Extraction automatique de relations sémantiques dans les définitions : approche hybride, construction d'un corpus de relations sémantiques pour le français

Emmanuel Cartier
Université Paris 13 Sorbonne Paris Cité, LIPN UMR 7030, équipe RCLN
emmanuel.cartier@lipn.univ-paris13.fr

Résumé. Cet article présente une expérimentation visant à construire une ressource sémantique pour le français contemporain à partir d'un corpus d'environ un million de définitions tirées de deux ressources lexicographiques (Trésor de la Langue Française, Wiktionary) et d'une ressource encyclopédique (Wikipedia). L'objectif est d'extraire automatiquement dans les définitions différentes relations sémantiques : hyperonymie, synonymie, méronymie, autres relations sémantiques. La méthode suivie combine la précision des patrons lexico-syntaxiques et le rappel des méthodes statistiques, ainsi qu'un traitement inédit de canonisation et de décomposition des énoncés. Après avoir présenté les différentes approches et réalisations existantes, nous détaillons l'architecture du système et présentons les résultats : environ 900 000 relations d'hyperonymie et près de 100 000 relations de synonymie, avec un taux de précision supérieur à 90% sur un échantillon aléatoire de 500 relations. Plus de 2 millions de prédications définitoires ont également été extraites.

Abstract.
Automatic Extraction of Semantic Relations from Definitions : en experiment in French with an Hybrid Approach
This article presents an experiment to extract semantic relations from definitions. It is based on approximately one million definitions from two general dictionaries (Trésor de la Langue Française, French Wiktionary) and from the collaborative Wikipedia. We aim at extracting from these data several semantic relations : hyperonymy, synonymy, meronymy and other semantic relations. The methodological approach combines the precision of lexico-syntactic patterns and the recall of statistical analysis. After a survey of the state-of-the-art methods in this area, we detail our system and give the overall outcomes : about 900 000 hypernymy and 100 000 synonymy relations are extracted with a precision above 90% on a sample of 500 pairs for each relation. About 2 millions of definitory predicates are also extracted.

Mots-clés : relations sémantiques, patrons lexico-syntaxiques, distributionnalisme, prédication, hyperonymie, synonymie, méronymie, définition.
Keywords: semantic relations, lexico-syntactic patterns, distributionnalism, predication, hypernymy, synonymy, meronymy, definition.

1 Introduction

Aujourd'hui, nous pourrions dresser le tableau suivant des recherches de TAL en sémantique : depuis une dizaine d'années, la disponibilité de corpus de plus en plus imposants a permis aux approches distributionnelles de gagner du terrain dans différents domaines, mais sans vision globale de leur potentiel et de leurs limites ; les approches par apprentissage automatique, qui ont pris le pas sur les approches symboliques, ne permettent pas d'en induire un modèle sémantique unifié et sont extrêmement hermétiques à toute interprétation théorique. Les approches symboliques, enfin, ont réduit leurs prétentions initiales pour focaliser sur des moyens d'expressions spécifiques - spécialement les patrons lexico-syntaxiques liés à telle ou telle information linguistique, avec un certain succès ; par ailleurs, différents modèles et réalisations de la structuration sémantique du lexique ont émergé, autour de WordNet, de FrameNet et de ses dérivés, du modèle sens-texte, de la théorie des qualia ou encore autour des projets lexico-encyclopédiques de type BabelNet ou NELL.

Cet article détaille une expérience visant à repérer automatiquement dans un corpus de définitions tirées de deux dictionnaires et d'une encyclopédie collaborative, différentes relations sémantiques. Pour ce faire, nous utiliserons une méthode hybrique guidée par une analyse fréquencielle de patrons dénotant l'hyperonymie, l'hyponymie, la méronymie, la synonymie et d'autres relations sémantiques.

En dehors de proposer un corpus de relations sémantiques du français contemporain, cet article cherche aussi à s'interroger sur la structuration sémantique du lexique et sa modélisation, d'une part, et à évaluer les méthodes de

repérage automatique, en choisissant finalement une méthode liant l'expertise linguistique et un calcul stattistique sur corpus.

2 Etat de l'art sur le repérage de relations sémantiques

Dans cette section, nous parcourons les différentes approches et travaux utilisés jusqu'ici pour mettre au jour les relations sémantiques entre lexies, en commençant par l'exploitation des ressources sémantiques existantes et des modèles sous-jacents. Nous présentons ensuite les approches basées sur les contextes dénotant des relations sémantiques, puis l'approche distributionnelle.

2.1 Ressources sémantiques et encyclopédiques existantes pour le français

Les dictionnaires du français ont une longue et riche histoire, et il est imaginable d'en faire l'exploitation informatique. Dans le cadre de cette étude, nous n'évoquerons que les ressources existantes liées au français contemporain, accessibles numériquement et librement pour la recherche.

2.1.1 Méthode lexicographique manuelle

Il s'agit de la méthode traditionnelle. Trois ouvrages, initialement sous format papier, sont exploitables pour le français : le Trésor de la Langue Française[1], le Littré[2] et les différentes versions du dictionnaire de l'Académie Française[3]. Ces ressources présentent les défauts classiques de ce type d'ouvrage : les informations sémantiques n'y sont pas décrites systématiquement, ni dans un langage formalisé. Par ailleurs, ces ouvrages n'ont pas de mise à jour disponible pour la période la plus contemporaine[4]. On peut tout de même imaginer exploiter pour le TAL les "définitions" de ces ouvrages, dans un objectif de modélisation de ce type d'informations, et pour extraire des informations sémantiques pour le français dans une perspective diachronique.

Deux ressources sémantiques plus récentes ont été spécifiquement développées dans le cadre du TAL. D'une part *Wordnet* (Miller, 1990), qui propose un réseau lexical hiérarchisé sur la base d'un noyau de relations sémantiques (hyperonymie, hyponymie, synonymie, antonymie), ensuite étendu à d'autres relations (méronymie principalement). Pour le français, une seule réalisation manuelle, EuroWordnet (Vossen, 1998), réunissant une quinzaine de langues européennes, a été produite sur la base du modèle Wordnet, avec une couverture très limitée et un droit d'accès restreint.

Sur la base du modèle Sens-Texte, plusieurs réalisations lexicographiques ont été produites manuellement, toutes très limitées quantitativement étant donné la richesse descriptive du modèle. Pour le français, le projet RELIEF mené à l'ATILF vise à produire une ressource sémantique centrée sur un noyau de fonctions lexicales, pour un noyau de lexies. (Lux-Pogodalla et al., 2011; Polguère, 2014).

Un autre modèle a été utilisé, rendant compte non plus des relations sémantiques mais des structures argumentales des lexies cibles. Cette ressource, issue des principes de la linguistique cognitive (Fillmore, 1982 par exemple) a donné naissance à FrameNet (Baker et al., 1998), une ressource construite manuellement à partir d'une analyse de corpus, et d'un modèle des rôles argumentaux. Une ressource similaire n'est actuellement pas disponible pour le français, mais un projet est en cours (Candito et al., 2014).

2.1.2 Méthode collaborative

Un second moyen de mettre en place une ressource lexicographique consiste à la construire collaborativement. Nous renvoyons à (Simko et Bielikova, 2014) pour une présentation détaillée de cette approche, qui ne sera pas explorée dans cet article.

2.1.3 Méthode automatique à base de ressources lexicales

Une troisième méthode consiste à "transformer" une ressource existante pour la rendre informatiquement exploitable. Concernant le français, deux ressources ont ainsi été exploitées : d'une part, *WordNet*, d'autre part le *Wiktionnaire*.

---

1 http://atilf.atilf.fr
2 http://www.littre.org
3 http://atilf.atilf.fr/academie9.htm
4 Le TLF a une couverture qui s'arrête dans les années 1970; le Littré est un ouvrage de la fin du XXème siècle; la huitième édition du dictionnaire de l'académie date de 1932-1935 et la neuvième est en cours de numérisation

WordNet est la ressource sémantique de référence en TAL, même si le modèle adopté, basé sur des relations sémantiques hors contexte, prête à de nombreuses critiques. Les chercheurs ont développé différentes techniques pour construire un *WordNet* français automatiquement. La technique de base consiste à traduire les lexies de WordNet vers la langue cible au moyen de lexiques bilingues (Sagot et Fišer, 2008 ; Mouton et de Chalendar, 2010). La difficulté réside dans les mots polysémiques, puisque les lexiques bilingues ne reprennent pas la nomenclature de WordNet. JAWS (Mouton et de Chalendar, 2010) utilise des modèles syntaxiques distributionnels pour désambiguïser les différents sens des lexies dans WordNet. Un autre système, WOLF (Sagot et Fišer, 2011) effectue la désambiguïsation des polysèmes de WordNet au moyen de corpus parallèles. Depuis, les auteurs ont produit de nombreuses améliorations (Apidianaki et Sagot, 2012 ; Sagot et Fišer, 2012), mais la ressource est encore loin d'être exploitable.

