Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques

Cécile Fabre, Didier Bourigault

Université Toulouse-Le Mirail – ERSS, CNRS
{cecile.fabre ; didier.bourigault}@univ-tlse2.fr

Résumé

Nous étudions les relations de proximité sémantique entre les noms et les verbes à partir de données calculées sur un corpus de 200 millions de mots par un programme d'analyse distributionnelle automatique. Nous exposons les résultats d'une méthode d'extraction de couples Nom/Verbe, qui combine un indice de proximité distributionnelle et un indice de cooccurrence : un couple est extrait si le nom et le verbe apparaissent avec les mêmes arguments sur l'ensemble du corpus, d'une part, et s'ils apparaissent au moins une fois dans un même paragraphe munis du même argument, d'autre part. L'article élabore une typologie des 1441 couples extraits et démontre l'intérêt de prendre en compte les couples non liés morphologiquement, qui constituent 70 % des données.

Mots-clés : relations sémantiques, ressources lexicales, analyse distributionnelle.

Abstract

In this paper, we study the semantic relations that hold between nouns and verbs. We benefit from the data provided by Upery, a program that automatically extracts word associations from a 200 million words corpus by means of distributional analysis. We present the results of an experiment in which noun-verb associations are extracted by crossing two criteria : distributional proximity and cooccurrence. The 1441 couples share the same arguments in the corpus and appear at least once with the same argument within the same paragraph. We present a typology of these noun-verb couples, showing the necessity to take into account non-mophologically related couples which amounts to 70 % of the data.

Keywords: semantic relations, lexical resources, distributional analysis.

1. Introduction

Le fait que les ressources lexicales se concentrent généralement sur le recensement des relations sémantiques entre mots de même catégorie morphosyntaxique est souvent signalé comme une limite (Vossen, 1998 ; Claveau, 2003). Ainsi, un des objectifs du projet EuroWordNet avait été de jeter des ponts entre les catégories, de manière à identifier des liens entre des *synsets* que le critère catégoriel isolait les uns des autres. Pour l'essentiel, ce projet s'est traduit dans les faits par la seule mention des relations intercatégorielles confortées par une ressemblance morphologique (ex. : *adorn* / *adornment*), qu'elles soient de type synonymique ou antonymique, sans qu'un travail de description extensif ait été, à notre connaissance, mené dans ce but. Nous voulons aller plus loin dans cette voie, en nous intéressant aux relations de proximité sémantique entre les noms et les verbes dans une perspective plus large, non limitée aux proximités morphologiques. Nous cherchons à établir des liens de nature paradigmatique (liens de substitution, d'équivalence) entre des unités lexicales pour lesquelles ce type de lien n'est généralement pas envisagé, du fait, précisément, de leur appartenance à des catégories distinctes. L'étude des rapprochements entre noms et verbes se limite en effet très souvent au niveau syntagmatique, à travers la question des collocations (Wanner, 2004). Pour remplir cet objectif, nous bénéficions des résultats fournis

par un programme d'analyse distributionnelle, à partir de l'analyse d'un corpus de 200 millions de mots. Nous reprenons l'hypothèse harrissienne selon laquelle la proximité distributionnelle – et, plus précisément, le partage d'arguments syntaxiques – constitue un indice de proximité sémantique, et nous nous proposons de caractériser les relations Nom/Verbe qui émergent d'une telle analyse. Cette étude vise à évaluer l'intérêt d'une approche distributionnelle comme outil d'aide à la description des relations lexicales intercatégorielles, ainsi que, d'un point de vue plus applicatif, comme moyen d'étendre au cas des relations intercatégorielles le repérage de zones textuelles sémantiquement proches.

