Étude des manifestations de la relation de méronymie dans une ressource distributionnelle

François Morlane-Hondère Cécile Fabre
CLLE-ERSS, Université de Toulouse - Le Mirail, 5, allées Antonio Machado - Toulouse Cedex 9
francois.morlane@univ-tlse2.fr, cecile.fabre@univ-tlse2.fr

RÉSUMÉ

Cette étude vise à étudier les manifestations de la relation de méronymie dans une ressource lexicale générée automatiquement à partir d'un corpus de langue générale. La démarche que nous adoptons consiste à recueillir un jeu de couples de méronymes issus d'une ressource externe que nous croisons avec une base distributionnelle calculée à partir d'un corpus de textes encyclopédiques. Une annotation sémantique des mots qui entrent dans ces couples de méronymes montre que la prise en compte de la nature sémantique des mots composant les couples de méronymes permet de mettre au jour des inégalités au niveau du repérage de la relation par la méthode d'analyse distributionnelle.

ABSTRACT

Study of meronymy in a distribution-based lexical resource

In this paper, we study the way meronymy behaves in a distribution-based lexical resource. We address the question of the evaluation of such resources through a semantic-based approach. Our method consists in collecting meronyms from a resource which we cross with a distribution-based lexical resource made from an encyclopedic corpus. Meronyms are then sub-categorized manually : firstly following the sub-relation they bear (STUFF/OBJECT, MEMBER/COLLECTION, etc.), then following the semantic class of their members. Results show that distributional analysis identifies meronymic relations in different proportions according to the semantic classes of the words involved in the meronymic pairs.

MOTS-CLÉS : analyse distributionnelle, sémantique lexicale, méronymie, évaluation.

KEYWORDS: distributional analysis, lexical semantics, meronymy, evaluation.

1 Introduction

Cette étude s'inscrit dans la problématique générale de l'évaluation des méthodes d'analyse distributionnelle (dorénavant *AD*) pour l'acquisition d'informations sémantiques (Baroni et Lenci, 2010). Ces méthodes de sémantique distributionnelle consistent à mesurer le degré de proximité sémantique entre mots sur la base du recouvrement de leurs contextes syntaxiques. La qualité des résultats fournis par ces méthodes s'avère difficile à mesurer du fait de la grande quantité de couples de mots générée par l'analyse et de la diversité des relations mises au jour (Sahlgren, 2006). Le typage des relations sémantiques calculées par l'AD est donc un enjeu pour optimiser l'utilisation des ressources distributionnelles dans des applications de TAL (van der Plas, 2008). Différents travaux ont cherché à mesurer l'efficacité des méthodes distributionnelles pour le repérage des relations lexicales en employant des méthodes impliquant des ressources de

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 169–182,
Grenoble, 4 au 8 juin 2012. ©2012 ATALA & AFCP

référence (Lin, 1998; Turney, 2008; Baroni et Lenci, 2011). En particulier, Baroni et Lenci ont soumis les résultats du calcul distributionnel à un large éventail de tâches sémantiques. Pour le français, les études ont été principalement consacrées au repérage de la synonymie (Bourigault et Galy, 2005; Ferret, 2010; Muller et Langlais, 2011), en particulier du fait de la disponibilité de lexiques de synonymes. Dans une étude précédente, nous nous sommes intéressés au cas de l'antonymie (Morlane-Hondère et Fabre, 2010). Tous ces travaux confirment la grande diversité des relations que peut détecter l'AD, et montrent aussi la nécessité de mieux comprendre sous quelles conditions le critère distributionnel opère, autrement dit, quels types d'informations distributionnelles doivent être pris en compte selon la nature de la tâche sémantique que l'on veut réaliser.

Dans cet article, nous nous focalisons sur le cas d'une relation lexicale particulière, la relation de méronymie (ou relation partie/tout), pour étudier la façon dont elle est repérée par un programme d'analyse distributionnelle. La relation de méronymie est intéressante à plusieurs titres : tout d'abord, elle constitue l'une des relations visées par les méthodes d'acquisition de ressources lexicales et terminologiques, au même titre que les relations plus souvent étudiées que sont l'hyperonymie et la synonymie. Ensuite, elle a la particularité de recouvrir un ensemble varié de relations (LIEU/ZONE, CONSTITUANT/OBJET, ÉTAPE/ACTIVITÉ, MEMBRE/COLLECTION, etc.), ce qui offre un terrain d'observation particulièrement riche pour étudier les modalités d'application de l'AD. Enfin, contrairement à la synonymie et à l'hyperonymie, la méronymie ne relie pas des mots relevant systématiquement de la même classe sémantique : c'est le cas par exemple du couple de mots *tête* et *enfant*, le premier étant une partie du corps, le second un être humain. Un tel cas de figure semble *a priori* défavorable au repérage par l'AD. C'est un des points que nous cherchons à examiner dans cet article.

La démarche que nous avons adoptée s'appuie sur un jeu de couples de méronymes issu du réseau JeuxDeMots (désormais *JDM*) (Lafourcade, 2007). Nous avons croisé ces données avec une base distributionnelle construite à partir d'un corpus issu de l'encyclopédie en ligne Wikipédia. Après une présentation de la relation de méronymie et des sous-relations qui la composent (2), nous décrivons les deux ressources que nous avons utilisées (3). Nous présentons ensuite la phase d'annotation, qui a donné lieu à deux procédures successives (4). Dans la section consacrée aux résultats (5), les couples ainsi annotés sont analysés du point de vue distributionnel, ce qui nous permet de dégager des classes de relations selon leur propension à être détectées par l'AD, et d'analyser ces différences.

2 La relation de méronymie : définition et typologie

La relation de méronymie est la relation qui s'établit entre une *partie* et son *tout*. Elle est asymétrique et sa réciproque, la relation entre un tout et l'une de ses parties, est l'holonymie. C'est une relation qui opère principalement entre deux noms, bien que Winston *et al.* (1987) proposent une relation FEATURE/ACTIVITY pour les couples désignant une étape dans un processus comme *paying/shopping*. La définition que donne Cruse (1986) de la méronymie est la suivante "X is a meronym of Y if and only if sentences of the form *A Y has Xs / an X* and *An X is a part of a Y* are normal when the noun phrases *an X, a Y* are interpreted generically." Ainsi, *La main est une partie du bras* est vrai, et ce même s'il existe des bras dont la main a été coupée.