L'exploitation du Wiktionnaire a été faite principalement par (Sajous et al., 2014). Le Wiktionnaire comprend environ 2 millions d'entrées, dont environ 1,4 millions d'entrées lexicales pour 186 00 lemmes. Parmi les nombreuses informations que chaque article peut contenir, figurent dans la très grande majorité des cas des définitions et, de manière sporadique, des relations sémantiques. L'écueil principal concerne le format non systématique de la source, qui rend la ressource difficilement exploitable. Cependant, les auteurs sont parvenus à extraire un certain nombre d'informations phonétiques et morphosyntaxiques utiles. Dans le cadre du présent travail, nous avons développé un programme permettant d'extraire un certain nombre d'informations à valeur sémantique (définitions et relations sémantiques) à partir de cette ressource.

Il faut enfin noter de nombreux travaux concernant la compilation et l'unification de différentes ressources (dictionnaires et encyclopédies) : BabelNet (Navigli et Ponzetto, 2012), Freebase (Bollacker et al., 2008), YAGO (Suchanek et al., 2007) et DBPedia (Auer et al., 2007). Ces outils génèrent des ressources multilingues à grande échelle, qui fournissent malheureusement des résultats encore trop bruités.

2.1.4 Méthodes automatiques à base de corpus textuels

Ces approches, privilégiées étant donné le coût prohibitif de la confection manuelle d'une ressource sémantique, visent à identifier automatiquement des relations sémantiques entre lexies sur gros corpus. Plusieurs techniques ont été utilisées.

Les approches symboliques reposent sur le postulat qu'il existe des propriétés linguistiques internes (par exemple des radicaux communs explicitant des dérivations lexicales) et surtout externes (contextuelles : patrons lexico-syntaxiques) signalant des relations sémantiques. On identifie alors les patrons dénotant les relations cibles, on les formalise, puis on reconnaît sur corpus les occurrences de ces patrons. La première réalisation de cette technique revient à (Hearst, 92), qui a particulièrement travaillé sur la relation d'hyperonymie. Elle indique quelles doivent être les propriétés de ces patrons :

> *« We identify a set of lexico-syntactic patterns that are easily recognizable, that occur frequently and across text genre boundaries, and that indisputably indicate the lexical relation of interest. » ((Hearst, 92, p.539)*

Pour « découvrir » les patrons, elle propose un algorithme d'amorçage (*bootstrapping*) dont nous parlerons plus loin. En français, de nombreux travaux ont été mené par l'équipe de Jean-Pierre Desclès depuis 1990, avec une technique similaire[5], pour reconnaître des relations aussi diverses que les relations causales (Garcia, 1998), la méronymie (Jackiewicz, 1999), les relations temporelles et aspectuelles (Battisteli, 2009) ou les expressions définitoires (Cartier, 2004). (Morin et Jacquemin, 2004) ont également repris les travaux de Hearst, en automatisant la découverte des patrons, pour mettre en place un système de repérage de relations sémantiques. Cette technique a les inconvénients suivants : difficulté à expliciter des relations « génériques », ambiguité de certains patrons, coût de développement.

Les méthodes semi-supervisées utilisent des corpus annotés pour apprendre, dans le même esprit mais de manière automatique des propriétés linguistiques signalant les relations visées. Les algorithmes d'apprentissage sont variés : Machine à Vecteur de Support (SVM) (Zhao and Grishman, 2005; Bunescu and Mooney, 2006), régression logistique (Kambhatla, 2004), parsing augmenté (Miller et al, 2000), Champs Conditionnels Aléatoires (CRF) (Culotta et al, 2006). Ces méthodes présentent deux défauts principaux : temps de développement des corpus d'apprentissage, problème d'adaptabilité à de nouveaux domaines. Voir (Bach et Badaskar, 2007) pour une présentation détaillée de ces méthodes.

Amorçage : (Hearst, 1992) en présentait déjà le fonctionnement générique : on identifie plusieurs termes représentatifs de la relation sémantique visée (appelés *seeds*), on repère en corpus les occurrences des termes (proximité en nombre de mots, ou en terme de distance syntaxique), puis on récupère les patrons correspondants - candidats patrons pour la relation étudiée-, le processus étant itératif, jusqu'au point d'arrêt (nombre de patrons atteint, découverte de patrons épuisée, etc.). Sur cette base de nombreux travaux ont été menés (par exemple Pantel et Pennachioti, 2006 ; Kozareva, Riloff, and Hovy 2008 ; 2010).

---

[5] Similaire seulement, car la méthode d'exploration contextuelle comprend deux temps : le premier consiste à marquer dans le texte des indicateurs de relations sémantiques, puis le second applique des patrons lexico-syntaxiques pour repérer les arguments de la relations sémantique. Voir (Desclès, 2006)

Ce type de système a l'avantage de ne pas nécessiter une intervention humaine prohibitive, mais il présente un inconvénient majeur, un faible rappel. En effet, certains patrons peuvent être ambigus, un patron ambigu récupérant de mauvais exemplaires, qui vont à leur tour récupérer de mauvais patrons. A notre connaissance, aucune méthode n'a encore été trouvée pour gérer ces glissements de sens.

Utilisation des contextes définitoires : une autre voie pour accéder aux relations sémantiques lexicales consiste à utiliser les contextes définitoires dans les textes. A l'évidence, en corpus, ces fragments textuels sont quantitativement moins importants que les patrons lexico-syntaxiques exprimant telle ou telle relation sémantique, mais on peut penser que les informations y sont plus fiables.

Plusieurs travaux ont été menés dans cette direction, dans différentes langues et sur différents corpus. (Pearson, 1998) décrit les différents propriétés linguistiques des contextes définitoires en anglais. (Meyer, 2001) fait de même dans un contexte terminologique. (Storrer et Wellinghoff, 2006 ; Walter et Pinkal, 2006) ont travaillé sur l'allemand, (Pinto et Oliveira, 2004) sur le portugais, (Leu et Ko, 2007) sur le chinois. Plusieurs systèmes ont été mis en place : Definder (Klavans et Muresan, 2001) sur les textes médicaux anglais, (Alarcón et al., 2008, 2009) sur l'espagnol. (Navigli et al., 2007, 2010) ont développé *GlossExtractor*. Les auteurs proposent de travailler à partir des *glosses*[6] de Wikipedia. Ce projet a donné lieu à des thesaurus en cinq langues. (Cartier, 2004) a étudié la structure lexico-syntaxique des définitions en français en partant d'un corpus issu de l'Encyclopédie Universalis et d'articles techniques et scientifiques. Des travaux ont été mené pour extraire automatiquement du TLF des relations hyperonymiques à partir des définitions (Barque et al., 2010). (Rebeyrolle, 2000) a également travaillé sur les contextes définitoires.

La plupart des travaux élaborent manuellement les patrons lexico-syntaxiques signalant des contextes définitoires. Mais certains auteurs automatisent le processus. Par exemple (Navigli, 2010) utilisent les Word Classes Lattices. (Serra, 2009) et (Cartier, 2004) insistent sur la dispersion des contextes définitoires et le continuum qui existe entre une expression définitoire proprement dite, un contexte propre à telle ou telle relation sémantique, et même toute mention du terme défini. Au final, les deux auteurs considèrent que « toute occurrence d'une lexie est de facto définitoire », reprenant la notion de *définition implicite* établie par (Gergonne, 1818).

Méthodes distributionnelles : depuis les années 90, le paradigme statistique, issu des intuitions du distributionnalisme (Harris, 1954) puis de la linguistique de corpus (Firth, 1957; Miller et Charles, 1991), domine les travaux en Traitement Automatique des Langues (TAL), avec des réalisations convaincantes dans tous les domaines : reconnaissance automatique des unités (poly-)lexicales, des parties du discours, des relations sémantiques, modèles probabilistes du langage, etc. Ces études ont mis à jour des phénomènes linguistiques au travers des notions de collocations, de « multiword expressions », de « word sketches » (Kilgariff et al., 2004) ou encore de « similarité sémantique ».

