Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques

Agnès TUTIN
LIDILEM, Université Grenoble 3, BP 25, 38040 Grenoble Cedex 09
agnes.tutin@u-grenoble3.fr

Résumé. Dans cette étude sur le lexique transdisciplinaire des écrits scientifiques, nous souhaitons évaluer dans quelle mesure les méthodes distributionnelles de TAL peuvent faciliter la tâche du linguiste dans le traitement sémantique de ce lexique. Après avoir défini le champ lexical et les corpus exploités, nous testons plusieurs méthodes basées sur des dépendances syntaxiques et observons les proximités sémantiques et les classes établies. L'hypothèse que certaines relations syntaxiques - en particulier les relations de sous-catégorisation – sont plus appropriées pour établir des classements sémantiques n'apparaît qu'en partie vérifiée. Si les relations de sous-catégorisation génèrent des proximités sémantiques entre les mots de meilleure qualité, cela ne semble pas le cas pour la classification par voisinage.

Abstract. In this study about general scientific lexicon, we aim at evaluating to what extent distributional methods in NLP can enhance the linguist's task in the semantic treatment. After a definition of our lexical field and a presentation of our corpora, we evaluate several methods based on syntactic dependencies for establishing semantic similarities and semantic classes. Our hypothesis that some syntactic relations – namely subcategorized relations – is more relevant to establish semantic classes does not entirely appears valid. If subcategorized relations produce better semantic links between words, this is not the case with neighbour joigning clustering method.

Mots-clés : corpus – écrits scientifiques - classes sémantiques – analyse distributionnelle.

Keywords: corpus – scientific writings – semantic classes – distributional analysis.

1 Introduction

Le traitement sémantique des éléments du lexique constitue un préalable dans de nombreuses applications du TAL. Dans une application d'aide à la rédaction en Français Langue Etrangère (Kraif & Tutin 2006), nous souhaitons ainsi effectuer un traitement du lexique transdisciplinaire des écrits scientifiques et de ses collocations. Dans ce cadre, nous souhaiterions proposer une approche onomasiologique de ce lexique (i.e. avec un accès par le sens plutôt que par la forme), dont l'étude pourrait être facilitée si les approches « machinales » (Habert & Zweigenbaum 2003) de traitement sémantique à partir d'analyse distributionnelle se révélaient concluantes pour le travail du linguiste. Dans cette étude, nous désirons plus précisément évaluer la pertinence des méthodes d'analyse distributionnelle basées sur des dépendances syntaxiques pour la constitution de classes sémantiques homogènes de noms transdisciplinaires des écrits scientifiques. Nous voudrions en particulier déterminer dans quelle mesure cette méthode, qui s'est révélée adaptée à des sous-langages spécifiques pour la terminologie du droit (Bourigault & Lame 2002), de l'immunologie (Harris *et al*. 1989) ou de la médecine (Nazarenko *et al.* (2001), peut être appliquée au lexique du genre des écrits scientifiques qui présente davantage de polysémie. Nous faisons l'hypothèse

que certaines relations syntaxiques de dépendance, plus contraintes sur le plan syntaxique et sémantique, produiront des associations sémantiques de meilleure qualité.

Dans un premier temps, nous définirons le lexique transdisciplinaire des écrits scientifiques, et présenterons un premier classement sémantique manuel basé sur des propriétés linguistiques. Dans un second temps, nous évaluerons les résultats de la méthode distributionnelle employée par Didier Bourigault (Bourigault 2002 ; Bourigault et Lame 2002) à notre lexique, méthode qui dissocie les « voisins en tête » des « voisins en expansion », et les comparerons au classement manuel. Puis, nous nous pencherons sur une seconde méthode basée sur les dépendances syntaxiques que le mot soit recteur ou régi (à l'instar de Grefenstette 1996). Nous comparerons enfin les associations établies avec les relations syntaxiques de sous-catégorisation et les associations issues des relations de modification. Nous finirons par une évaluation et une réflexion sur les méthodes distributionnelles « machinales » pour la tâche linguistique qui nous intéresse.