Après avoir fait le point sur les tentatives de description dans le domaine du TAL des relations sémantiques entre noms et verbes (section 2), nous présentons dans la section 3 Upery, dispositif d'acquisition de données distributionnelles, les résultats qu'il fournit sur un corpus de très grande taille, ainsi que la méthode de sélection contextuelle que nous avons élaborée pour filtrer au plus près une liste de couples Nom/Verbe. Cette méthode croise deux critères : un nom et un verbe sont rapprochés si un nombre suffisant de mots apparaissent à la fois comme tête de l'argument objet du verbe (ex. : *écourter le mandat*) et comme tête du complément prépositionnel du nom (ex. : *raccourcissement du mandat*) ; en outre, le nom et le verbe doivent apparaître munis d'un argument commun dans au moins un paragraphe du corpus. Le critère de cooccurrence sert donc à confirmer le critère distributionnel. Ce filtrage par les cooccurrences s'inspire de méthodes mises au point pour la validation d'informations morphologiques (Hathout et Tanguy, 2005). Dans la section 4, nous présentons une première typologie de ces couples, dans le but d'évaluer quelle proportion s'apparente à une relation de proximité sémantique.

2. Proximité Nom/Verbe, enjeux en TAL et en linguistique

Dans le champ du Traitement Automatique des Langues, notre objectif est la construction de ressources lexicales susceptibles de contribuer aux travaux sur le repérage de la paraphrase ou plus généralement de la variation sémantique dans les systèmes d'accès à l'information (Ferret *et al.*, 2001 ; Barzilay et Mac Keown, 2001). Le travail que nous décrivons ici doit permettre d'établir un lien entre des zones de textes de contenu équivalent mais qui font appel l'une à une expression nominale, l'autre à une expression verbale (ex. : *anniversaire du débarquement* / *célébrer le débarquement*). Comme nous l'avons dit en introduction, le repérage de ces reformulations intercatégorielles est un des objectifs que P. Vossen assignait à EuroWordnet : « From an information retrieval point of view the same information can be coded in an NP or in a sentence. By unifying higher-order nouns and verbs in the same ontology it will be possible to match expressions with very different syntactic structures but comparable content. » (Vossen, 1998). L'expérience que nous décrivons ici fait suite à plusieurs projets antérieurs de prise en compte de la variation verbo-nominale : dans (Fabre et Jacquemin, 2000), le but était de constituer une grammaire permettant d'identifier les reformulations verbales de termes nominaux, de manière à étendre les types de variation couverts par Fastr. Il s'agissait donc de contrôler la proximité sémantique des deux formulations, de manière à récolter comme variantes les zones verbales jugées sémantiquement équivalentes au terme nominal (ex. : *technique de mesure* / *mesurer à l'aide d'une technique*), et à rejeter celles qui ne présentaient pas de proximité sémantique avec le terme de départ (ex. : *amélioration de technique* / *technique permettant d'améliorer*). Dans (Bouillon *et al.*, 2000), il s'agissait d'extraire, par des techniques d'apprentissage, les verbes sémantiquement associés à un nom, de manière à identifier pour un nom donné les verbes désignant les activités typiques dans lesquelles il est impliqué (ex. : *écrou* / *serrer*, *prise* /

brancher). Le projet final et l'exploitation de ce type de ressources en recherche d'information sont décrits dans (Claveau, 2003).