Ces différentes approches se basent sur l'hypothèse que le meilleur moyen d'accéder au fonctionnement des langues est de se baser sur les traces matérielles, et la répétition des séquences constitutives. Plusieurs hypothèses « sémantiques » en sont dérivées (Turney et Pantel, 2010 ; Baroni et Lenci, 2010 ; Clark, 2015) : tout d'abord, le calcul des répétitions de séquences permet de mettre au jour les préférences sélectionnelles, et donc l'usage, des lexies. Par exemple (Kilgarriff, 2004) déduit, sur cette base, les structures argumentales les plus fréquentes des verbes, qu'il appelle *Word Sketches*. (Hanks, 2013) théorisera cette approche avec la *Théorie des Normes et des Exploitations*, les normes (usages) pouvant être décrites par les constructions les plus fréquentes utilisées avec une lexie donnée, et les exploitations (ruptures ou évolutions d'usage) par des constructions attestées moins fréquentes. Une autre hypothèse dérivée se résume ainsi : deux lexies partageant un grand nombre de contextes sont « similaires » sémantiquement (Harris, 1954). C'est ce flou dans l'identification exacte de la relation sémantique, et l'absence actuelle de mesure pour les distinguer qui rend la notion de « similarité » difficilement opérationnalisable.

Cependant, il est évident que le calcul des répétitions nous informe *en quelque manière* sur le sens des lexies, et la répétition de séquences dénotant des relations sémantiques nous informe également sur les patrons les plus employés. Dans ce cadre nous utiliserons l'outil SDMC (Béchet et al ., 2013), outil venu de la fouille de données, qui permet de repérer les séquences répétitives en utilisant une combinaison de traits (forme linguistique, lemme et partie du discours) pour découvrir les patrons lexico-syntaxiques les plus fréquents. Nous utiliserons cet outil pour guider la recherche manuelle des patrons dénotant des relations sémantiques dans les définitions.

L'approche suivie dans l'expérimentation qui suit se positionne donc dans la perspective de la découverte de patrons lexico-syntaxiques, étant donné la précision qu'ils permettent d'obtenir. Pour éviter l'écueil de la dispersion des patrons sur corpus, nous focaliserons sur des énoncés privilégiés, les définitions. Pour contourner la difficulté liée au temps de développement prohibitif des ressources, nous utiliserons SDMC, un outil permettant d'extraire les patrons lexico-syntaxiques les plus fréquents sur corpus.

[6] les premières phrases de chaque article dont l'objectif éditorial est d'expliciter les caractéristiques essentielles du concept décrit

3 Expérimentation : extraction automatique des relations sémantiques dans les définitions

Nous présentons ci-après l'expérimentation menée pour repérer des relations sémantiques dans les définitions en français. Après avoir brièvement présenté le modèle définitoire retenu, nous détaillons le corpus, l'architecture du système et les différents processus constitutifs. Nous terminons par la présentation des résultats et leur évaluation.

3.1 Modélisation de la définition

La notion de *définition* a été étudiée par la philosophie, la logique, la linguistique et la psychologie. Nous partirons ici du modèle « classique » de la définition, représenté par (Arnauld et Nicole, 1662). Dans cette conception, qui a impacté et impacte fortement la lexicographie, une définition (de mot) est une proposition dénotant les propriétés essentielles d'une unité lexicale. On appelle *definiendum (DFN)* l'unité définie, et *definiens (DFS)* le segment textuel définissant. Appelons également *relateur définitoire (DFR)* la séquence linguistique mettant en relation d'identification les deux éléments précédents. Par exemple, en partant de :

> *(1) Le chat domestique (Felis silvestris catus) est un mammifère carnivore de la famille des félidés. Il est l'un des principaux animaux de compagnie et compte aujourd'hui une cinquantaine de races différentes reconnues par les instances de certification. (Wikipedia)*

nous aboutissons à :

DFN : Chat domestique *DFR : .. est un... DFS : mammifère carnivore de la famille des félidés*

DFN : Chat domestique *DFR : ...est un des principaux... DFS : animaux de compagnie*

Ici, deux définitions « classiques » peuvent être identifiées, avec deux relateurs différents. Le dernier segment (*et compte... certification.*) peut être caractérisé comme une prédication *essentielle*, étant donné le contexte définitoire. Nous appellerons ce type de segment *prédication définitoire*. Le segment nominal entre parenthèses (*Felis silvestris catus*) est un synonyme du terme défini, dont le relateur est complexe : *DFN (SYN)*.

Le *definiens* se compose, dans la vision classique de la définition, de deux parties : un hyperonyme et des différences spécifiques. Ce modèle utilise les notions problématiques et floues de « différences » (ou encore propriétés) et « spécifiques », que nous tenterons de préciser après l'expérimentation. Ce modèle classique, aristotélicien, du *definiens*, a été étendu à d'autres informations définitoires : définition synonymique, méronymique, causale, télique, etc. Voir par exemple (Martin, 1992) pour une typologie des définitions lexicographiques.

L'énoncé définitoire, comme toute énoncé, comprend également une prise en charge énonciative et une modalisation implicites ou explicites : *Selon Sisley, un Composant G est un... / XXX considère que le composant G est un...* Dans le corpus qui est le nôtre, ces prises en charge énonciatives resteront implicites.

Enfin, une restriction de domaine peut limiter la portée de la définition : *En astrophysique, on appelle XXX* …

3.2 Corpus

Les définitions se rencontrent principalement dans les dictionnaires et les encyclopédies. Le corpus utilisé devra également être dans le domaine public, être disponible sous format électronique, et décrire le vocabulaire général du français ; enfin, étant donné que cette expérimentation vise à mettre en place un prototype d'extracteur, nous avons restreint notre corpus aux définitions de noms.
Les trois ressources suivantes seront donc utilisées :

- **Trésor de la Langue Française informatisé (TLFI)** : le laboratoire ATILF a cordialement accepté de nous fournir une version électronique du TLFI, qui comprend 61 234 unités lexicales nominales pour un total de 90 348 définitions;
- **Version française du Wiktionnaire (FRWIK)** : nous avons mis au point un utilitaire permettant d'en extraire les termes définis, la partie du discours, ainsi que les différentes définitions. Au total, ce corpus est constitué de 140 784 noms,pour un total de 187 041 définitions ;
- **Version française de Wikipedia (WIKP)** : nous avons également utilisé Wikipedia, même s'il s'agit d'une ressource de type « encyclopédie collaborative » ; mais d'autres chercheurs (Navigli and al, 2008) ont montré l'intérêt des premières phrases de chaque article, qui « définissent » les aspects essentiels du concept décrit. Nous avons donc extrait de cette ressource 610 013 entrées nominales[7] et la première phrase de chaque article.

3.3 Architecture du système

Le système mis en place comprend cinq étapes de traitement :

1. Nettoyage du corpus : il s'agit ici de transformer le fichier de départ en un fichier texte propre comportant une définition par ligne ; cette étape a aussi consisté à convertir le wiktionnaire de son format web à un format

[7] A noter que (Navigli et al. 2008) n'utilisent que 410 000 définitions car ils opèrent un filtrage sur la présence du marqueur définitoire *être un*.

exploitable, en extrayant les seules définitions ; il s'agit également de préparer le corpus pour l'étape suivante, notamment par segmentation en mots en en phrases ;
2. Analyse morphosyntaxique du corpus : elle a été conduite avec TreeTagger (Schmid, 1994);
3. Identification des marqueurs définitoires : il s'agit de marquer dans le texte les lexies marquant une relation sémantique donnée ;
4. Simplification/ canonisation des phrases : cette étape vise à simplifier les énoncés définitoires pour faciliter l'application ultérieure des patrons;
5. Extraction des relations sémantiques au moyen de patrons lexico-syntaxiques.

Dans cet article, nous ne détaillerons pas les deux premières étapes, nous concentrant sur les trois dernières.