2 Le lexique transdisciplinaire des écrits scientifiques : un premier classement manuel

Le lexique transdisciplinaire des écrits scientifiques, qui apparaît dans les articles de recherche, les monographies scientifiques, les mémoires, les thèses et les rapports de recherche, est le lexique partagé par la communauté scientifique mis en œuvre dans la description et la présentation de l'activité scientifique. Ce lexique peut être considéré comme un lexique de genre, n'intégrant pas la terminologie du domaine, mais renvoyant aux concepts mis en œuvre dans l'activité scientifique (*examiner, prouver, réfuter, concluant, hypothèse, examen, encourageant ...*) (Cf. aussi les définitions un peu différentes du VGOS de Phal (1971) et les travaux de Pecman (2004) sur le lexique des écrits des sciences « dures »). Nous nous intéressons en particulier au lexique méthodologique partagé par l'ensemble des disciplines scientifiques, qu'il s'agisse des sciences expérimentales, des sciences appliquées ou des sciences humaines.

L'étude de ce lexique permet d'approfondir au plan linguistique et épistémologique la spécificité de l'écrit scientifique en repérant un ensemble de traces lexicales emblématiques du genre. Ce traitement peut également déboucher sur des applications didactiques comme l'aide à la rédaction en langue maternelle et en langue étrangère. Dans cette perspective, nous souhaiterions proposer des outils facilitant le choix lexical pour les apprenants étrangers, basés sur un accès onomasiologique (accès par l'analogie ou la classe sémantique) ou sémasiologique (par la forme) (Cf. Kraif & Tutin 2006). A cet effet, un premier relevé basé sur les noms fréquents et communs à des corpus de plusieurs disciplines a été effectué puis filtré[1]. Dans un second temps, ces noms ont été répartis dans des grandes classes sémantiques, à partir de propriétés syntaxiques, morphologiques et sémantiques, un peu à la façon de Flaux et van de Velde (2000) pour les noms abstraits. Pour les 83 noms les plus fréquents, sept grandes classes ont été dégagées :

[1] Ont été retenus un ensemble de noms (catégorisation de Cordial) apparaissant plus de 15 fois en médecine, linguistique et économie dans un corpus de 2 millions de mots.

1 **Les noms de processus de l'activité scientifique** (*analyse, application, choix, ...*) sont des noms extensifs (se combinent avec *lors, durant*, des verbes phasiques, souvent avec *faire*), et ont un agent humain.
2 **Les noms d'objets construits par l'activité scientifique** (*approche, argument, concept, conception, démarche, ...*) ne sont pas extensifs, ont un agent humain, se combinent avec des verbes comme *élaborer, construire*.
3 **Les noms d'observables de l'activité scientifique** (*cas, données, échantillon, exemple, facteur,...*) ne sont pas extensifs, se combinent avec le support *être* et avec les verbes *analyser, examiner, étudier*.
4 **Les noms de supports de la rédaction scientifique** (*article, chapitre, conclusion, document, figure, ...*) sont à la fois concrets et abstraits non extensifs. Ils se combinent avec la préposition *dans*, et sont sujets du verbe *présenter*.
5 **Les noms de caractérisation** (*caractère, caractéristique, différence, difficulté, fonction,…*) sont des noms intensifs, se combinant souvent avec le support *avoir* et sont généralement accompagnés d'un adjectif.
6 **Les noms d'acteurs de l'activité scientifique** (*auteur, chercheur, …*) sont des noms humains, souvent sujets des verbes d'activité scientifique (*examiner, décrire, observer …*).
7 **Les noms de relation logique** (*but, cause, conséquence, corrélation, effet, influence, liaison, lien, rapport, relation…*), qui sont abstraits et non extensifs, se combinent avec les supports *être* et *avoir* et apparaissent souvent dans des structures : Nlogique de N..

Les noms polysémiques comme *rapport* ou *étude* sont bien entendu rattachés à plusieurs classes. Ce premier classement sera notre étalon pour l'évaluation des méthodes distributionnelles automatiques.

3 Le corpus des écrits scientifiques

Les méthodes distributionnelles machinales sont tributaires des données textuelles exploitées. La qualité des associations lexicales extraites dépend en effet très largement de l'homogénéité et de la représentativité des corpus traités. Pour cette étude, nous avons constitué un corpus de 2 millions de mots comprenant plusieurs genres d'écrits scientifiques du français (articles scientifiques, thèses, rapports, cours) dans trois disciplines assez différentes : la linguistique, l'économie et la médecine (Le tableau 1 indique le nombre de mots pour chaque type de texte). Le corpus d'articles scientifiques est extrait du corpus KIAP[2] élaboré par l'équipe de Kjersti Fløttum, de l'Université de Bergen. Notre objectif sera d'observer comment s'effectuent les regroupements des noms transdisciplinaires qui ont des comportements syntaxiques analogues.