Parallèlement, en linguistique théorique et descriptive et en psycholinguistique, des travaux cherchent à dépasser des cloisonnements traditionnels entre catégories, dans la mesure où ceux-ci n'auraient pas de réalité du point de vue fonctionnel. Nous poursuivons en particulier l'objectif formulé par (Nespoulous et Virbel, 2003) de sortir de la dichotomie nom-verbe pour étendre l'observation de relations d'équivalence en discours à des entités que la description grammaticale traditionnelle isole les unes des autres. La porosité de la frontière entre catégories en ce qui concerne les fonctionnements lexicaux a été montrée par (Pustejovsky, 1995) à travers le dispositif de la structure *qualia* ; elle est également attestée par (Murphy, 2003) : « […] if lexical relations relate word-concepts, rather than lexical entries, the relations can utilize information about words that transcends the lexical boundaries among words. ». Jusqu'à présent, lorsque l'on sort du champ balisé des relations sémantiques intracatégorielles, pour lesquelles on dispose d'outils de description éprouvés, et que l'on s'intéresse à la proximité sémantique de mots de catégories différentes, c'est en général seulement la proximité d'ordre formel, morphologique qui permet de poser des passerelles entre les catégories. La proximité entre les verbes et les noms, lorsqu'elle s'appuie sur une ressemblance morphologique, a été largement décrite, et certaines ressources lexicales les prennent en compte. C'est aussi le cas d'EuroWordnet qui code certaines « relations intercatégorielles explicites » (*Explicit Cross-Part-Of-Speech relations*), et, dans une démarche plus modeste mais plus exhaustive, de *Verbaction*[1]. Il reste alors un vaste champ à explorer, celui des relations intercatégorielles sans lien morphologique. Certaines fonctions lexicales du *Dictionnaire Explicatif et Combinatoire* (Mel'čuk *et al.*, 1995) permettent de les coder (fonctions de dérivation syntaxique et sémantique S_0, V_0, S_i), mais leur étude systématique et la mise au point de méthodes de repérage dans les textes pour leur exploitation en TAL restent à mener.

3. Technique d'extraction des couples Nom/Verbe

3.1. Méthode d'analyse distributionnelle : Upery

L'outil d'analyse distributionnelle Upery (Bourigault, 2002) s'appuie sur les résultats de l'analyseur syntaxique de corpus Syntex (Bourigault et Fabre, 2000 ; Bourigault *et al*., 2005). Syntex prend en entrée un corpus de phrases étiquetées, et calcule pour chaque phrase les relations de dépendance syntaxique entre les mots (sujet, complément d'objet, complément prépositionnel, épithète, etc.). À partir de l'analyse syntaxique, Upery extrait des triplets (gouverneur, relation, dépendant). Par exemple, de l'analyse syntaxique de la phrase *« il mange la pomme »*, il extrait le triplet *(manger, obj, pomme)*. Au cours de cette étape d'extraction de triplets, sont effectuées automatiquement un certain nombre de normalisations syntaxiques, visant soit à passer de relations syntaxiques de surface à des relations syntaxiques plus profondes, soit à expliciter certaines relations :

- coordination : *Jean mange la pomme et la poire → (manger, obj, pomme) ; (manger, obj, poire)*
- passif : *La pomme a été mangée → (manger, obj, pomme)*

[1] Le lexique Verbaction a été élaboré par Nabil Hathout (ERSS) à partir de la nomenclature du *Trésor de la Langue Française*. Il comporte environ 9000 couples Nom/verbe : http://www.univ-tlse2.fr/erss/ressources/verbaction/main.html

- antécédence relative : *Jean qui dort → (dormir, suj, Jean)*
- verbe à contrôle de l'infinitif : *Jean décide de nager → (nager, suj, Jean) ; Jean ordonne à Marie de travailler → (travailler, suj, Marie)* ; etc.
- préposition : *il mange avec une fourchette → (manger, avec, fourchette)*
- expansion des syntagmes nominaux : *il mange une pomme rouge → (manger, obj, pomme) et (manger, obj, pomme rouge)*

La méthode d'analyse distributionnelle que nous développons distingue deux types d'unités : les prédicats et les arguments. À chaque triplet extrait *(gouverneur, relation, dépendant)* correspond un couple *(prédicat, argument)* : le prédicat est constitué du gouverneur auquel on « accole » la relation, et l'argument est le dépendant. Ainsi, à partir du triplet *(manger, obj, souris)*, on construit le couple *(manger_obj, souris)*. On procède ensuite à une double analyse distributionnelle : on rapproche les prédicats qui se construisent avec les mêmes arguments et on rapproche les arguments qui se construisent avec les mêmes prédicats. Par exemple, à partir du corpus LM10 (cf. section 3.2), les prédicats *sculpter_obj* et *modeler_obj* sont rapprochés parce qu'ils se construisent de façon régulière avec les arguments *visage, corps, espace, personnage, forme, image*. De la même façon, les arguments *tempête* et *tornade* sont rapprochés parce qu'ils se construisent avec les prédicats *emporter_suj*, *déclencher_obj*, *résister_à*, *effet_de*, *provoquer_obj*. Nous disons que *(sculpter_obj, modeler_obj)* est un couple de « voisins prédicats » et que *(tempête, tornade)* est un couple de « voisins arguments ». La proximité entre prédicats (resp. arguments) est calculée à partir de la mesure du Jaccard, qui fait intervenir le nombre d'arguments (resp. prédicats) partagés par les deux prédicats (resp. arguments).