3.3.1 Identification des marqueurs définitoires

Cette étape consiste à identifier et marquer dans la phrase source les unités lexicales (simples ou composées) permettant d'identifier un énoncé définitoire ou l'un de ses composants. Reprenant des travaux précédents (Rebeyrolle, 2000 ; Cartier, 2004), nous avons utilisé la liste de marqueurs suivants :

Type de marqueur	Échantillon de marqueurs	Nombre total
Relateur définitoire 1 (REL1)	*être ,":" , c'est-à-dire, ou...*	11
Relateur définitoire 2 (REL2)	*se définir comme, vouloir dire, s'appeler, signifier...*	34
Marqueur de catégorisation (CAT)	*catégorie, famille, espèce...*	47
Marqueur de spécification (SPEC)	*exemple, exemplaire, prototype...*	23
Marqueur de propriété (PROP)	*se caractériser, avoir pour caractéristique...*	42

TABLEAU 1 : liste des marqueurs linguistiques de la définition en français

Le nombre de marqueurs est relativement bas. Comme nous le verrons dans la suite, leur couverture est maximale. Il faut néanmoins noter que la mise en place de cette liste nécessite un travail de dépouillement et une expertise non négligeables.

3.3.2 Simplification des phrases

La simplification/canonisation des phrases est une étape innovante dans le cadre des travaux sur l'extraction d'information à base de patrons lexico-syntaxiques. Son objectif principal est de réduire les énoncés à une forme « canonique », en repérant et supprimant les éléments accessoires et en ne conservant que les éléments essentiels au modèle définitoire. Cet objectif est réalisé en trois étapes :

1. Identification et suppression des éléments circonstanciels et accessoires des phrases, pour autant qu'ils ne font pas partie d'une partie constitutive du modèle définitoire adopté ;
2. Identification, *extraction* et suppression des informations définitoires périphériques : restrictions de domaine, marqueurs énonciatifs, relations sémantiques « annexes » au prédicat définitoire principal : ces informations sont conservées puis retirées de l'énoncé définitoire à analyser ;
3. Unification des groupes nominaux, puisqu'ils sont dans le cadre de cette expérimentation, les éléments cibles de la partie définissante, et que leur variété est un facteur de perte de précision pour les étapes suivantes.

Prenons un exemple de ces différents traitements. Soit la définition de *abaisse* dans le Wiktionnaire, après analyse morphosyntaxique et identification des marqueurs définitoires :

> *(2) en/P cuisine/NC et/CC en/P pâtisserie/NC ,/PONCT un/DET DEFINIENDUM être/V/DEF_REL un/DET pièce/NC de/P pâte/NC aplatir/VPP ,/PONCT généralement/ADV au/P+D rouleau/NC à/P pâtisserie/NC ou/CC un/DET laminoir/NC ./PONCT*

Elle sera réduite à la séquence suivante :

> *un/DET DEFINIENDUM être/V un/DET pièce/NC de/P pâte/NC aplatir/VPP ,/PONCT au/P+D rouleau/NC à/P pâtisserie/NC ou/CC un/DET laminoir/NC ./PONCT*

Et nous récupérons la restriction de domaine : *en/P cuisine/NC et/CC en/P pâtisserie/NC.*

La suppression concerne les adverbiaux, les compléments circonstanciels et les propositions subordonnées circonstancielles, qui sont des informations accessoires. La difficulté consiste à distinguer entre les informations circonstancielles portant sur l'énoncé dans son entier, qui sont la cible de ce processus, et les informations circonstancielles qui sont liées à l'un des composants de la définition, et qu'il faut conserver.

Par exemple, nous souhaitons supprimer la parenthèse dans :

(3) DEFINIENDUM ***(/PONCT proche/ADJ de/PREP la/DET capitale/NC)/PONCT*** *être/V un/DET atoll/NC de/P le/DET république/NC du/P+D Kiribati/NPP ./PONCT*

Mais pas la proposition relative dépendante du *definiens* :

DEFINIENDUM être/V du/P+ enzymes/NC ***qui/PROREL contrôler/V le/DET structure/NC topologique/ADJ de/P l'ADN/NC...***

Pour identifier les structures lexico-syntaxiques correspondant aux différents éléments à supprimer, nous avons calculé avec SDMC les patrons lexioc-syntaxiques les plus fréquents en deux positions : au début des énoncés (c'est-à-dire avant le *definiendum*) et entre le terme défini et le relateur définitoire. La zone du *definiens* comprend généralement des structures définissantes qui ne sont pas touchées par le processus.

Trois cas méritent mention :

- Dans Wikipedia, la grande majorité des éléments qui s'insèrent entre le terme défini et le relateur définitoire explicitent une relation sémantique synonymique : *DEFINIENDUM* ***(/PONCT parfois/ADV Apaiang/NPP ,/PONCT même/ADJ prononciation/NC)/PONCT*** *être\/V\/DEF_REL un/DET...*
- Les morphèmes adverbiaux de négation doivent être conservés ;
- Un certain nombre de compléments circonstanciels dénotent une restriction de domaine, notamment lorsqu'ils sont placés en tête d'énoncé. Pour les repérer et enregistrer leur contenu avant de les retirer de l'énoncé initial, nous avons modélisé sous forme d'expressions régulières, à partir des résultats de SDMC les patrons les plus fréquents:

  ^((?:en|dans|à|sur|selon|pour|chez|par).{5,150}?)\t, \/PONCT\t/
  DEFINIENDUM\t, \/PONCT\t((?:en|dans|à|sur|selon|pour|chez|par).{5,150}?)\t, \/PONCT}

3.3.3 Unification/réduction de la variété des syntagmes nominaux

Intuitivement, la grande majorité des définitions obéit au modèle hyperonymique, où le *definiens* est composé d'un hyperonyme suivi des différences spécifiques exprimées sous forme adjectivale (doté ou non de ses expansions) et/ou de proposition relative. Le premier élément du *definiens* est donc l'hyperonyme de la lexie définie. Mais plusieurs phénomènes complexifient l'identification de l'hyperonyme nominal et principalement : les déterminants composés (*est une des … ; est un ensemble de* ...), des adjectifs axiologiques (*est le principal moyen de....*), des termes marquant une relation spécifique (*est un genre de …*).

Pour traiter ces exceptions, nous avons listé la plupart des déterminants composés qui sont pris en compte lors de la segmentation en mots, et les marqueurs de relation sémantique sont identifiés en tant que tels afin d'éviter d'être les cibles de la relation sémantique à extraire.
Ensuite, les différentes formes de noms simples ou composés sont unifiées. Pour cela, en nous basant sur un certain nombre d'études (Ramish, 2015 ; Mathieu-Colas, 1996), nous reconnaissons les formes les plus fréquentes en français (ici sous forme d'expressions régulières):

NC (ADJ) ? de/P NC (ADJ) ?
NC (ADJ){0,3}
NPP+

Ce traitement est important car, d'une part, il facilite l'extraction des cibles argumentales des relations sémantiques définitoires principales (hyperonymie ou synonymie), et, d'autre part, il rend plus efficace l'extraction ultérieure des autres relations sémantiques.

3.3.4 Identification des relations sémantiques par patrons lexico-syntaxiques

La mise au point des patrons lexico-syntaxiques est la dernière étape, permettant de repérer automatiquement les relations sémantiques suivantes : hyperonymie, synonymie, méronymie, et prédications définitoires. Les traitements précédents ont fortement simplifié cette reconnaissance.
La mise au point des patrons a combiné l'expertise linguistique et l'analyse fréquencielle du corpus par SDMC. Cet outil est lancé sur les séquences définitoires issues des traitements précédents, éventuellement tronquées (voir plus loin). La recherche de séquences répétées a été faite en utilisant le paramétrage suivant[8] : recherche d'un maximum de cinq cooccurrents, fenêtre de quinze mots, combinant les traits: forme graphique, lemme, partie du discours, avec présence d'un marqueur de la relation en question. Le calcul statistique nous a permis d'aboutir pour chaque corpus a un noyau de patrons, correspondant à chaque type de relations sémantiques. L'expertise linguistique, à partir des résultats fréquenciels bruts, a permis d'une part de ne retenir que les patrons les plus fréquents et les moins ambigus (par sondage), puis, parmi les patrons les moins fréquents, de ne retenir que les patrons non ambigus pour la relation sémantique considérée.