La relation de méronymie est prise en compte dans la construction de thésaurus et d'ontologies (Van Campenhoudt, 1996; Keet et Artale, 2008). Elle se décline en plusieurs sous-relations.

Winston *et al.* (1987) définissent six sous-types de méronymes en s'appuyant sur trois critères :
– la *fonctionnalité* : la partie a-t-elle une fonction vis-à-vis du tout ? Par exemple, *poignée* est fonctionnel vis-à-vis de *porte*, mais pas vis-à-vis de *maison*.
– l'*homéomérité* : la partie et le tout sont-ils matériellement identiques (comme *tranche/gâteau*, contrairement à *arbre/forêt*) ?
– la *séparabilité* : la partie et le tout sont-ils séparables ? C'est le cas de *anse* et *tasse*, mais pas d'*acier* et *vélo*.

Relation	Exemple	Critères		
		Fonct.	Homéo.	Sépar.
ÉLÉMENT/OBJET	anse/tasse	+	−	+
MEMBRE/COLLECTION	arbre/forêt	−	−	+
PORTION/MASSE	tranche/gâteau	−	+	+
CONSTITUANT/OBJET	acier/vélo	−	−	−
ÉTAPE/ACTIVITÉ	payer/magasiner	+	−	−
LIEU/ZONE	oasis/désert	−	+	−

TABLE 1 – Sous-types de la relation de méronymie définis par Winston *et al.* (1987).

La combinaison de ces trois critères leur permet de dégager les relations rapportées au tableau 1. En marge de cette première série, Winston *et al.* (1987) décrivent des relations qui s'apparentent à de la méronymie sans en être tout à fait. Ces relations sont les suivantes :
– l'inclusion topologique : l'holonyme est un contenant (*prisonnier/cellule*), une zone (*Berlin Ouest/Allemagne de l'Est*) ou exprime une durée temporelle (*réunion/matin*).
– l'inclusion de classe : il s'agit ici de la relation d'hyponymie (*rose/fleur*, *peur/émotion*, etc.).
– la relation d'attribution : il s'agit d'une relation de type modifieur entre un mot et un adjectif (*tour/haute*, *blague/drôle*, etc.).
– la relation d'attachement : elle porte sur deux objets attachés l'un à l'autre (*boucle d'oreille* et *oreille*).
– la relation d'appartenance : elle relie des mots comme *millionnaire* et *argent* ou *auteur* et *copyright* et peut être confondue avec la méronymie à cause de l'ambiguïté du patron *X a Y*, qui peut exprimer l'appartenance (*Camille a un vélo* vs. *Un vélo a des roues*).

La différence entre certaines de ces relations et la méronymie *stricto sensu* est parfois assez fine. Nous verrons à la section 4 que beaucoup des paires annotées relèvent de l'une ou l'autre de ces relations pseudo-méronymiques.

3 Description des données

Cette étude repose sur la confrontation de deux types de données : les Voisins de Wikipédia, qui est une base lexicale générée automatiquement par analyse distributionnelle à partir de corpus, et une ressource lexicale construite de manière collaborative, JeuxDeMots.

3.1 Les voisins de Wikipédia

La base distributionnelle utilisée dans cette étude a été calculée à partir d'un corpus constitué de l'intégralité des articles de l'encyclopédie en ligne Wikipédia dans une version datant d'avril 2007. Ce corpus compte environ 194 millions de mots. L'analyse syntaxique du corpus a été

effectuée par le programme Syntex (Bourigault, 2007) à partir d'une version du corpus Wikipédia précédemment étiquetée morpho-syntaxiquement par TreeTagger. L'analyse distributionnelle a été réalisée par l'outil Upery développé par Didier Bourigault (2002) [1]. À partir des relations de dépendance syntaxique calculées par Syntex, le programme Upery extrait dans un premier temps des triplets de structure (mot1, RELATION, mot2). Les relations syntaxiques prises en compte sont les relations sujet, objet, complément prépositionnel, modification adjectivale. Les mots sont des unités simples ou des syntagmes, sous une forme lemmatisée, par exemple : *utiliser*, OBJET, *voiture*. Ces triplets servent de base au calcul distributionnel, qui rapproche les couples d'éléments qui partagent les mêmes contextes syntaxiques. Ces éléments sont de deux types : prédicats ou arguments. L'argument correspond au mot régi par la relation (ex : *voiture*). Le prédicat résulte de l'association du mot recteur et de la relation (ex : *utiliser*_OBJ). Upery rapproche donc les prédicats qui partagent les mêmes arguments, ainsi que les arguments qui partagent les mêmes prédicats. Par exemple, *voiture* est rapproché de *véhicule* parce que ces deux mots partagent, en position argument, les contextes suivants : *loueur*_DE, *pneu*_DE, *garer*_OBJ, *percuter*_SUJ, etc. La mesure de similarité utilisée est basée sur l'indice de Lin (1998). Le score de similarité de deux prédicats ou arguments varie – de 0 à 1 – en fonction de plusieurs facteurs : le nombre de contextes partagés, le nombre de triplets différents dans lesquels chacun de deux mots apparaît (indice de productivité), le degré de spécificité du contexte qui permet d'effectuer le rapprochement (se reporter à (Bourigault, 2002) pour les détails de la procédure de calcul). La base de voisins distributionnels de Wikipédia compte 2 441 118 paires de mots.

3.2 JeuxDeMots

Le jeu de couples que nous avons utilisé est issu de la base JeuxDeMots, qui est un réseau lexical enrichi de façon collaborative (Lafourcade, 2007) : des utilisateurs – experts et non-experts – se connectent à une interface en ligne [2] et ont pour tâche de proposer un ensemble de mots pour une relation et un mot-cible donnés. Les réponses communes à deux joueurs sont ajoutées au réseau, et si le lien était déjà présent, alors il est renforcé selon un système de pondération. Les relations proposées par le jeu incluent aussi bien des relations lexicales classiques que des relations moins usuelles (CHOSE/LIEU, ACTION/INSTRUMENT, etc.). Nous nous sommes tournés vers cette ressource car elle est librement accessible et est une des rares en français à inclure des couples portant la relation partie-tout.