	Linguistique	**Economie**	**Médecine**
Articles de revues (corpus KIAP)	285 881 mots	374 516 mots	164 315 mots
Thèses, rapports, cours	364 812 mots	286 653 mots	492 173 mots
Total	650 693 mots	661 169 mots	656 488 mots

Tableau 1 : Corpus des écrits scientifiques

[2] KIAP : Kulturell Identitet i Akademisk Prosa. Cf. http://kiap.aksis.uib.no/

4 La méthode distributionnelle du linguiste et l'analyse distributionnelle machinale

Pour établir des associations sémantiques, l'intérêt de l'analyse distributionnelle paraît aller de soi, puisqu'il est classique dans la tradition de la sémantique lexicale, en particulier européenne (Cruse 1986, par exemple), de considérer que des mots qui ont des environnements syntaxiques comparables partagent des propriétés sémantiques non triviales, allant de la synonymie pour les associations les plus fortes à la co-hyponymie (Cf. aussi l'étude réalisée par Galy & Bourigault (à paraître)). Le recours aux distributions syntaxiques pour mettre en évidence les propriétés sémantiques permet au linguiste de s'appuyer sur des critères tangibles, palpables, et non plus des approximations notionnelles.

Cependant, la méthode distributionnelle du linguiste, qui fait appel en partie à son intuition de sujet parlant et catégorisant, diffère assez largement de l'approche distributionnelle « orthodoxe », en particulier dans Harris et al. (1989), entièrement basée sur les observables du corpus. En effet, le linguiste choisit tout d'abord les contextes lexicaux qui lui apparaissent les plus pertinents pour circonscrire la notion qui l'intéresse (Cf. par exemple, la notion de classe d'objets, chez Gaston Gross (1994)). Dans notre champ lexical, par exemple, on pourra ainsi repérer comme 'objets construits par l'activité scientifique' des noms qui se combinent régulièrement avec les verbes de la série *élaborer, construire, concevoir*. Le linguiste laissera de côté les associations lexicales qui lui apparaissent moins déterminantes, contrairement à l'approche automatique qui ne peut pas sélectionner *a priori* les contextes lexicaux qui seront les plus révélateurs. En outre, le linguiste effectue naturellement la désambiguïsation des notions, par exemple *conclusion* comme partie du texte, ou comme aboutissement d'un raisonnement, opération qui sera beaucoup plus délicate avec une méthode machinale. Enfin, le linguiste complète les données lacunaires du corpus ou écarte les associations jugées atypiques. Si un contexte n'est pas observable dans les textes, il recourt à son intuition pour vérifier si le contexte est possible. En bref, le linguiste s'aide du corpus, mais s'en abstrait partiellement pour les besoins interprétatifs si besoin est.

La méthode distributionnelle « orthodoxe » apparaît plus contrainte, puisqu'elle doit permettre de tirer toutes les observations du corpus et rien que du corpus. Le corpus doit donc à la fois être exhaustif pour la représentativité des associations lexicales (donc de grande taille), et très homogène pour éviter la polysémie. Cette approche donne généralement de bons résultats dans le domaine de la terminologie (Bourigault & Lame 2002 ; Harris *et al*. 1989 ; Nazarenko *et al.* 2001) où le lexique présente peu de variations. Nous souhaitons évaluer la même méthode dans notre champ lexical, en exploitant des relations syntaxiques de dépendance. Nous faisons l'hypothèse qu'en sélectionnant certains types de relation, à l'instar de la méthode distributionnelle « manuelle », nous obtiendrons des résultats de meilleure qualité.

5 Évaluation de méthodes d'analyse distributionnelle machinale basées sur des dépendances syntaxiques

Dans les méthodes d'analyse distributionnelle machinale, plusieurs définitions de la distribution ont été proposées. Les plus rustiques (Cf. par exemple Grefenstette (1996) peuvent simplement prendre en compte les mots pleins partagés dans une fenêtre de quelques mots. Les distributions basées sur les relations syntaxiques partagées donnent cependant de meilleurs résultats sur les lexèmes les plus fréquents, donc les plus significatifs (Grefenstette Ibid.). Nous adopterons cette dernière méthode en exploitant les dépendances syntaxiques obtenues sur

notre corpus à l'aide des résultats de l'analyseur Syntex (Bourigault *et al.* 2005). Nous évaluerons les proximités sémantiques établies et les classes sémantiques obtenues à l'aide des coefficients de similarité entre les mots.