La recherche fréquencielle de patrons est précédée de réductions de la phrase source selon la relation sémantique visée. Pour ce qui concerne les relations d'hyperonymie, de synonymie et de méronymie, pour s'assurer que la relation liait le *definiendum* à un syntagme nominal, le calcul a été effectué en tronquant l'énoncé définitoire sur la droite au niveau de

[8] Le choix des paramètres n'a pas de justification objective : il s'agit de paramètres maximisant la couverture des patrons repérés. Nous renvoyons à (Kiela et Clark, 2014) pour une évaluation des paramètres optimaux, notamment fenêtre de recherche et nombre de cooccurrents,, qui recoupe les intuitions choisies ici.

la séquence définissante, à la première occurrence d'un pronom relatif, d'un adjectif suivi de ses arguments ou d'un complément nominal doté d'un déterminant (DEFINIENDUM est un GN | PROPREL ; DEFINIENDUM est un GN | ADJ PREP ; DEFINIENDUM est un GN | PREP DET ...). Ces éléments signalent en effet le début d'une autre information définitoire liée à la prédication, et fournissent une frontière droite aux groupes nominaux à rechercher.

Pour ce qui concerne la relation de prédication, la recherche statistique s'est faite après la reconnaissance des autres relations sémantiques, permettant ainsi de cibler la recherche sur le reste de la définition. Par exemple, pour le definiendum *Baclage*, à partir de la phrase source :

(4) DEFINIENDUM :/PONCT redevance/NC au/P+D officier/NC préposer/VPP à/P ce/PRO arrangement/NC ./PONCT

Nous obtenons, après marquage de la relation d'hyperonymie :

DEFINIENDUM :/PONCT/DEF_REL ***GN (redevance)/HYPER*** *au/P+D GN(officier/NC) préposer/VPP à/P ce/PRO GN(arrangement/NC) ./PONCT*

Ce qui permet de faire une recherche seulement dans ce qui complète l'hyperonyme, où sont explicitées les différences spécifiques de la lexie. La recherche des patrons de prédications définitoires ne s'appuie sur aucun marqueur spécifique, et aboutit à identifier trois patrons très génériques correspondant aux expansions syntaxiques d'un nom en français : complément prépositionnel de nom (*(1) de la famille des félidés ;* (4) : *à l'officier*) groupe adjectival suivi ou non de ses expansions nominales ou verbales (*(1) carnivore ; (5) … affecté au transport urbain*), proposition relative. phrase coordonnée sans marqueur (*(1) … et compte aujourd'hui...*). Le coordination d'éléments a été partiellement modélisée.

3.4 Résultats et analyse

Nous présentons ci-après les résultats globaux puis les résultats par relation sémantique.

3.4.1 Résultats globaux

Le tableau 2 présente les totaux de repérage pour chaque relation sémantique et pour chaque corpus.

	Wiktionnaire (WIKT) : 186 502 lexies-définitions				**TLF : 90 190 lexies-définitions**				**Wikipedia (WIKIP) : 610 013 lexies-définitions**			
	Nbre	Moyenne par déf.	Nbre de déf. sans extraction	% sans relation	Nbre	Moyenne par déf	Nbre de déf. sans extraction	% sans relation	Nbre	Moyenne par déf	Nbre de déf. sans extraction	% sans relation
hyperonymie	187934	1,02	2435	1,31%	90605	1,03	2364	2,62%	592311	1,03	35459	5,81%
synonymie	785	1,53	185988	99,72%	376	1,2	89876	99,65%	83276	1,24	542911	89,00%
méronymie	1313	1	185189	99,30%	1733	1,11	88627	98,27%	0	0	610013	100,00%
prédications	469325	2,55	2435	1,31%	232300	2,65	2364	2,62%	1398178	2,43	35459	5,81%
domaine	9413	1,02	177245	95,04%	5779	1,43	86143	95,51%	31834	1,05	579598	95,01%

TABLEAU 2 : résultats globaux d'extraction des relations sémantiques

Ces chiffres appellent plusieurs commentaires :

– la relation d'hyperonymie est la plus représentée dans le corpus, de très loin devant la synonymie et la méronymie ; on note tout de même une plus grande représentation de la synonymie dans Wikipedia, sachant que dans un certain nombre de cas, les gloses comprennent à la fois un synonyme et un ou des hyperonymes ; ce constat est à rapprocher d'un modèle définitoire comprenant dans la quasi totalité des cas un hyperonyme et des propriétés spécifiques ;

– la moyenne d'extraction d'une relation donnée par définition est intéressante, puisqu'on note que, tandis que l'hyperonymie, la méronymie, et les domaines sont proches d'une instance par définition, les synonymes ont une tendance à se multiplier dans une même définition (1,53 pour WIKT) ; les prédications sont évidemment multiples dans chaque définition, avec une moyenne de 2,5 prédications extraites, le TLF ayant la plus grande diversité ;

– les restrictions de domaine sont présentes de manière stable entre les corpus : environ 5% des définitions comprennent cette information[9];

– Parmi les définitions pour lesquelles aucune extraction d'hyperonyme n'a eu lieu (colonnes 3 et 4 pour chaque ressource), les causes en sont diverses :

- absence du patron qui aurait permis le repérage : *abondement/NC :/PONCT/DEF_REL lorsque/CC du/P+D GN(salarié/NC) acheter/V du/P+D GN(action/NC) de/P leur/DET GN(propre/ADJ société/NC) ,/PONCT le/DET GN(abondement/NC)* ***correspondre/V/DEF_REL au/P+D*** *GN(versement/NC complémentaire/ADJ) verser/VPP par/P le/DET GN(société/NC) ./PONCT*

[9] A noter que le TLF explicite cette information dans un champ spécifique différent de la définition

- énoncé définitoire incomplet : *absence/NC :/PONCT/DEF_REL se/PRO employer/V/DEF_REL absolument/ADV ./PONCT*
- énoncé définitoire trop complexe pour être analysé : *alloux/NC :/PONCT/DEF_REL en/P GN(droit/NC foncier/ADJ) allodial/ADJ un/DET GN(alloux/NC)* ***être/V/DEF_REL le/DET GN(inverse/NC) de/P*** *un/DET GN(fief/NC) en/P GN(droit/NC féodal/ADJ) ./PONCT*
- erreur d'analyse morphosyntaxique : *acide/NC :/PONCT/DEF_REL* ***liquider/VS*** *capable/ADJ de/P attaquer/VINF et/CC de/P dissoudre/VINF le/DET GN(métal/NC) ,/PONCT certain/PRO GN(roche/NC) ./PONCT*

3.4.2 Hyperonymie

Les résultats d'extraction par patrons sont détaillés dans le tableau 3.
On constate :

- La dispersion des patrons est plus importante dans le corpus Wikipedia, ce qui s'explique par le caractère collaboratif et moins normé de cette ressource.
- Dans les corpus dictionnaires, en additionnant les patrons dérivés (par exemple, la coordination de groupes nominaux pour le patron générique *DEF être un GN*, ou encore la formulation à pronom relatif présentatif : *celui qui/ce qui...*), un seul patron couvre plus de 90% des occurrences : le patron *DEF être un/le GN/HYPER*. Seulement trois patrons se partagent la totalité des occurrences. En plus du précédent : *DEF : genre de GN/HYPER*, *DEF : nom de GN/HYPER*.
- Le corpus encyclopédique a une bien plus grande dispersion d'expressions : le patron avec *être* représente un peu plus de 50% des cas, et on rencontre également le patron *nom de GN*. La particularité de ce corpus ressortit à l'utilisation massive d'une hyperonymie par apposition et parenthésage après le terme défini, qui représente environ 35% des cas. Mais, de manière globale, on constate que le nombre de patrons est limité.