3.2. Extraction de couples Nom/Verbe à partir du corpus LM10

Dans cette étude, nous exploitons les résultats fournis par cette méthode d'analyse distributionnelle appliquée à un corpus journalistique de très grande taille, constitué des articles du journal *Le Monde* couvrant la période 1991-2000 (corpus LM10[2]), soit environ une centaine de milliers d'articles. La taille de ce corpus est d'environ 200 millions de mots. Le nombre de couples prédicat/arguments de fréquence supérieure ou égale à 5 est de 2 987 284, et le nombre de couples de voisins prédicats et de voisins arguments calculés à partir de ces couples dont le coefficient de Jaccard est supérieur à 0.1 est de 6 183 862. Dans cette étude, nous nous focalisons sur les couples de prédicats qui impliquent un nom et un verbe, puisque l'étude des proximités Nom/Verbe n'a de sens que pour ce qui concerne les noms en situation prédicative, munis de compléments, et dont le fonctionnement s'apparente alors à celui du verbe. Nous nous intéressons plus précisément aux cas où la relation associée au prédicat nominal est la préposition *de* et la relation associée au prédicat verbal est la relation *obj*[3]. Nous retenons les couples tels que :

- le prédicat verbal apparaît parmi les 10 premiers voisins du prédicat nominal
- le prédicat nominal apparaît parmi les 100 premiers voisins du prédicat verbal.

On obtient ainsi un ensemble de 3 667 couples. Nous donnons dans le tableau 1 deux exemples de couples de voisins prédicats impliquant un nom et un verbe. Le premier couple

[2] Le corpus LM10 a été construit à partir de ressources obtenues auprès de l'agence ELDA, à l'aide de programmes de nettoyage et de balisage réalisés par B. Habert du LIMSI, que nous remercions ici.

[3] Ce choix est le résultat d'observations empiriques : les rapprochements sur la base de relations prépositionnelles au niveau du verbe sont marginales. Seule la relation Nom_de/V_suj est quantitativement équivalente à la relation Nom_de/V_obj, mais on constate qu'elle est nettement plus bruitée.

réunit un verbe et son déverbal d'action, alors que dans le second couple, le lien morphologique est absent.

(protection_de, protéger_obj)	*kurde de Irak, intérêt financier, population kurde, espace naturel, espèce animal, convoi humanitaire, patrimoine naturel, santé humain, opération humanitaire, …*
(défense_de, protéger_obj)	*intérêt fondamental, intérêt vital, identité culturel, acquis social, droit de personne, intérêt national, droit de minorité, actionnaire minoritaire, peuple serbe, faible, …*

Tableau 1. Exemples de couples de prédicats impliquant un Nom et Verbe, avec la liste des arguments (lemmatisés) partagés

3.3. Filtrage par cooccurrence

La liste de 3 667 couples ainsi constituée sur la base de critères distributionnels pourtant relativement sévères comporte encore un grand nombre d'intrus. En effet, le partage de dépendants identiques n'est pas toujours un critère suffisant pour détecter une proximité sémantique. Un cas de figure en particulier contrarie cet objectif : lorsque les dépendants forment une classe trop restreinte, le voisinage porte sur des aspects du sens trop périphériques. Par exemple pour le rapprochement (*carcasse_de*, *intercepter_obj)*, la liste des dépendants partagés est : {*véhicule*, *camion*, *voiture*, *navire*, *bateau*}. Ces deux prédicats n'ont en commun que le fait de pouvoir s'appliquer à des véhicules. Un filtrage supplémentaire s'impose, pour améliorer la précision de la collecte.