Nous avons donc récupéré de JDM (dans sa version du 10/05/2011) les couples de noms entretenant une relation de méronymie ou d'holonymie. Ces derniers ont été produits par des joueurs auxquels il était demandé de "Donner des TOUT/PARTIES" des mots-cibles qui leur étaient proposés. Cette consigne étant relativement floue, nous verrons que les couples produits se trouvent souvent à la frontière de ce que l'on considère comme de la méronymie au sens strict (section 4.1).

3.3 Croisement des deux ressources

Nous avons croisé la base de méronymes avec les voisins distributionnels afin de repérer les paires de méronymes qui ont été repérées par l'AD. Pour cela, nous avons dans un premier temps éliminé de la base de méronymes les paires dont un mot au moins était absent du vocabulaire

1. La constitution du corpus et l'application de Syntex et Upery à ce corpus ont été réalisées par Franck Sajous, qui en a également assuré la mise en ligne : `http://redac.univ-tlse2.fr/voisinsdewikipedia/`.
2. `http://www.jeuxdemots.org/jdm-accueil.php`

des voisins. Nous avons ensuite symétrisé les couples de méronymes : pour tout couple A/B où A est méronyme de B nous avons généré un couple B/A où B est holonyme de A. La base obtenue compte 15 912 couples dont 34 % (5380) sont présents parmi les voisins. Ce taux de recouvrement est comparable à celui que nous avons pu observer pour d'autres relations (synonymie, antonymie).

Dans un deuxième temps, nous avons calculé le rapport de productivité (cf. section 3.1) entre les deux membres de chaque couple contenu dans la base de méronymes afin de ne conserver que les couples dont les deux membres ont des productivités comparables. En effet, de nombreux couples de mots ne sont pas repérés par l'AD parce que leurs productivités sont trop déséquilibrées. Cette étape vise donc à atténuer les effets liés au calcul des voisins en ne conservant que les couples qui sont potentiellement repérables par l'AD. Le seuil du rapport de productivité a été fixé à 0,60, ce qui signifie que, dans un couple, un mot ne pourra pas avoir une productivité 40 % plus élevée ou plus basse que l'autre mot. La base obtenue (désormais $\text{JDM}_{méro}$) compte 1520 paires dont 55 % (829) sont captées par l'AD.

4 Phase d'annotation

La phase d'annotation doit permettre de prendre en compte parmi les couples de $\text{JDM}_{méro}$ la diversité des sous-relations qu'inclut la méronymie. Nous nous sommes appuyés dans un premier temps sur la typologie de Winston *et al.* (1987), décrite à la section 2.

4.1 Typologie de Winston *et al.* (1987)

Cette annotation étant entièrement manuelle, nous n'avons dans un premier temps annoté qu'une partie de la base, soit 481 couples de $\text{JDM}_{méro}$ en prenant pour critère un seuil de productivité supérieur ou égal à 0,85. La répartition des paires dans les catégories a été rapportée au tableau 2.

Les relations décrites comme méronymiques dans la typologie figurent dans la partie haute du tableau, les relations pseudo-méronymiques apparaissent en bas. On peut constater que c'est la relation ÉLÉMENT/OBJET qui prévaut (elle englobe plus d'un tiers de l'ensemble des couples). Elle correspond à un vaste éventail de cas où la partie a un rôle fonctionnel vis-à-vis de son tout : *pince/crabe*, *clavier/piano*, Les autres relations sont nettement minoritaires. Dans la partie basse du tableau, les relations pseudo-méronymiques représentent 44 % des couples annotés. La dernière relation, identifiée dans le tableau par un point d'interrogation, regroupe tous les couples qui, selon nous, portent une relation autre que toutes celles qui ont été identifiées par (Winston *et al.*, 1987) : *chemin/voyage*, *corps/femme*, *électricité/fil*, *activité/temps*, etc. Ces couples sont au nombre de 151, ce qui représente 31,4 % de l'ensemble. Certains d'entre eux passent le test de l'insertion dans un patron de type *X a Y* (*chauffeur/taxi*, *billet/montant*, *carte/couleur*). Beaucoup des paires appartenant à cette catégorie sont constituées de noms exprimant des concepts abstraits pour lesquels les critères de fonctionnalité, d'homéomérité et de séparabilité sont difficilement applicables comme dans *calcul/chiffre* ou *lumière/univers*. Malgré le caractère périphérique d'une partie des relations, nous prenons comme objet d'étude le jeu de couples dans son ensemble, dans la mesure où ils ont été produits par des locuteurs qui les ont perçus comme relevant de la relation partie/tout.

Le bilan que l'on peut tirer de cette première annotation est que la typologie montre ses limites

	Relation	Fréq.	Proportion
Relations méronymiques	ÉLÉMENT/OBJET	177	36,8 %
	CONSTITUANT/OBJET	35	7,3 %
	LIEU/ZONE	34	7,1 %
	MEMBRE/COLLECTION	14	2,9 %
	ÉTAPE/ACTIVITÉ	6	1,2 %
	PORTION/MASSE	1	0,2 %
Autres relations	?	151	31,4 %
	INCLUSION TOPOLOGIQUE	31	6,4 %
	CLASSE	18	3,7 %
	SYNONYMIE	11	2,3 %
	APPARTENANCE	3	0,6 %

TABLE 2 – Résultats de l'annotation basée sur la typologie de Winston *et al.* (1987).

lorsqu'elle est confrontée aux données de JDM, pour deux raisons :
– une seule relation – ÉLÉMENT/OBJET – concentre près de 66 % des couples considérés comme relevant strictement de la méronymie. Cette classe apparaît manifestement comme trop englobante dans la mesure où elle porte sur des couples de nature hétérogène.
– 31,4 % des couples ne relèvent pas d'une des relations définies dans la typologie de référence, même en l'augmentant avec la série des relations pseudo-méronymiques.
Nous avons donc décidé de délaisser une typologie préétablie et d'adopter une approche *bottom-up* : nous nous focalisons cette fois sur le sens des mots composant les paires de méronymes afin de faire émerger des combinaisons de classes sémantiques.