Patron générique	**Patron**	**Nb**	**%**	**Nb**	**%**	**Nb**	**%**
		TLF		**WIKT**		**WIKP**	
DEF_:_DEF_REL_(DET) ?_ GN/HYPER	DEF_:_(le) ?_GN	62969	78,45%	93676	80,67%	570	
	DEF_:_(le) ?_GN_,_GN_et_GN	573	0,71%	526	0,45%		
	DEF_:_(le)_GN_et_GN	5601	6,98%	4858	4,18%		
	DEF_:_(tout)_CE_PROREL	1290	1,61%	800	0,69%		
	DEF_:_*_DEF_REL_*_GN	3966	4,94%	5087	4,38%		
	DEF_:_CELUI_PROREL	1221	1,52%	2945	2,54%		
	DEF_:_PROREL	1938	2,41%	170	0,15%		
	DEF_REL|DEF_,_GN_,	225	0,28%	224	0,19%	61099	
	DEF_(_(ou) ?_GN_)					55638	
	DEF_(_GN_)					75219	
	DEF_,_(ou) ?_GN_, ?					24804	
	DEF_,_GN					3744	
	GN_(de)_DEF					29668	
	GN_OU_DEF					1385	
	Être_..._GN					26668	
	être_un_GN					277803	
	être_un_GN_et_un_GN					28861	
	être_un_GN_,_un_GN_et_un_GN					8162	
	Être_un_ADJ_GN					6322	
	Être_un_ADJ_et_ADJ_GN					5	
(DEF_ :_) ?_*_(genre, espèce...)_*_GN/HYPER	DEF_:_genre_de_GN	700	0,87%	2432	2,09%		
	être_un_genre_de_GN					14991	
	DEF_:_genre_de_GN_,_de_GN_et_de_GN	3	0,00%	9	0,01%		
	DEF_:_genre_de_GN_et_de_GN	40	0,05%	93	0,01%		
	genre_de_(ADJ)_GN	129	0,16%	544	0,47%		
	genre_de_GN_,_de_GN_et_de_GN	2	0,00%	14	0,01%		
	genre_de_GN_et_de_GN	33	0,04%	47	0,04%		
(DEF_ :_) ?_*_(nom,	DEF_REL_*_nom_*_de_GN	326	0,41%			3272	

dénomination...)_*_GN/HYPER							
	DEF_REL_*_nom_donne_*_a_GN	138	0,17%	622	0,54%	274	
	nom_(de)_"_GN_"	10	0,01%	27	0,02%		
	nom_*_de_GN	1005	1,25%	1813	1,56%		
	nom_*_donne_*_a_GN	62	0,08%	322	0,28%		
	DEF_REL_*_nom_*_de_GN			1768	1,52%		
signifier	signifier_"_GN_"	35	0,04%	136	0,12%		
	signifier_"_GN_"_ou_"_GN_"			4	0,00%		
	Total	**80266**		**187934**		**592311**	

TABLEAU 3 : distribution des patrons de l'hyperonymie sur les trois corpus

3.4.3 Méronymie/holonymie

La méronymie est généralement exprimée comme information principale de l'énoncé définitoire, dans un patron générique du type : *DEF est un(e) (partie, ...) de GN*.
L'holonymie est exprimée de manière converse par le patron : *DEF est (composée, constituée de...) GN.*
Il est également possible de rencontrer un troisième patron combinant expression hyperonymique et méronymique : *DEF est un(e) HYPER (composé de) GN...*
Comme montré dans le tableau 2, très peu de relations méronymiques ont été repérées, pourtant, comme l'évaluation le montrera, la précision est très bonne. Une analyse plus fine devra être menée pour recueillir plus de patrons et les affiner.

3.4.4 Synonymie

Cette relation sémantique est l'une des techniques de définition dans le cadre lexicographique, mais exclut le plus souvent une définition par hyperonyme, dans les sources lexicographiques.
Dans Wikipedia, les gloses contiennent très souvent à la fois une définition par hyperonyme et une ou des expressions synonymes, principalement par le biais d'une apposition ou d'une indication entre parenthèses qui suit le terme défini (voir exemple 1).
Dans le TLF et le Wiktionnaire, l'expression synonymique est exclusive de l'expression hyperonymique. Dans ces cas, la définition est une énumération (parfois réduite à un seul membre) de groupes nominaux coordonnés.

3.4.5 Prédicats définitoires

Ce que nous avons appelé *prédicats définitoires* dénotent les différences spécifiques et ont été repérés sur la base des constructions syntaxiques expansions de groupes nominaux. Elles ne sont donc pas typées *a priori* sémantiquement mais sont articulées autour d'un prédicat verbal ou adjectival, qui exprime la relation sémantique. Un nombre important de patrons dénotent la relation sémantique de *fonction* (*servant à, utilisé pour, etc.*). Ces prédications devront dans un prochain travail être consolidées afin de mettre au jour la structuration sémantique des termes définis.

3.5 Evaluation

En l'absence de ressource de référence, nous avons mis au point un protocole d'évaluation manuelle par trois linguistes. Cinq cent définitions ont été extraites du corpus, et les différentes informations définitoires (hyperonymie, synonymie, méronymie, domaine, prédications) ont été annotées par trois annotateurs experts. Les instructions étaient les suivantes : à partir des extractions automatiques (type de relation sémantique, lexies mises en relation) et de la définition source (où se trouve la *référence*), indiquer si la relation sémantique est correcte ou erronée ; dans le second cas, il était également demandé aux évaluateurs d'indiquer en commentaire les raisons de l'erreur et, le cas échéant, les bonnes lexies dans la relation. donnée La totalité des annotations divergentes entre annotateurs (trente-deux cas) ont été résolues d'un commun accord. Au total, le corpus de référence, la précision et le rappel des extractions sont les suivants :

Information définitoire	**Corpus de référence**	**Extractions automatiques**	**Extractions correctes**	**Précision**	**Rappel**
hyperonymes	512	489	477	0.975	0.931
synonymes	137	123	109	0.886	0.795
méronymes	67	43	35	0.813	0.522
prédications définitoires	976	1012	953	0.941	0.976
domaine	18	16	16	100	0.888

TABLEAU 4 : résultats de l'évaluation manuelle sur un échantillon de 500 définitions

On notera le rappel assez faible pour la relation de méronymie, qui s'explique par des patrons trop imprécis et lacunaires. Les extractions fautives pour les synonymes et les hyperonymes proviennent le plus souvent d'une reconnaissance partielle des groupes nominaux correspondants (*animal* au lieu de *animal de compagnie* , *fille* au lieu de *jeune fille*), ainsi que d'erreurs d'analyse morphosyntaxique. Pour les prédications, le découpage sur la base de patrons génériques génèrent un taux d'erreur minimal qui pourrait encore être amélioré par une analyse en dépendance.

Extractions communes entre dictionnaires : nous avons comparé les extractions entre les différentes ressources pour les mêmes lexies, en partant de l'hypothèse que si des relations sémantiques étaient repérées plusieurs fois dans plusieurs sources, cela avait toute chance de valider la relation extraite. Cette comparaison a été faite pour les relations de synonymie et d'hyperonymie, les autres informations extraites pouvant difficilement être comparées étant donné la variabilité des expressions et l'absence de généralisation effectuée. Les résultats sont présentés dans le tableau 5.

	Total éléments	1 dictionnaire	%	2 dictionnaires	%	3 dictionnaires	%
Lexies	886705	806497	90,95%	33358	3,76%	46850	5,28%
Hyperonymes	870850	781864	89,78%	52802	6,06%	36184	4,16%
Synonymes	84437	79978	94,72%	2600	3,08%	1859	2,20%

TABLEAU 5 : extractions communes entre dictionnaires

On note un recoupement assez faible des informations sémantiques entre dictionnaires. Pourtant, comme l'a montré l'évaluation manuelle, la précision des repérages est importante. Cela signifie que les dictionnaires expriment des relations sémantiques en utilisant soit des variantes (pour la synonymie et l'hyperonymie), soit spécifient des hyperonymes en se plaçant à différents niveaux d'abstraction. De cette évaluation, nous pouvons tirer un corpus de relations sémantiques consolidées de référence, puisque nous avons, en cumulant les éléments communs à deux ou trois dictionnaires, 88 986 (10,22%) hyperonymes communs et 4 459 (5,28%) synonymes communs.

Extractions communes avec d'autres ressources : une autre expérimentation à consisté à comparer les extractions de SemDef avec celles contenues dans Wolf et dans le Wiktionnaire. Ces deux ressources n'ont elles-mêmes pas été validées, mais nous pouvons partir de la même hypothèse que pour l'évaluation précédente. Pour le Wiktionnaire, nous sommes partis de l'extraction présentée dans cet article ; parmi les lexies étiquetées « nom commun »[10], 114004 ont au moins une relation sémantique explicitée ; nous avons conservé celles qui s'apparentent aux deux relations étudiées ici[11], soit 38 018 relations de synonymie et 124152 relations d'hyperonymie. Pour Wolf, nous sommes partis de la version 1.0b4 au format XML disponible en ligne. Nous en avons extrait les synsets ayant des réalisations en français de type nominaux (42427) : la relation de synonymie provient des réalisations en français d'un synset (soit 50130), et les relations d'hyperonymie ont été extraites directement (58359). Le tableau 6 détaille les résultats de la comparaison des ressources, en partant de la ressource cumulée SemDef obtenue par l'expérience précédente. On constate que le cumul donne des résultats comparables à ceux obtenus entre dictionnaires.