L'idée que nous avons mise en œuvre dans cette étude est celle d'un filtrage des voisins distributionnels (cooccurrence du second ordre) par une contrainte de cooccurrence dans une même zone textuelle (cooccurrence du premier ordre). On retient un couple de prédicats si les *deux prédicats apparaissent ensemble dans au moins un paragraphe, avec le même argument* (en tant que dépendant). Notons d'emblée qu'il faut distinguer deux cas de figure, selon que les deux prédicats partagent la même occurrence du mot argument (cooccurrence liée) ou apparaissent chacun avec une occurrence différente du mot argument (cooccurrence disjointe). Quelques exemples de cas de cooccurrence sont donnés dans le tableau 2. Ce filtrage par le contexte textuel permet de retenir 1 441 couples de prédicats. Ce sont ces couples que nous analysons dans la section suivante.

Outre qu'il permet de sélectionner des couples *a priori* plus pertinents, le grand intérêt de ce mode de filtrage contextuel est de nous permettre de projeter et de visualiser les couples dans des contextes restreints, de manière à guider l'évaluation des relations sémantiques mises au jour. Ce n'est en effet parfois qu'en examinant les usages discursifs de ces unités lexicales que l'on est à même de caractériser les relations sémantiques associées. C'est le cas par exemple pour le couple *(image_de, voir_obj)*, dont nous n'aurions pas prédit qu'il donne accès dans certains cas à des zones textuelles très proches sémantiquement (cf. tableau 2).

Cooccurrence disjointe	
(adepte_de, pratiquer_obj) argument partagé : *marche*	*Les Finlandais et les Suédois sont en effet de fervents* ***adeptes*** *de la marche à pied. Ils distancent dans ce domaine leurs voisins européens. Et ce à tout âge. Dans chacun de ces deux pays, 76 % et 74 % des plus de 55 ans, notamment****, pratiquent*** *la marche au moins une fois tous les quinze jours, contre 40 % en moyenne dans l'Union.*

(raccourcissement_de, écourter_obj) argument partagé : *mandat*	*M. Mitterrand n'a pas eu l'occasion, dimanche 12 avril, de donner son avis sur un éventuel **raccourcissement** du mandat présidentiel. Il a, en revanche, rappelé qu'élu pour sept ans, c'était à lui seul d'apprécier s'il devait volontairement **écourter** son mandat dans le cas où il estimerait n'avoir plus " la capacité " de l'exercer.*
(image_de, voir_obj) argument partagé : *métier*	*Et surtout l'**image** de beaucoup de métiers industriels reste peu attractive. […] Familles et enseignants **voient** encore trop souvent les métiers de la production avec les yeux de Zola.*
Cooccurrence liée	
(greffe_de, prélever_obj) argument partagé : *moelle*	*Cette nouvelle procédure thérapeutique consistait, schématiquement, à administrer une chimiothérapie à très forte dose impliquant une toxicité dangereuse dont les effets secondaires ne pouvaient être combattus que par une **greffe** de moelle osseuse. **prélevée** auparavant sur le malade*
(signature_de, conclure_obj) argument partagé : *accord*	*La **signature** de cet accord de paix, **conclu** à Estoril, est de bon augure.*

Tableau 2. Exemples de cooccurrences de couples de prédicats

4. Typologie des couples extraits

Nous pouvons maintenant entamer l'étude précise, en contexte, des 1 141 couples ainsi sélectionnés. Nous estimons qu'une telle étude est en soi riche d'enseignements sur le plan linguistique, et qu'elle doit précéder toute tentative d'évaluation de l'exploitation de ces couples dans une application TAL particulière, comme l'expansion de requête en recherche d'information. Nous proposons dans cette section une première typologie des couples sélectionnés.

