974). Regular polysemy. *Linguistics*, *14(2),* 5-32.

Asher, N., & Pustejovsky J. (2006). A type composition logic for generative lexicon. *Journal of Cognitive Science*, *6*(1), 1-38.

Barque, L., Haas, P., Huyghe, R., Tribout, D., Candito, M., Crabbé, B., & Segonne, V. (2020). FrSemCor: Annotating a French corpus with supersenses. *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 5912-5918.

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

Bonami, O., & Tribout, D. (2021). Échantinom: a hand-annotated morphological lexicon of French nouns. In F. Namer, N. Hathout, S. Lignon, M. Ševčíková, & Z. Žabokrtský (Eds), *Proceedings of the 3rd International Workshop on Resources and Tools for Derivational Morphology*, 42-51.

Ciaramita, M., & Johnson, M. (2003). Supersense tagging of unknown nouns in WordNet. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 168-175.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press.

Cruse, D. A. (1995). Polysemy and related phenomena from a cognitive linguistic viewpoint. In P. Saint-Dizier, & E. Viegas (Eds), *Computational Lexical Semantics* (pp. 33–49). Cambridge University Press.

Fellbaum C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

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 (Eds), *Competition in Inflection and Word-Formation* (pp. 67-93). Springer International Publishing.

Fradin, B. (2021). Caractériser les paradigmes dérivationnels. *Verbum*, *XLIII*(1), 149-178.

Godard, D., & Jayez, J. (1996). Types nominaux et anaphores : le cas des objets et des événements. *Cahiers Chronos*, *1,* 41-58.

Goudet, L., Paveau, M.-A., & Ruchon, C. (2018). Zoo-anthroponymes. Quand l'animal est le nom de l'humain, *Realista*, https://realista.hypotheses.org/1581

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

Huguin, M., Barque, L., Haas, P., Namer, F., & Tribout, D. (2022). *Guide d'annotation sémantique des noms construits*, Projet ANR Demonext, https://hal.archives-ouvertes.fr/hal-03638962

Jakubicek, M., Kilgarriff, A., Kovar, V., Rychly, P., & Suchomel, V. (2013). The TenTen corpus family. *Proceedings of the 7th International Corpus Linguistics Conference*, Lancaster University, 125-127.

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

Lignon, S., Namer, F., & Villoing, F. (2014). De l'agglutination à la triangulation ou comment expliquer certaines séries morphologiques. In F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschaefer, & S. Prévost (Eds), *SHS Web of Conferences, 8*, 1813-1835. EDP Sciences. https://doi.org/10.1051/shsconf/20140801324

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

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

Roché, M. (2011). Pression lexicale et contraintes phonologiques dans la dérivation en *-aie* du français. *Linguistica*, *51*(1), 5-22. https://doi.org/10.4312/linguistica.51.1.5-22

Salvadori, J., & Huyghe R. (2022). When morphology meets regular polysemy. *Lexique*, *31*, 85-113.

Salvadori, J., & Huyghe, R. (2023). Affix polyfunctionality in French deverbal nominalizations. *Morphology*, *33*, 1-39. https://doi.org/10.1007/s11525-022-09401-4

Sanacore, D., Hathout, N., & Namer, F. (2020). Représentation sémantique des familles dérivationnelles au moyen de frames morphosémantiques. In *Actes de TALN 2020*, 342-350. https://hal.science/hal-02784784v2

Thuilier, J., Tribout, D., & Wauquier, M. (2023). Affix rivalry in French demonym formation: The role of linguistic and non-linguistic parameters. *Word Structure*, *16*(1), 114-145. https://doi.org/10.3366/word.2023.0223