5.1 Proximités sémantiques établies à l'aide de la méthode de D. Bourigault (2002)

La méthode distributionnelle a été appliquée avec succès par Didier Bourigault et ses collègues à plusieurs domaines dont la terminologie du droit (Bourigault & Lame 2002). Cette approche présente deux originalités : d'une part, elle dissocie les mots proches, appelés « voisins », selon qu'ils sont recteurs (ou têtes) ou régis (dans l'expansion) ; d'autre part, comme elle vise les applications terminologiques, elle prend en compte aussi bien les unités que les syntagmes dans les relations syntaxiques de dépendance. Le système, appelé Upery, basé sur les résultats de l'analyse syntaxique du logiciel Syntex (Bourigault et *al.* 2005), extrait des triplets contenant le terme (unité lexicale simple ou complexe), la relation de dépendance, et le contexte (le syntagme ou élément lexical régi). Il rapproche ensuite, en utilisant des mesures de proximité comme le jaccard, les termes selon le nombre de contextes différents qu'ils partagent[3]. Par exemple, les mots *article* et *chapitre*, qui apparaissent à la première ligne du tableau, partagent 6 contextes identiques (= a) lorsqu'ils sont accompagnés d'un adjectif[4] (par exemple *présent, suivant, dernier ...*). *article* apparaît lui-même dans 18 contextes adjectivaux différents (= n1), alors que *chapitre* apparaît lui-même dans 12 contextes adjectivaux (= n2). Le coefficient jaccard utilisé ici calcule la proximité sémantique entre les mots avec la formule suivante : $a/(n_1+n_2-a)$. Seuls sont sélectionnés les voisins pour lesquels le coefficient de jaccard dépasse 0,10 et qui ont au moins quatre types de contextes communs.

contexte1	rel1	contexte2	rel2	a	n1	n2	jaccard
article	ADJ	chapitre	ADJ	6	18	12	0.25
article	EPI	section	EPI	6	11	19	0.25
tableau	EPI	chapitre	EPI	21	84	21	0.25

Tableau 2 : Exemples de voisins en tête extraits à l'aide de l'outil Upery de Didier Bourigault

Upery a été appliqué à notre corpus d'écrits scientifiques et sur le lexique des 85 noms transdisciplinaires classés. Nous avons ensuite évalué les couples extraits à partir des classes établies manuellement, en examinant tour à tour les voisins en expansion et les voisins en tête.

Les voisins en tête associent des mots qui sont des recteurs et qui partagent des contextes semblables avec une relation syntaxique donnée. Pour la liste de noms sélectionnés, on obtient 516 résultats. Nous avons observé pour chaque couple de voisins établi si les deux éléments associés appartenaient à la même classe dans notre classification manuelle. Si tel était le cas, nous avons considéré que la réponse était acceptable et l'avons rejetée dans le cas inverse. Par exemple, l'association *figure-chapitre* a été considérée comme satisfaisante car les deux noms font partie de la classe des 'supports écrits de l'activité scientifique', mais l'association

[3] La méthode ne prend pas en compte le nombre d'occurrences pour chaque contexte, contrairement à d'autres approches comme celle de Grefenstette (1996) mais seuls sont retenus les contextes apparaissant plus de deux fois.

[4] Les relations pourraient ici être différentes pour les deux éléments rapprochés.

hypothèse-section n'apparaît pas valide car les deux éléments appartiennent à des classes différentes.

L'observation des résultats révèle que 50,5 % des voisins en tête extraits relèvent de la même classe, ce qui est *a priori* assez peu, étant donné le caractère assez lâche des classes établies manuellement. Les voisins en tête mettent en jeu de nombreuses relations de modification[5], facultatives, et peu contraintes sur le plan sémantique, comme la relation d'épithète ou d'attribut. Par exemple, les noms *cas* et *modèle*, assez distincts sur le plan sémantique, apparaissent dans 19 contextes adjectivaux communs. Un examen plus poussé montre que nombre de ces adjectifs sont très peu contraints du point de vue de leur sélection nominale (par exemple, *autre, dernier, tel, général, précédent*) et donc probablement peu informatifs du point de vue sémantique.