4.1. Part du lien morphologique

31 % des couples (soit 455 sur 1441) correspondent à une relation entre un nom déverbal et le verbe morphologiquement associé. Parmi ceux-ci, 92 % sont du type déverbal d'action/verbe, comme (*réduction_de, réduire_obj*), les autres cas correspondant pour l'essentiel à une association entre un verbe et son déverbal d'agent, comme (*lecteur_de*, *lire_obj*). La localisation au sein d'un même paragraphe du verbe et de son déverbal munis du même dépendant permet de repérer des variantes sémantiques (ex. : *augmentation de salaire* / *augmenter le salaire*), autrement dit des zones de reformulation. L'extrait présenté dans le tableau 3 illustre le cas où le titre d'un article comporte une expression verbale qui est reprise sous une forme nominale dans la première phrase du corps de l'article (relation discursive d'élaboration).

*CAPITAL BSN **assouplit** son projet de protection du capital* *Dans un entretien aux Échos du 11 septembre, M. Antoine Riboud, président de BSN, a annoncé un **assouplissement** de son projet de protection contre une offre publique d'achat (OPA).*

Tableau 3. Cooccurrence du couple (assouplissement_de, assouplir_obj)

4.2. Typologie des couples Nom/Verbe sans lien morphologique

Plus des 2/3 des couples extraits par Upery, puis filtrés contextuellement, sont des couples dans lesquels le nom et le verbe ne présentent aucune ressemblance formelle. C'est bien sûr le cas lorsque le nom est morphologiquement non construit *(patron_de, diriger_obj)*, mais en position prédicative ces noms sont très minoritaires. On observe donc surtout deux cas de figure. Dans le premier, le nom déverbal est lié à une série de verbes dans laquelle sa base verbale figure, souvent au premier rang. C'est le cas du prédicat nominal *attribution_de* dans le tableau 4. Dans ce cas, la proximité du nom avec le verbe morphologiquement associé est réelle, mais elle est loin d'épuiser l'ensemble des relations de proximité que le nom entretient avec des verbes. Dans le second cas, le nom déverbal est lié à un ou plusieurs verbes, *mais pas* au verbe qui lui est morphologiquement lié. C'est le cas du prédicat nominal *hausse_de* dans le tableau 4. C'est le cas aussi du prédicat *vainqueur_de* qui est associé à *gagner_obj* et *remporter_obj*, mais pas à *vaincre_obj*. Ces données nous permettent donc d'observer des proximités sémantiques réellement attestées dans les usages.

Le tableau 5 indique de quelle manière se répartissent les relations sémantiques que nous avons identifiées au sein des couples sans lien morphologique. Ces estimations chiffrées portent sur l'examen des 200 premiers couples extraits (par ordre décroissant d'occurrences en nombre de paragraphes dans le corpus), observés chacun dans les dix premiers paragraphes dans lesquels ils cooccurrent.

Voisins de *attribution_de*	nb de paragraphes	Voisins de *hausse_de*	nb de paragraphes
attribuer_obj	50	*augmenter_obj*	105
accorder_obj	20	*baisser_obj*	54
distribuer_obj	8		
allouer_obj	5		
octroyer_obj	1		

Tableau 4. Voisins verbaux du prédicat nominal attribution_de

Synonymie	Antonymie	Actant / activité typique	Phases successives	Catégorisation	Collocations / expressions figées	Erreurs[4]
15 %	**15 %**	**10 %**	**13 %**	**8 %**	**20 %**	**19 %**

Tableau 5. Typologie des couples Nom/Verbe sans lien morphologique

15 % des couples sans lien morphologique peuvent être considérés comme synonymes et permettent de repérer à l'examen des textes des zones de reformulation *(relèvement_de, augmenter_obj)*, *(octroi_de, attribuer_obj)*. On trouve dans les mêmes proportions des antonymes, comme *(violation_de, respecter_obj)* ou *(reprise_de, suspendre_obj),* parmi lesquels on identifie des rapports de réciprocité (*obtention_de*, *délivrer_obj*). Cette proportion vérifie la remarque de C. Fellbaum citée par (Murphy, 2003) : *"[r]egardless of their syntactic*