4.2 Annotation en classes sémantiques

La deuxième procédure d'annotation consiste à attribuer une classe sémantique à chaque mot des couples de méronymes, afin de mettre au jour de nouvelles configurations distributionnelles. Elle s'inspire du point de vue de Murphy (2003), qui rejette l'idée selon laquelle il existerait plusieurs déclinaisons de la méronymie et qui considère que la seule chose qui change entre les différents sous-types est la nature des mots sur lesquels porte la relation.

Les couples que nous avons utilisés sont ceux de la base $\text{JDM}_{méro}$ (section 3.2) : la méthode d'annotation étant cette fois semi-automatisée, nous avons utilisé un ensemble plus élevé de paires que dans la section précédente. Suite aux résultats obtenus lors de l'annotation effectuée à partir de la typologie de Winston *et al.* (1987), nous avons choisi de retirer manuellement les couples d'hyperonymes et de synonymes. $\text{JDM}_{méro}$ compte désormais 1334 paires dont 53 % (711) sont détectés par l'AD (contre 1520 paires dont 55 % de voisins dans sa version précédente).

En guise de classe sémantique, nous avons associé chaque mot à l'un de ses hyperonymes *de haut niveau* dans WordNet (Fellbaum, 1998). La raison pour laquelle nous avons utilisé cette ressource plutôt que les hyperonymes de JDM est que les relations d'hyperonymie y sont présentes de façon plus systématique que dans JDM (tous les mots de notre sous-ensemble n'ont pas forcément d'hyperonyme dans JDM). Cette démarche a, dans un premier temps, consisté à

traduire le lexique de JDM$_{méro}$ en anglais[3]. Nous avons ensuite associé chaque mot à l'ensemble des hyperonymes de sa traduction anglaise dans le réseau, et ce pour chacune de ses acceptions recensées dans WordNet. Ainsi, *église* est associé aux quatre chemins suivants :

ENTITY>ABSTRACT_ENTITY>ABSTRACTION>EVENT>HUMAN_ACTIVITY>ACTIVITY>CEREMONY

ENTITY>ABSTRACT_ENTITY>ABSTRACTION>SOCIAL_GROUP>ORGANIZATION>INSTITUTION>RELIGION

ENTITY>ABSTRACT_ENTITY>ABSTRACTION>SOCIAL_GROUP>GATHERING>BODY

ENTITY>PHYSICAL_ENTITY>PHYSICAL_OBJECT>WHOLE>ARTIFACT>STRUCTURE>EDIFICE>PLACE_OF_WORSHIP

L'étape suivante consiste à procéder à un *élagage* de l'arborescence. Par défaut, la granularité de WordNet est bien trop fine pour nous permettre d'obtenir des classes de taille satisfaisante (par exemple, dans nos données, *église* est le seul mot à figurer en position hyponyme de CEREMONY). L'élagage vise à obtenir une arborescence moins complexe. Ainsi, nous avons choisi de couper les noms abstraits (hyponymes de ABSTRACT_ENTITY) au troisième niveau de profondeur et les noms concrets (hyponymes de PHYSICAL_ENTITY) au cinquième. Ce choix se justifie par un nombre plus important de noms concrets dans nos données. Dans le cas de *église*, cela entraîne la disparition de la nuance entre les deux acceptions du mot en tant que groupe social :

ENTITY>ABSTRACT_ENTITY>ABSTRACTION>EVENT

ENTITY>ABSTRACT_ENTITY>ABSTRACTION>SOCIAL_GROUP

ENTITY>PHYSICAL_ENTITY>PHYSICAL_OBJECT>WHOLE>ARTIFACT>STRUCTURE

Les mots ainsi annotés sont ensuite désambiguïsés manuellement en fonction du mot avec lequel ils entretiennent une relation de méronymie. Dans l'exemple précédent, cette démarche consiste à associer *église* à un type de bâtiment (STRUCTURE) dans le couple *église/village* et à un groupe social (SOCIAL_GROUP) dans le couple *fidèle/église*. Toujours dans la même optique, nous avons procédé à différents ajustements consistant à opérer des regroupements entre certaines classes de mots. Cette étape s'est faite de façon empirique en fonction notamment du nombre d'éléments contenus dans chaque catégorie : par exemple, les éléments appartenant à des catégories de moins de 10 membres ont été systématiquement déplacés sous l'hyperonyme de niveau supérieur. Dans l'exemple ci-dessous, le premier chemin correspond à celui de *doigt*, le deuxième à celui de *nez* :

ENTITY>PHYSICAL_ENTITY>THING>PIECE>BODY_PART>EXTERNAL_BODY_PART>MEMBER>DIGIT

ENTITY>PHYSICAL_ENTITY>THING>PIECE>BODY_PART>ORGAN>SENSORY_RECEPTOR>CHEMORECEPTOR

Après regroupement, les deux mots se retrouvent au même niveau dans la hiérarchie : ils sont directement subordonnés à BODY_PART. La répartition finale des mots de JDM$_{méro}$ après désambiguïsation et élagage de la hiérarchie a été rapportée à la figure 1 pour les noms concrets et à la figure 2 pour les noms abstraits. Sur ces figures, on constate clairement que la profondeur de la hiérarchie varie selon les classes. La classe ABSTRACT_ENTITY n'a qu'un niveau de profondeur. Elle se divise en quatre classes : les événements (EVENT : *exposition*, *procès*), les groupes sociaux (SOCIAL_GROUP : *peuple*, *famille*), les collections (COLLECTION : *flotte*, *galaxie*) et *autres* (OTHER : *trou*, *valeur*). La catégorie *autres* est un ajout de notre part, elle contient des mots appartenant à des classes comme les jours de la semaine, les unités monétaires ou les notes de musique qui contiennent trop peu de membres pour avoir une existence autonome dans notre classification. La classe des noms concrets regroupe un nombre de noms beaucoup plus important que la classe des noms abstraits (2082 contre 460). Elle est structurée de façon plus complexe et possède trois niveaux de profondeur. Le premier niveau distingue les parties du corps (LIVING_PART, qui

3. Cette étape a été facilitée par l'utilisation de Google Traduction (`http://translate.google.fr/`). Les traductions ont ensuite été vérifiées manuellement. La trentaine de cas d'ambiguïtés liés à la traduction – *plat* traduit *flat* plutôt que *dish*, ou *car* traduit *because* au lieu de *bus* – ont été également désambiguïsés.