	Total éléments	1 dictionnaire	%	2 dictionnaires	%	3 dictionnaires	%
Lexies	300895	271799	90,33%	20531	6,82%	8565	2,85%
Hyperonymes	693356	682259	98,40%	10281	1,48%	816	0,12%
Synonymes	145248	136617	94,06%	7957	5,48%	674	0,46%

TABLEAU 6 : extractions communes entre SemDef, Wolf et Wiktionaire-relations-sémantiques

[10] Nous insistons sur ce point car certains noms communs sont en réalité des noms propres, ou en sont dérivés, comme les gentilés.

[11] Les relations explicitées comme hyperonymes ou hyponymes ont été assimilées à la relation d'hyperonymie, et les relations suivantes à la synonymie : Synonymes, Quasi-synonymes, Noms_vernaculaires, Variantes, Variantes_dialectales, Variantes_orthographiques, Abréviations, Anciennes_orthographes, Diminutifs, Synonymes_pour_la_définition

4 Conclusions, perspectives

Dans cet article, nous avons présenté une expérimentation visant à extraire automatiquement des relations sémantiques dans des définitions issues de différents corpus. A partir de près d'un million de définitions de noms extraites du TLF, du Wiktionnaire et de Wikipedia, nous avons pu extraire près de 900 000 relations d'hyperonymie et près de 100 000 relations de synonymie, avec un taux de précision supérieur à 90% sur un échantillon aléatoire de 500 relations. Nous avons également extrait près de 2 millions de prédications dont le statut définitoire est avéré étant donné le contexte définitoire, mais qui demandent une étude approfondie. Les différentes ressources seront mises à disposition de la communauté scientifique à l'occasion du colloque TALN, ainsi que les corpus source Wiktionnaire et Wikipedia.

Cette expérimentation a combiné deux approches complémentaires : une approche symbolique centrée sur la notion de patron lexico-syntaxique, ici dénotant une relation sémantique lexicale, avec pour objectif d'en décrire les différentes formes dans le corpus choisi ; une approche statistique basée sur le calcul des répétitions de séquence, combinant différents niveaux (forme graphique, lemme, partie du discours, marque sémantique) qui a permis d'accélérer la mise au point manuelle des patrons. Notre approche a été bonifiée par deux prétraitements : un prétraitement linguistique des énoncés définitoires, afin de les canoniser, en éliminant les informations accessoires à la prédication définitoire principale ; une décomposition des contextes à étudier statistiquement selon l'information recherchée, afin de réduire le bruit de calculs statistiques bruts sur corpus. Cette combinaison de techniques nous semble prometteuse pour tout type de relations sémantiques, même si elle doit être éprouvée sur corpus libre, et étendue à d'autres types d'informations sémantiques.

Du point de vue de la structuration sémantique du lexique, les résultats obtenus montrent que le réseau lexical ne se limite pas aux relations classiques, « classificatoires » (hyperonymie-hyponymie, synonymie, antonymie) qui sont le cœur d'un réseau lexical de type Wordnet : les définitions explicitent certes clairement ces relations de catégorisation, mais d'autres relations relient les lexies entre elles : méronymie-holonymie, fonction, et prédications définitoires qui décrivent différents aspects essentiels. Ces derniers éléments sont peut-être la matière la plus intéressante du corpus étudié ici, car elles expriment des propriétés essentielles des lexies, mais, à ce stade de l'étude, elles ne peuvent pas être catégorisées.

Cette étude nous engage à mener des travaux complémentaires dans différentes directions : d'une part, préparer une évaluation à plus grande échelle, afin de valider les relations sémantiques extraites et fournir à la communauté scientifique un corpus de référence plus conséquent encore. Cette évaluation pourra être conduite à la fois par une validation collaborative en ligne, mais également en étudiant les relations sémantiques sur gros corpus à partir de l'hypothèse distributionnelle.

Remerciements

Merci aux évaluateurs de leurs commentaires et suggestions, qui ont permis de grandement bonifier cet article.

Références

ALARCÓN R., BACH C. AND SIERRA G. (2008) “Extracción de contextos definitorios en corpus especializados: Hacia una elaboración de una herramienta de ayuda terminográfica”. *Revista Española de Lingüística*. Madrid, 2008. 247-278.

APIDIANAKI M. ET SAGOT B. (2014) Data-driven Synset Induction and Disambiguation for Wordnet Development. *Language Resources and Evaluation Journal*, Springer Netherlands, Vol. 48(4), pp. 655-677.

ARNAULD A. ET NICOLE P. (1662) *La logique ou l’art de penser*, édition critique par D. Descotes, Paris: Champion, 2011.

AUER S, BIZER C, KOBILAROV G, LEHMANN J, IVES Z (2007) DBpedia: A nucleus for a web of open data. In: *Proceedings of 6th International Semantic Web Conference*, Springer, Busan, Korea, pp 11–15

BACH, N., AND BADASKAR S. (2007) “A Review of Relation Extraction,” Literature review for Language and Statistics II, 2007

BAKER, C.F., FILLMORE C.J., AND LOWE J.B. (1998) The Berkeley FrameNet project." COLING-ACL '98: Proceedings of the Conference. Montreal, Canada 1998. 86-90.

BARONI, M., AND LENCI A. (2010) “Distributional Memory : A General Framework for Corpus-Based Semantics,” Computational Linguistics, 36-4 (2010), 50

BARQUE L., NASR A., POLGUÈRE A. (2010) From the Definitions of the Trésor de la Langue Française to a Semantic Database of the French Language, Proceedings of the 14th EURALEX International Congress, Leewarden.

BATTISTELLI D. (2009) *La temporalité linguistique : circonscrire un objet d'analyse ainsi que des finalités à cette analyse*. HDR, Université de Nanterre - Paris X, 2009.

BÉCHET N., CELLIER P., CHARNOIS T., CRÉMILLEUX B., QUINIOU S. (2013). SDMC : un outil en ligne d'extraction de motifs séquentiels pour la fouille de textes. Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC'13), Jan 2013, Toulouse, France.

BOLLACKER K, EVANS C, PARITOSH P, STURGE T, TAYLOR J (2008) Freebase: A collab- oratively created graph database for structuring human knowledge. In: Pro- ceedings of the 2008 ACM SIGMOD International Conference on Management of Data, ACM, New York, NY, USA, SIGMOD '08, pp 1247–1250

BUNESCU R, MOONEY R (2006) Subsequence kernels for relation extraction. In: Weiss Y, Scholkopf B, Platt J (eds) Advances in Neural Information Processing Systems 18, MIT Press, Cambridge, MA, pp 171–178

CANDITO, M. AMSILI, P., BARQUE, L., BENAMARA, F., CHALENDAR, G., DJEMAA, M., HAAS, P., HUYGHE, R., MATHIEU, Y., MULLER, P., SAGOT, B. & VIEU, L., (2014), Developing a French FrameNet: methodology and first results, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland, 2014

CARTIER E. (2004) : Repe rage automatique des expressions de finitoires : mode lisation de l'information de finitoire, me thode d'exploration contextuelle, me thodologie de de veloppement des ressources linguistiques, description des expressions du franc ais contemporain, mise en oeuvre informatique, thèse de doctorat, Université Paris IV-Sorbonne.

CLARK, S. (2015) "Vector Space Models of Lexical Meaning," in *Handbook of Contemporary Semantics*, second edition, ed. by Shalom Lappin and Chris Fox (Wiley-Blackwell, 2015), pp. 1–43

CULOTTA A, MCCALLUM A, BETZ J (2006) Integrating probabilistic extraction mod- els and data mining to discover relations and patterns in text. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Association for Computational Linguistics, New York, New York, pp 296–303

DESCLÈS J.-P. (2006) «Contextual Exploration Processing for Discourse Automatic Annotations of Texts», FLAIRS 2006, Melbourne, Floride, 11-13 mai, Invited Speakers, p. 281-284.