Nous avons ensuite comparé ces résultats avec les voisins en expansion, c'est-à-dire les cas où les noms transdisciplinaires sont régis dans une relation de sujet ou de complément. Nous faisons l'hypothèse que ces relations qui mettent souvent en jeu des arguments sous-catégorisés – mais pas uniquement –, souvent obligatoires, seraient davantage significatives pour établir des proximités sémantiques. Les résultats obtenus, bien que peu nombreux, semblent aller dans ce sens. Utilisant les mêmes seuils que pour les voisins en tête, 52 paires de voisins sont dégagées, dont 34 apparaissent valides (65,5% des paires). L'examen plus détaillé des contextes partagés montre que les associations Nom-Verbe apparaissent souvent plus significatives que dans les contextes Nom-Adj, à l'exception des relations où le verbe *être* apparaît.

5.2 Proximités sémantiques établies à l'aide de l'ensemble des relations syntaxiques

La méthode de Didier Bourigault dissocie les voisins qui apparaissent comme têtes des voisins qui apparaissent comme régis (dans l'« expansion »). Ce traitement séparé permet de mettre en lumière des associations spécifiques, comme l'association *examiner des données* et *l'examen des données*, qui seraient autrement noyées dans l'ensemble des relations. Ce type d'observation n'étant pas essentiel pour notre étude, nous avons observé, à l'instar de Grefenstette (1996) les proximités sémantiques établies à partir de l'ensemble des relations syntaxiques, que le nom transdisciplinaire soit recteur (Ex : *analyse des données*) ou régi (Ex : *confirmer l'analyse ...*). Les contextes ont ici été réduits aux verbes, noms et adjectifs qui entretenaient une relation syntaxique avec le nom transdisciplinaire, et non plus à tous les éléments (mots simples ou syntagmes) apparaissant en cooccurrence. L'idée était ici de vérifier si une fusion des relations, en produisant un plus grand nombre de contextes communs, pouvait améliorer la qualité des résultats.

La méthode employée (avec les mêmes seuils qu'en 5.1) produit 292 paires, dont 177 (soit 60,5%) apparaissent correctement appariées. Les résultats apparaissent donc meilleurs que pour les voisins en tête, mais cependant inférieurs à ceux des voisins en expansion.

[5] Mais pas uniquement. On repère aussi des relations de compléments de noms comme dans *l'efficacité de cette méthode* ou *l'élaboration du modèle*.

En outre, une classification par voisinage (neighbour joining cluster) a été effectuée à partir d'une une matrice contenant tous les coefficients de proximité (jaccard) – sans seuil – liant les mots (Cf. Fig. 1.a). Sur les 27 classes finales obtenues, 20 constituent des sous-ensembles des 7 classes définies manuellement (2 sous-ensembles ont des éléments uniques). Les sous-classes révèlent des associations lexicales fines, qui apparaissent pour la plupart appropriées pour notre approche onomasiologique.