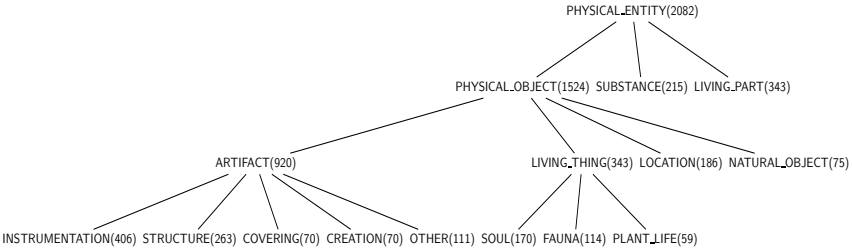

FIGURE 1 – Répartition des mots de JDM$_{méro}$ dans la classe PHYSICAL_ENTITY.

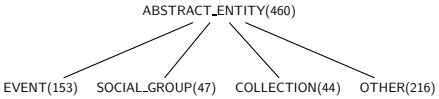

FIGURE 2 – Répartition des mots de JDM$_{méro}$ dans la classe ABSTRACT_ENTITY.

regroupe en fait les parties de corps humain, animal, et les parties de végétaux : *langue*, *patte*), les substances ou matières (SUBSTANCE : *fer*, *laine*), et les objets physiques. Cette dernière classe, la plus volumineuse, regroupe les objets naturels (NATURAL_OBJECT : *torrent*, *volcan*), les lieux (LOCATION : *grotte*, *quartier*), les entités vivantes (LIVING_THING) et les artéfacts (ARTIFACT). Ces deux dernières classes possèdent enfin un dernier niveau de profondeur. La classe des entités vivantes se subdivise en trois sous-classes regroupant les noms se rapportant à des humains (SOUL : *joueur*, *pompier*), des animaux (FAUNA : *canard*, *requin*) et des végétaux (PLANT_LIFE : *olive*, *rose*). La classe des artéfacts comprend les noms d'instruments (INSTRUMENT : *lampe*, *pneu*), de bâtiments (STRUCTURE : *lycée*, *magasin*), de vêtements (COVERING : *cape*, *chapeau*), les créations (CREATION : cette classe englobe les productions littéraires, artistiques comme *fresque* ou *roman*) et une catégorie *autres* (OTHER : de la même façon que la catégorie éponyme dans la classe des entités abstraites, elle regroupe des objets de nature hétérogène comme *brique* ou *savon*).

L'abandon d'une typologie préétablie au profit de classes sémantiques va nous permettre de mener des analyses plus précises : dans la section suivante, nous analysons les propriétés distributionnelles des couples de classes les plus fréquents dans notre jeu de méronymes.

5 Analyse des couples

À ce stade de l'étude, nous nous intéressons à deux propriétés des 1334 couples de la base JDM$_{méro}$:
– chaque couple a été catégorisé selon qu'il a été capté par l'AD ou non (section 3.2),
– chaque membre de chaque couple de méronymes est associé a une étiquette sémantique qui lui est propre (section 4.2).

classe méro.	classe holo.	nb. de couples	% voisins
SOUL	SOCIAL_GROUP	54	96,2
LOCATION	LOCATION	19	90,0
EVENT	EVENT	29	85,3
STRUCTURE	LOCATION	94	82,6
STRUCTURE	STRUCTURE	12	82,5
SOUL	STRUCTURE	25	66,7
INSTRUMENTATION	STRUCTURE	11	51,9
SUBSTANCE	SUBSTANCE	27	51,2
INSTRUMENTATION	INSTRUMENTATION	82	41,5
LIVING_PART	PLANT_LIFE	76	41,2
LIVING_PART	LIVING_PART	62	33,9
SUBSTANCE	INSTRUMENTATION	17	33,3
LIVING_PART	SOUL	54	18,4
LIVING_PART	FAUNA	41	14,6

TABLE 3 – Couples de classes qui englobent au moins 10 couples dans $\text{JDM}_{méro}$.

Nous croisons à présent ces deux aspects afin de mettre au jour des couples de classes de mots et d'expliquer pourquoi certaines sont mieux captées par l'AD que d'autres. Nous avons rapporté au tableau 3 les couples de classes représentés par au moins 10 paires de méronymes dans la base (nous avons supprimé la classe OTHER à cause du caractère hétérogène des relations qu'elle englobe). Ils sont classés par ordre de proportion de voisins décroissante. On peut constater qu'il y a de fortes disparités entre les différentes combinaisons : alors que 96,2 % des couples dont le méronyme est un humain et l'holonyme un groupe social sont repérés par l'AD, cela n'est vrai que de 14,6 % des couples dont le méronyme est une partie du corps et l'holonyme un animal. Dans cette section, nous nous focalisons sur l'observation des propriétés distributionnelles de ces différents types de classes afin d'expliquer pourquoi certaines sont plus compatibles que d'autres. Les différentes combinaisons de classes sont analysées en deux temps. La première section est consacrée à l'étude des couples constitués de deux mots appartenant à des classes identiques – couples dits *homogènes*. Les couples constitués de deux mots relevant de deux classes différentes – couples *hétérogènes* – sont analysés à la deuxième section. Ce découpage est motivé par le fait que, contrairement aux autres relations classiques, la méronymie possède la particularité de pouvoir associer deux mots qui ont des natures sémantiques différentes (*vache/troupeau*, *métal/épée*). Or, on sait que l'AD basée sur l'analyse des contextes syntaxiques présente une tendance à rapprocher des mots qui sont sémantiquement similaires. Les données dont nous disposons nous donnent la possibilité de mettre au jour les conditions dans lesquelles se principe se vérifie ou ne se vérifie pas.