FILLMORE, C. J. (1982) Frame semantics. Linguistics in the Morning Calm. Seoul, South Korea: Hanshin Publishing Co., 1982. 111-137

FIRTH, J. R. (1957). A synopsis of linguistic theory 1930–1955. In Studies in Linguistic Analysis, pp. 1–32. Blackwell, Oxford.

GARCIA, D. (1998), Analyse automatique des textes pour l'organisation causale des actions, Re alisation du syste me Coatis. The se d'informatique, Universite Paris IV.

GERGONNE J. (1818) « Variétés, essai sur la théorie des dé finitions », *Annales de Mathématiques pures et appliquées*, tome 9 (1818-1819), p.1-35.

HANKS, P. (2013) *Lexical Analysis: Norms and Exploitations*. Cambridge. MIT Press.

HARRIS, Z. (1954). Distributional structure. Word, 10(23), 146–162.

HEARST M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the 15th International Conference on Computational Linguistics (COLING 1992), p. 539–545, Nantes.

JACKIEWICZ A. (1999), « La causalité dans les textes », in Semantyka i kontrontacja jezykowa(Sémantique et Confrontation des Langues), Varsovie, Pologne, SOW, Académie des Sciences de la Pologne, 1999, vol.2, pp.147-164

KAMBHATLA N. (2004) Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In: Proceedings of the ACL 2004, Association for Computational Linguistics, Morristown, NJ, USA.

KIELA D. AND CLARK S. (2014) A Systematic Study of Semantic Vector Space Model Parameters, in *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, 2014, pp. 21–30

KILGARRIFF, A., RYCHLY P., SMRZ P., AND TUGWELL D.. (2004). The Sketch Engine. In Proceedings of Euralex, pages 105–116, Lorient.

KLAVANS J. AND MURESAN S. (2001) "Evaluation of the DEFINDER System for Fully Automatic Glossary Construction". In Proceedings of the American Medical Informatics Association Symposium. ACM Press, New York, 2001. 252- 262

KOZAREVA Z, HOVY E (2010) Learning arguments and supertypes of semantic relations using recursive patterns. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Uppsala, Sweden, pp 1482–1491

KOZAREVA Z, RILOFF E, HOVY E (2008) Semantic class learning from the web with hyponym pattern linkage graphs. In: Proceedings of ACL-08: HLT, Association for Computational Linguistics, Columbus, Ohio, pp 1048–1056

LEU F. AND KO C. (2007) "An Automated Term Definition Extraction using the Web Corpus in Chinese Language". In Proceedings of the Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'07), 2007. 435-440.

LUX-POGODALLA V. & POLGUÈRE A. (2011). Construction of a French Lexical Network : Methodological Issues. In Proceedings of the First International Workshop on Lexical Resources, WoLeR 2011. An ESSLLI 2011 Workshop, p. 54–61, Ljubljana.

MARTIN R. (1992) *Pour une logique du sens*, Paris, P.U.F., coll. Linguistique nouvelle, 2e édition revue et augmentée, 1992

MATHIEU-COLAS M. (1996) « Essai de typologie des noms composés français », Cahiers de lexicologie, 69, 1996-II, pp. 71-125.

MEYER I. (2001) "Extracting Knowledge-rich Contexts for Terminography". In Recent Advances in Computational Terminology. D. Bourigault, C. Jacquemin and M.C. L'Homme (eds.).. John Benjamin's, Amsterdam, 2001. 278- 302.

MILLER G. A., BECKWITH R., FELLBAUM C., GROSS D. & MILLER K. J. (1990). Introduction to WordNet : An On-line Lexical Database. International Journal of Lexicography, 3(4), 235–244.

MILLER S, FOX H, RAMSHAW L, WEISCHEDEL R (2000) A novel use of statistical parsing to extract information from text. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, Morgan Kaufmann Publishers Inc., Seattle, Washington, pp 226–233

MILLER, G. AND CHARLES W. (1991) « Contextual correlates of semantic similarity. » *Language and Cognitive Processes*, 6:1–28

MORIN E. AND JACQUEMIN C. (2004). Automatic Acquisition and Expansion of Hypernym Links. Computers and the Humanities (CHUM), Kluwer, 38(4), 363–396.

MOUTON, C. AND DE CHALENDAR, G. (2010). JAWS: Just Another WordNet Subset. In Proc. of TALN'10, Montreal, Canada.

NASTASE V., NAKOV P., SÉAGHDHA D.O., AND SZPAKOWICZ S. (2013), *Semantic Relations Between Nominals*, Synthesis Lectures on Human Language Technologies, April 2013, Vol. 6, No. 1 , Pages 1-119

NAVIGLI R. AND S. PONZETTO (2012) BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, Elsevier, 2012, pp. 217-250

NAVIGLI R. AND P. VELARDI (2010) Learning Word-Class Lattices for Definition and Hypernym Extraction. Proc. of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Uppsala, Sweden, July 11-16, 2010, pp. 1318-1327

NAVIGLI R. AND VELARDI P. (2007) "GlossExtractor: A Web Application to Automatically Create a Domain Glossary". In Lecture Notes in Computer Science 4733, 2007. 339-349

PANTEL, P. AND PENNACCHIOTTI M.. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In Proceedings of 44th AnnualMeeting of the Association for Computational Linguistics joint with 21st Conference on Computational Linguistics (COLING-ACL), pages 113–120, Sydney

PEARSON J. (2001) Terms in Context. John Benjamin's, Amsterdam.

POLGUÈRE, A. (2014) "Principes de Modélisation Systémique Des Réseaux Lexicaux," in 21ème Conférence TALN, pp. 79–90

RAMISCH C. (2015) "Multiword Expressions Acquisition: A Generic and Open Framework", Theory and Applications of Natural Language Processing series XIV, Springer, ISBN 978-3-319-09206-5, 230 p., 2015.

REBEYROLLE, J. (2000) Forme et fonction de la définition en discours, Thèse de doctorat, Université Toulouse-Le Mirail.

SAGOT B. & FISER D. (2008). Construction d'un WordNet libre du français à partir de ressources multilingues. In Actes de TALN 2008 (Traitement automatique des langues naturelles), Avignon : LIA.

SAGOT, B., AND FIŠER D. (2011), "Extending Wordnets by Learning from Multiple Resources," LTC'11: 5th Language and Technology Conference, 2011

SAGOT B. ET FIŠER D. (2012). Cleaning noisy wordnets. In Proceedings of LREC 2012, Istanbul, Turquie

SAJOUS, F., HATHOUT N., CALDERONE B. (2014), "Ne Jetons Pas Le Wiktionnaire Avec L'Oripeau Du Web ! Études et Réalisations Fondées Sur Le Dictionnaire Collaboratif," 8 (2014), 663–680

SCHMID H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

SERRA, G. (2009) "Extraccion de Contextos Definitorios En Textos de Especialidad a Partir Del Reconocimiento de Patrones Linguisticos," *Linguamatica*, 2009, 13–38

SIMKO J. ET BIELIKOVA M. (2014) *Semantic Acquisition Games :Harnessing Manpower for Creating Semantics*, Springer International Publishing Switzerland,

STORRER A. AND WELLINGHOFF S. (2006) "Automated Detection and Annotation of Term Definitions in German Text Corpora". In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06). Genève, 2006. 2373- 2376.

SUCHANEK F.M., KASNECI G, WEIKUM G (2007) YAGO: A core of semantic knowl- edge. In: Proceedings of WWW-07, pp 697–706

TURNEY, P.D., AND P. PANTEL. (2010) "From Frequency to Meaning: Vector Space Models of Semantics." Journal of Arti ficial Intelligence Research 37 (1): 141–188.

VOSSEN, P. (1998) EuroWordNet: Building a Multilingual Database with Wordnets for European Languages. In: K. Choukri, D. Fry, M. Nilsson (eds), The ELRA Newsletter, Vol3, n1, 1998. ISSN: 1026-8200.

WALTER S. AND PINKAL M. (2006) "Automatic Extraction of Definitions from German Court Decisions". In Proceedings of the Workshop on Information Extraction Beyond the Document. 21st International Conference on Computational Linguistics (COLING'2006). Sydney, 2006. 20–28.

ZHAO S, GRISHMAN R (2005) Extracting relations with integrated information using kernel methods. In: ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, pp 419–426