[4] Nous regroupons dans cette catégorie les résultats erronés dûs à des erreurs d'étiquetage et d'analyse syntaxique, ainsi que les couples donnant lieu à des relations de sens trop hétérogènes pour pouvoir être caractérisés (mots polysémiques, verbes supports, etc.).

category, words expressing semantically opposed concepts tend to be used together in the same sentences". Le rapport actants/activité typique, sur lequel se focalisait l'étude de (Bouillon *et al.*, 2000) pour sa propension à renvoyer également des zones d'équivalence sémantique compte pour 10 % des couples (*responsable_de, diriger_obj*), *(marché_de/vendre_obj)*. 13 % des couples comportent des éléments qui désignent des phases successives d'une activité *(signature_de, conclure_obj), (restitution_de, confisquer_obj)*. Enfin, pour 8 % des couples, le nom et le verbe n'ont en commun que de se rapporter à des dépendants de même catégorie : par exemple, le couple *(entrée_de, situer_obj)* est rapproché parce que les deux prédicats s'appliquent à des bâtiments.

Au total, 3 des 5 relations que nous avons identifiées (synonymie, activité typique, succession temporelle), soit 38 % des couples sans lien morphologique) sont susceptibles de permettre le repérage de zones de textes comparables.

Pour 20 % des couples (catégorie collocation / figement), on ne peut que constater les limites de l'analyse distributionnelle qui ne permet pas de dissocier les couples entretenant des relations d'ordre paradigmatique de ceux qui sont pris dans des associations de nature simplement collocative. Néanmoins, l'examen des contextes de cooccurrence du nom et du verbe pour ces couples montrent qu'ils apparaissent exclusivement ou très majoritairement dans des contextes de cooccurrence liée (cf. section 4.3). Par exemple, pour le couple *(levée_de, décréter_obj)*, quand le prédicat nominal et le prédicat verbal apparaissent dans le même paragraphe, c'est systématiquement avec la même occurrence du mot argument (***levée*** *de l'*<u>*embargo*</u> ***décrété*** *par*, ***levée*** *des* <u>*sanctions*</u> ***décrétées*** *par*). Ces couples pourraient donc être éliminés de façon automatique.

5. Conclusion

Ces premiers résultats montrent l'intérêt de combiner une approche de type distributionnel, à travers le recueil systématique d'informations sur l'ensemble du corpus, et une approche contextuelle, via la projection sur le corpus des couples extraits et une étude de leurs cooccurrences proches. En effet, l'analyse distributionnelle à elle seule ne permet pas de distinguer des relations Nom/Verbe de type syntagmatique (collocations) et paradigmatique. L'examen des contextes ouvre une voie pour une meilleure différenciation de ces types d'association. Une perspective immédiate de notre travail consiste à vérifier la corrélation que nous avons commencé à observer entre l'apparition du nom et du verbe dans des zones de cooccurrence disjointe et leur aptitude à apparaître dans des zones textuelles sémantiquement équivalentes.

Sur le plan linguistique, et plus précisément dans la perspective d'une « linguistique instrumentée » selon (Habert, 2005), cette expérience montre l'intérêt d'exploiter un dispositif tel qu'Upery pour observer à grande échelle certains fonctionnements sémantiques et discursifs. Nous avons ainsi constaté que les rapports sémantiques Nom/Verbe – en tout cas tels qu'ils s'élaborent dans un corpus de ce type – s'émancipent largement du lien morphologique, montrant ainsi la nécessité de ne pas se limiter au critère morphologique dans le recensement des liens intercatégoriels. Au-delà, nous disposons des moyens d'observer les relations discursives qui s'établissent entre zones verbales et zones nominales et d'étudier en discours les conditions du choix entre expression nominale et expression verbale.