5.1 Couples homogènes

Parmi les 14 types de couples rapportés au tableau 3, 6 sont homogènes. Leur proportion de voisins moyenne est de 64,1 %, ce qui est un peu plus élevé que celle des couples hétérogènes, qui est de 50,6 %.

5.1.1 Les classes les mieux repérées

Les couples composés de deux éléments appartenant aux classes LOCATION, EVENT ou STRUCTURE sont repérés par l'AD dans des proportions allant de 82,5 % à 90 %. Les couples dont les deux membres appartiennent à la classe LOCATION expriment une relation entre deux lieux, l'un étant localisé dans un second de taille supérieure (*Allemagne/Europe*, *place/village*). Ce sont les couples homogènes qui sont le mieux repérés par l'AD. Les mots qui les composent partagent la propriété d'exprimer des entités localisées spatialement. De fait, ils partagent des contextes comme la position objet de verbes de localisation – *(se) situer*, *se trouver* – *via* des prépositions complexes comme AU SUD DE ou AU CENTRE DE, ou encore la position complément du nom, quand le nom exprime un point cardinal (NORD DE, SUD DE, etc.). En plus de partager ce faisceau de contextes, les mots exprimant des lieux se distinguent par des contextes spécifiques qui permettent de distinguer des sous-classes le lieux. Par exemple, beaucoup des couples de lieux expriment différents niveaux de subdivisions administratives (*commune/canton*, *village/département*, etc.). Ils partagent des contextes comme *administration*_DE, *communauté*_DE, *population*_DE ou *territoire*_DE. De la même façon, l'analyse des contextes du couple *propriété/parc* montre qu'ils ont été rapprochés à la fois grâce au fait qu'ils sont des objets localisés dans l'espace (*limite*_DE, *s'étendre*_SUR, *superficie*_DE), mais aussi parce qu'il apparaissent comme des biens que l'on peut posséder (*revendre*_OBJ, *acheter*_OBJ, *gérer*_OBJ). Ces contextes spécifiques viennent renforcer la proximité distributionnelle entre les différents sous-ensembles de la classe des noms de lieux.

Le cas des couples d'événements est assez similaire si ce n'est que les mots expriment des valeurs temporelles et non plus spatiales : l'événement méronyme prend place dans un processus de plus grande ampleur exprimé par le second membre de la paire (*bataille/campagne*, *départ/course*, *victoire/combat*). Les noms exprimant une durée s'emploient dans des contextes comme *avoir lieu*_SUJ, *prendre fin*_SUJ et *se terminer*_SUJ, par l'intermédiaire de prépositions comme LORS DE, AU COURS DE, etc. Ici aussi, certains types d'événements se distinguent du fait, par exemple, que certains ont un aspect duratif alors que d'autres sont plus ponctuels (*mission* vs. *victoire*). Comme ça été le cas pour les noms de lieux, les contextes exprimant la localisation temporelle d'un événement sont associés à d'autres contextes exprimant des caractéristiques liées au sous-type d'événement.

Enfin, la classe STRUCTURE relie des noms de bâtiments ou de parties de bâtiments qui se situent au sein d'un autre bâtiment (*tour/château*, *hall/immeuble*, *salle/lycée*). Il semblerait que la classe des bâtiments ait une distribution moins bien circonscrite que celle des lieux et des événements. En effet, l'étude des paires de cette catégorie ne fait apparaître que peu de contextes transversaux, qui s'appliquent à l'ensemble des mots appartenant à la classe des bâtiments. Le contexte *construire*_OBJ en est un : il peut virtuellement s'appliquer à tout type de bâtiment mais ne permet pas, par exemple, de rapprocher la paire *salon/appartement*. De la même façon, le contexte *habiter*_OBJ est assez répandu mais ne s'applique, par définition, qu'aux structures destinées à être habitables (ce contexte n'apparaît pas dans les contextes communs de la paire de voisins *pièce/musée*, par exemple). Ainsi, la classe des bâtiments apparaît de façon assez floue, dans la mesure où l'emploi qui est fait des noms de bâtiments dans le corpus met l'accent sur leur aspect fonctionnel. Il semblerait que les classes qui émergent se situent à un niveau de granularité inférieur. Par exemple, les couples *appartement/immeuble* et *chambre/hôtel* possèdent en commun des contextes comme *habiter*_OBJ, *louer*_OBJ ou *se installer*_DANS. Ces contextes définissent un type de bâtiment bien particulier, à savoir les bâtiments destinés au logement. De la même façon, les couples *tour/château* et *fortification/fort* partagent des contextes comme

*protéger*_OBJ, *détruire*_OBJ ou *attaquer*_OBJ qui permettraient de dessiner les contours de la classe des bâtimens militaires.

Ainsi, les mots qui appartiennent à ces trois types de couples de classes sont particulièrement bien repérés par l'AD du fait que les propriétés sémantiques qu'ils partagent se répercutent sur le plan distributionnel. Nous avons vu que les distributions des couples de lieux et d'événements se caractérisaient par un ensemble de contextes compatibles avec la plupart des mots appartenant à chacune de ces classes. Ce constat se vérifie dans une moindre mesure sur la classe des bâtiments, pour laquelle les classes qui émergent se situent à un niveau plus fin.

5.1.2 Classes repérées en quantités moindres

Les couples composés de deux éléments appartenant aux classes SUBSTANCE, INSTRUMENTATION ou LIVING_PART sont captés par l'AD dans des proportions allant seulement de 33,9 % à 51,2 %. Nous avons donc affaire à des couples de mots qui, tout en possédant la même étiquette sémantique, se caractérisent par des propriétés distributionnelles différentes.