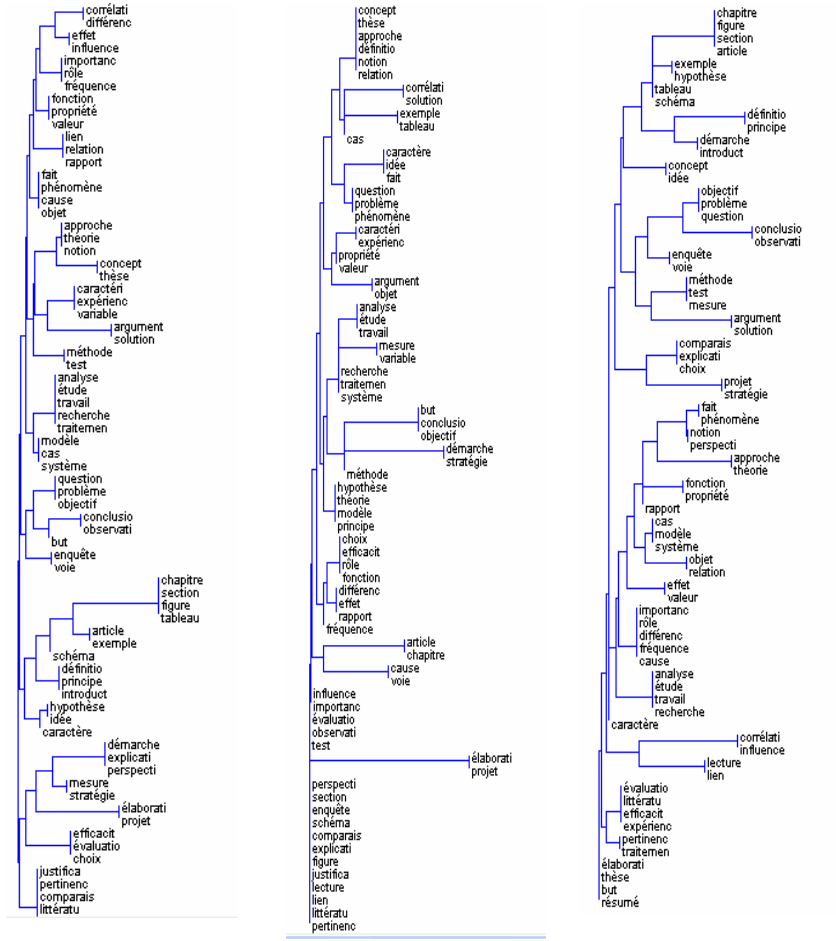
5.3 Proximités sémantiques établies à l'aide des relations de sous-catégorisation *vs* relations de modification

Nous faisons l'hypothèse que les relations syntaxiques mettant en jeu la sous-catégorisation sont plus déterminantes pour établir des proximités sémantiques que les relations de modification, parce que les arguments sont davantage contraints sur le plan syntaxique et sémantique par les restrictions sélectionnelles. Les voisins en expansion - correspondant pour la plupart à des relations de sous-catégorisation - obtenus avec la méthode de Didier Bourigault semblaient aller dans ce sens. Nous avons souhaité approfondir ce point en observant plus systématiquement quelques relations de sous-catégorisation. Les relations de sous-catégorisation observées ont été la relation objet (*confirmer une analyse*), la relation sujet (*les résultats infirment* ...), les compléments nominaux en *de*, que le nom soit recteur (ou tête) (*l'analyse des données*) ou régi (*l'efficacité de la méthode*)[6].

La méthode a dégagé 76 paires, dont 48 ont été considérées valides, soit 63 %. Nous avons ensuite comparé ces résultats avec les associations obtenues uniquement avec les modifieurs. Pour cela, nous avons sélectionné uniquement les relations liant l'adjectif épithète au nom, ainsi que la relation d'apposition. 582 paires ont été obtenues, parmi lesquelles 285 ont été validées, soit 49%. On remarque donc que le nombre de paires obtenues par les relations de sous-catégorisation apparaît nettement moins important que le nombre de paires obtenues à l'aide des relations de modification. Cette disparité des effectifs semble avoir une incidence sur les classes établies à l'aide de la même méthode qu'en 5.2 (Cf. Fig. 1.b et Fig. 1.c), puisqu'on relève que les classes obtenues par les relations de sous-catégorisation sont de moins bonne qualité (14 sur 23 classes sont des sous-classes de nos classes manuelles) que les classes obtenues à l'aide des relations de modification (20 sur 29 classes apparaissent valides).

Le type de relation – sous-catégorisation ou modification – semble donc avoir une incidence sur la qualité des associations produites avec la méthode distributionnelle lorsqu'on observe les proximités entre mots. Les relations adjectivales et apposition, plus lâches, permettent moins facilement de rendre compte du sens des noms. Les relations de sous-catégorisation paraissent plus adaptées pour cette tâche, mais la supériorité de l'analyse à l'aide des relations de sous-catégorisation n'apparaît cependant pas réelle si l'on observe les classes obtenues à l'aide des coefficients de proximité, probablement du fait d'un nombre de relations syntaxiques moins important pour ces distributions syntaxiques.

[6] Les relations incluant d'autres prépositions comme *sur* ou *dans* n'ont pas été retenues car elles mettent en jeu des relations de sous-catégorisation ou de modification selon le contexte. Le logiciel Syntex ne fait pas la différence entre ces deux types de relations.



(a) Ensemble des relations syntaxiques

(b) Relations de sous-catégorisation

(c) Relations de modification

Fig. 1 : Classification par voisinage à partir des coefficients de proximité (jaccard) entre mots

Le tableau 2 résume les résultats des méthodes employées.