Dans l'immédiat, dans une perspective plus applicative, nous voulons exploiter ces données en recherche d'informations, dans le cadre de techniques d'expansion de requêtes, afin de déterminer si les couples extraits peuvent servir de ressources afin de capter des zones de variation sémantique. On sait en effet que les requêtes courtes sont très majoritairement

exprimées sous forme nominale. L'ajout de formes verbales devrait donc, comme suggéré par les résultats de (Claveau, 2003) sur un échantillon plus limité de couples Nom/Verbe, permettre d'étendre les possibilités de formulation[5].

Références

BARZILAY R., MAC KEOWN K. (2001). « Extracting Paraphrases from a parallel Corpus ». In *Actes de ACL*. Toulouse.

BOUILLON P., FABRE C., SEBILLOT P., JACQMIN L. (2000). « Apprentissage de ressources lexicales pour l'extension de requêtes ». In Chr. Jacquemin (coord.), *Traitement automatique des langues* 41 (2) : 367-393.

BOURIGAULT D. (2002). « Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus ». In *Actes de TALN 2002*. Nancy : 75-84

BOURIGAULT D., FABRE C. (2000). « Approche linguistique pour l'analyse syntaxique de corpus ». In *Cahiers de Grammaire* 25 : 131-151.

BOURIGAULT D., FABRE C., FRÉROT C., JACQUES M.-P., OZDOWSKA S. (2005). « Syntex, analyseur syntaxique de corpus ». In *Actes de TALN 2005*. Dourdan.

CLAVEAU V. (2003). *Acquisition automatique de lexiques sémantiques pour la recherche d'information*. Thèse de l'Université de Rennes 1.

FABRE C. ET JACQUEMIN C. (2000). « Boosting Variant Recognition with Light Semantics ». In *Actes de COLING (Computational Linguistics)*. Sarrebrück : 264-270.

FELLBAUM C. (1995). « Co-occurrence and antonymy ». In *International Journal of Lexicography* 8 : 281-303.

FERRET O., GRAU B., HURAULT-PLANTET M., ILLOUZ G., JACQUEMIN C. (2001). « Utilisation des entités nommées et des variantes terminologiques dans un système de question-réponse ». In *Actes de TALN 2001*. Tours.

HABERT B. (2005). *Instruments et ressources électroniques pour le français*. Coll. L'essentiel français, Ophrys, Gap/Paris.

HATHOUT N., TANGUY L. (2005). « Webaffix : une boîte à outils d'acquisition lexicale à partir du Web ». In *Revue Québéquoise de Linguistique* 32 (1) : 61-84.

MEL'ČUK I., CLAS A., POLGUÈRE, A. (1995). *Introduction à la lexicologie explicative et combinatoire.* Duculot, Louvain-la-Neuve.

MURPHY L. (2003). *Semantic Relations and the Lexicon. Antonymy, Synonymy, and Other Paradigms*. Cambridge University Press.

NESPOULOUS J.-L., VIRBEL J. (éds) (2003.) « Vers une révision de la notion de lexicalisation – Contribution à une vision dynamique du lexique mental : 'stock' lexical, catégories vs 'réseau' lexico-sémantique ». In *Regards croisés sur l'analogie,* K. Duvignau, O. Gasquet et B. Gaume (éds), *Revue d'Intelligence Artificielle RSTIA série RIA* 17 (5-6) : 747-760.

PUSTEJOVSKY J. (1995). *The Generative Lexicon*, MIT Press, Cambridge.

VOSSEN P. (éd.) (1998). *EuroWordNet : a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Dordrecht.

WANNER L. (2004). « Towards automatic fine-grained semantic classification of verb-noun collocations ». In *Natural Language Engineering* 10 (2) : 95-143.

[5] Cette expérience sera menée dans le cadre du projet ARIEL (Projet TCAN, 2004-06) : « Adaptation d'une chaîne de Recherche d'Information sur la base de traitements Linguistiques », en collaboration avec l'IRIT (CNRS – Université de Toulouse 3).