Dans le cas de la classe SUBSTANCE, les mots reliés désignent deux substances ou matières (au sens large) dont l'une entre dans la composition de l'autre : *carbone/diamant*, *crême/beurre*, *éthanol/rhum*. Nous identifions deux phénomènes expliquant la raison pour laquelle ces couples de mots sont mal repérés par l'AD. Le premier est que leurs membres ne sont pas forcément employés comme des substances dans le corpus. *Rhum*, par exemple, apparaît comme un produit fini et non pas comme un ingrédient (sauf dans le contexte *baba*_À). Le second est que, même dans les – rares – cas où les deux mots sont employés comme des composants, ils n'entrent pas forcément dans la composition du même type d'objets : pour le couple *carbone/diamant*, les contextes comme *collier*_DE sont incompatibles avec *carbone*. Un couple comme *crème/beurre* fait exception à la règle. *Crème* et *beurre* ont été détectés comme voisins, ils partagent les contextes *mélanger*_OBJ, *incorporer*_OBJ, *verser*_OBJ, etc. Ces deux mots ont en commun qu'ils apparaissent comme des ingrédients de cuisine. Dans le cas de *carbone/diamant*, les contextes se recoupent moins dans la mesure où on a affaire, d'un côté, à un élément chimique et, de l'autre, à un minéral. Cette différence sémantique semble suffisamment importante pour qu'elle soit perceptible au niveau de leurs distributions respectives et que ces deux mots ne soient donc pas repérés comme des voisins.

Les couples dont les deux membres appartiennent à la classe INSTRUMENTATION sont repérés par les voisins à hauteur de 41,5 %. La notion d'*instrument* est à prendre au sens large, et les couples appartenant à cette classe expriment une relation où un élément fait partie d'un dispositif ou un système de plus grande ampleur : *écran/ordinateur*, *pédale/bicyclette*, *pneu/autobus*. Dans la plupart des cas, les distributions entre les deux mots sont trop éloignées pour que l'analyse permette de les rapprocher. Par exemple, les contextes dans lesquels apparaît le méronyme *réservoir* (*volume*_DE, *servir*_DE, *placer*_OBJ) diffèrent complètement de ceux dans lesquels apparaissent ses holonymes *automobile* et *moto* (*accident*_DE, *conduire*_OBJ, *modèle*_DE). Le cas du méronyme *moteur*, en revanche, illustre une situation où la distribution du méronyme et de l'holonyme se recoupent : les 7 paires dans lesquelles il prend place sont toutes repérées par les voisins. Il apparaît en position méronyme de *avion*, *bateau*, *machine*, *navire*, *train*, *véhicule* et *voiture*. L'analyse des contextes communs fait apparaître une certaine symbiose entre le moteur et la machine qu'il équipe, dans la mesure où ils partagent un éventail de contextes relativement étendu comme *panne*_DE, *bruit*_DE, *consommer_suj*, *puissance*_DE, *se arrêter*_SUJ, *fonctionner*_SUJ, etc. On pourrait analyser certains de ces contextes comme des cas de métonymie : le bruit produit

par l'avion est en fait le bruit du moteur, de même que la puissance de la voiture est celle de son moteur.

Les couples de mots appartenant tous deux à la classe LIVING_PART sont les couples homogènes les moins bien identifiés par l'AD. Ils relient deux parties du corps (corps humain, animal ou partie d'un végétal), dont l'une est elle-même une partie de l'autre : *chair/doigt*, *muscle/bras*, *peau/visage*. Une des raisons expliquant les différences de distribution parmi les parties du corps est que, dans la plupart des cas, l'on affaire à des sous-classes de parties du corps dont les fonctionnements diffèrent radicalement. Ainsi, le fait que les couples *nerf/jambe* ou *nerf/doigt* ne sont pas captés s'explique par le fait que *jambe* et *doigt* sont des membres du corps. Ils peuvent par conséquent apparaître en position objet de verbes comme *lever*, *croiser* ou *replier*, soit autant de contextes dans lesquels ne peut pas apparaître *nerf*.

Ainsi, nous pouvons conclure de l'analyse de ces trois types de couples que les catégories *substance*, *instrumentation* et *living_part* s'avèrent peu pertinentes du point de vue distributionnel. Elle sont constituées de mots dont les distributions sont particulièrement dissemblables. Ainsi, si on postule *a priori* l'existence d'une classe sémantique des parties du corps, l'analyse du corpus montre que les mots *jambe*, *bras*, *doigt*, etc. entrent en fait dans un paradigme différent de celui de *veine*, *nerf* ou *os*. Il y a donc un décalage entre les classes sémantiques que l'on pourrait dégager intuitivement et les classes distributionnelles qui émergent de l'analyse du texte.

5.2 Couples hétérogènes

Nous avons auparavant évoqué la tendance qu'a l'AD à faire émerger des rapprochements relevant de la similarité sémantique, c'est-à-dire des mots qui sont "le même genre de choses" (van der Plas, 2008). De ce fait, on pouvait s'attendre à ne pas trouver de couples hétérogènes parmi les voisins distributionnels. Les résultats montrent toutefois que certains couples de catégories différentes sont quasi-intégralement repérés par l'AD.

C'est notamment le cas des couples de mots dont le méronyme appartient à la classe des humains (SOUL) et l'holonyme à celle des groupes sociaux : *capitaine/marine*, *fils/famille*, *musicien/orchestre*. Ces couples sont repérés à 96,2 %. Cela s'explique par le fait que les mots de la classe SOCIAL_GROUP ont des distributions similaires à ceux de la classe SOUL. Ils partagent par exemple la propriété d'apparaître en position sujet des verbes d'actions. Ainsi, le couple *directeur/entreprise* a été rapproché sur la base de contextes comme *détenir*_SUJ, *conseiller*_OBJ ou *affirmer*_SUJ, qui sont clairement destinés à être employés avec des animés. Il en va de même pour *joueur* et *équipe*, qui ont été rapprochés *via* les contextes *se entraîner*_SUJ, *affronter*_SUJ ou *se qualifier*_SUJ. Nous avons ici aussi affaire à un fonctionnement de type métonymique, dans la mesure où l'ensemble est employé pour désigner les membres.

Les couples dont le méronyme est un bâtiment et l'holonyme un lieu – *château/canton*, *école/commune*, *immeuble/métropole* – sont également bien captés par l'AD (c'est le cas de 82,6 % d'entre eux). Cela peut s'expliquer par l'ambiguïté des noms de bâtiments, qui peuvent aussi bien être employés comme des noms de lieux. Ainsi, le recouvrement entre ces deux classes implique une certaine similarité au niveau des distributions de leurs membres.

À l'autre extrémité du spectre, on remarque que la catégorie LIVING_PART apparaît en position méronyme dans trois des configurations hétérogènes les moins bien repérées par l'AD. Cette classe est successivement associée à PLANT_LIFE (*pétale/marguerite*, *tige/rose*, *tronc/chêne*), SOUL (*bras/citoyen*, *doigt/bébé*, *main/professeur*) et FAUNA (*bec/canard*, *patte/chat*, *queue/loup*). Dans

les trois cas, le fait que les couples relevant de ces classes ne soient que peu repérés s'explique par le fait qu'ici, les *touts* sont des êtres animés, contrairement à leurs parties. La conséquence en est que leurs propriétés distributionnelles sont radicalement opposées à celles de leurs méronymes. Cela semble être un peu moins flagrant pour les végétaux (ce qui explique que les couples LIVING_PART/PLANT_LIFE sont mieux repérés que les couples LIVING_PART/SOUL et LIVING_PART/FAUNA). On est donc dans le cas attendu de mots relevant de sens différents et par conséquent dissemblables sur le plan distributionnel.

5.3 Conclusion

Dans le cadre de cette étude consacrée à l'acquisition de relations sémantiques par des techniques d'analyse distributionnelle, nous nous sommes concentrés sur le cas de la relation de méronymie. Nous avons adopté une méthode d'évaluation qualitative reposant sur l'annotation sémantique de couples de méronymes. Sur le plan méthodologique, cette étude a montré que la typologie habituellement utilisée pour décrire les différents types de relations méronymiques était peu adaptée pour catégoriser nos données. Une approche consistant à typer sémantiquement les couples de méronymes permet de mieux rendre compte de la diversité des relations qu'ils expriment. Sur le plan des résultats, nous avons montré que si la méronymie, considérée globalement, est repérée dans des proportions comparables à d'autres relations (environ 1/3 des méronymes de JDM sont détectés par le programme d'AD que nous avons utilisé), elle n'est pas repérée par l'AD de manière homogène : la nature sémantique des mots qui entrent dans la relation de méronymie constitue un facteur décisif pour leur détection par l'AD. Tout d'abord, nous avons constaté que l'AD privilégie le repérage des couples de méronymes dont les membres relèvent de la même classe sémantique. Ensuite, nous avons vu que certaines configurations étaient identifiées dans des proportions beaucoup plus fortes que d'autres. C'est le cas des paires associant deux lieux, deux événements ou deux structures, ou associant un humain et un groupe social ou un lieu et un bâtiment. D'autres relations méronymiques, comme celles impliquant les parties du corps, sont mal détectées par l'AD car elles mettent en jeu des termes qui ne fonctionnent pas de la même manière sur le plan distributionnel. Cette étude contribue donc à préciser les conditions d'application du critère distributionnel au repérage d'une relation sémantique donnée. À ce stade, elle laisse cependant ouverte la question de l'influence du corpus de test sur la prédominance de certaines configurations distributionnelles des résultats.

Références

BARONI, M. et LENCI, A. (2010). Distributional memory : A general framework for corpus-based semantics. *Computational Linguistics*, 36.

BARONI, M. et LENCI, A. (2011). How we BLESSed distributional semantic evaluation. *GEMS 2011*, pages 1–10.

BOURIGAULT, D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *In Actes de la 9econférence sur le Traitement Automatique de la Langue Naturelle*, pages 75–84, Nancy.

BOURIGAULT, D. (2007). *Un analyseur syntaxique opérationnel :* SYNTEX. Mémoire d'habilitation à diriger des recherches. Université Toulouse II – Le Mirail.

BOURIGAULT, D. et GALY, E. (2005). Analyse distributionnelle de corpus de langue générale et synonymie. *In 4esJournées de la linguistique de corpus*, Lorient.

CRUSE, D. A. (1986). *Lexical semantics*. Cambridge University Press.

FELLBAUM, C., éditeur (1998). *WordNet : an electronic lexical database*. MIT Press, Cambridge.

FERRET, O. (2010). Similarité sémantique et extraction de synonymes à partir de corpus. *In Actes de TALN 2010 Traitement Automatique des Langues Naturelles - TALN 2010*.

KEET, C. et ARTALE, A. (2008). Representing and reasoning over a taxonomy of part–whole relations. *Applied Ontology*, 3(1):91–110.

LAFOURCADE, M. (2007). Making people play for lexical acquisition. *In 7th Symposium on natural Language Processing*.

LIN, D. (1998). An information-theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304.

MORLANE-HONDÈRE, F. et FABRE, C. (2010). L'antonymie observée avec des méthodes de TAL : une relation à la fois syntagmatique et paradigmatique ? *In Actes de TALN 2010 Traitement Automatique des Langues Naturelles - TALN 2010*.

MULLER, P. et LANGLAIS, P. (2011). Comparaison d'une approche miroir et d'une approche distributionnelle pour l'extraction de mots sémantiquement reliés. *In Traitement Automatique des Langues Naturelles (TALN), Montpellier*, volume 1, pages 235–246.

MURPHY, M. L. (2003). *Semantic Relations and the Lexicon*. Cambridge University Press, New York.

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

TURNEY, P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. *In COLING*, pages 905–912.

VAN CAMPENHOUDT, M. (1996). Recherche d'équivalences et structuration des réseaux notionnels : le cas des relations méronymiques. *Terminology*, 3(1):53–83.

van der PLAS, L. (2008). *Automatic lexico-semantic acquisition for question answering*. Thèse de doctorat, Université de Groningen (Pays-bas).

WINSTON, M. E., CHAFFIN, R. et HERRMANN, D. (1987). A taxonomy of part-whole relations. *Cognitive Science*, 11(4):417–444.