	Ensemble des relations de dépendance	Relations de sous-catégorisation	Relations de modification
Nombre de paires dégagées	292	76	582
Qualité estimée pour les paires obtenues (avec la mesure jaccard)	60,5%	63%	49%
Précision des classes obtenues avec la classification par voisinage (calculée à partir du jaccard)	20/27 (74%)	14/23 (61%)	20/29 (69%)

Tableau 2 : Comparatif des méthodes employées

6 Conclusion

Les méthodes d'analyse distributionnelle automatique appliquées à notre champ lexical n'apparaissent qu'en partie concluantes. Les voisins obtenus à partir des distributions syntaxiques apparaissent valides à 60% si l'on tient compte de l'ensemble des relations syntaxiques. Nos résultats sont cependant pratiquement toujours meilleurs que ceux que Grefenstette (1996) obtient avec l'analyse syntaxique en comparant ses résultats à l'aune du thésaurus Roget. Nos classes sont cependant plus lâches.

La prise en compte des seules relations de sous-catégorisation augmente la précision (63%), mais le rappel est plus faible du fait du faible nombre de relations envisagées. Les résultats paraissent plus intéressants pour les classes obtenues par voisinage à l'aide du coefficient de proximité (jaccard), surtout si l'on prend en compte l'ensemble des relations syntaxiques (sans privilégier les relations de sous-catégorisation ou les relations de modification). Les classes obtenues confirment souvent la classification manuelle, tout en proposant des regroupements plus fins, probablement très utiles pour l'accès onomasiologique que nous envisageons pour notre application d'aide à la rédaction.

Deux types de traitement linguistique pourraient probablement améliorer les résultats. Tout d'abord, il serait souhaitable de normaliser les relations syntaxiques et les ramener à des relations plus sémantiques. Par exemple, il n'y a pas lieu de distinguer la relation entre l'adjectif épithète et le nom, et celle qui lie l'adjectif attribut et le nom. En outre, pour pallier le manque de données, il pourrait être utile de regrouper les relations par classes sémantiques, en utilisant la méthode distributionnelle de façon incrémentale. Enfin, il apparaît indispensable d'explorer d'autres mesures de similarité, comme la mesure prox, qui prend en compte la productivité de la relation syntaxique, ce qui n'est pas le cas de la mesure de jaccard.

Pour une application linguistique comme la nôtre, la méthode peut néanmoins apparaître utile, si les données obtenues sont validées manuellement. Le linguiste pourra ainsi partir des classifications obtenues automatiquement, observer les contextes partagés dans le corpus et corriger les données. Comme en terminologie, la méthode distributionnelle sera ainsi conçue comme une aide à la décision pour le lexicologue.

Remerciements

Tout d'abord, un très grand merci à Didier Bourigault qui m'a fourni les résultats de l'analyseur Syntex ainsi que les résultats du système d'analyse Upery et a relu une première version de ce papier. Merci également à Kjersti Fløttum, de l'Université de Bergen, qui m'a permis d'utiliser le corpus KIAP. Toute ma reconnaissance également à Christophe, le roi de Java, pour son aide. Merci aussi à Cécile Frérot pour ses conseils et à Olivier Kraif pour sa relecture d'une première version de ce papier.

Références

BOURIGAULT D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. Actes de *la 9ème conférence annuelle sur le Traitement Automatique des Langues* (TALN 2002), Nancy, 2002, 75-84.

BOURIGAULT D., FABRE C., FRÉROT C., JACQUES M.-P. & OZDOWSKA S. (2005), Syntex, analyseur syntaxique de corpus. Actes des *12èmes journées sur le Traitement Automatique des Langues Naturelles*, Dourdan, France.

BOURIGAULT D., LAME G. (2002). Analyse distributionnelle et structuration de terminologie. Application à la construction d'une ontologie documentaire du Droit, in *TAL*, 43-1.

CRUSE D.A. (1986). *Lexical Semantics*. Cambridge, London : Cambridge University Press (Cambridge Textbooks in Linguistics).

GALY E., BOURIGAULT D. (à paraître). Analyse distributionnelle de corpus de langue générale et synonymie. *Actes JLC* 2005. Lorient.

GREFENSTETTE G. (1996). Evaluation techniques for automatic semantic extraction : Comparing syntactic and window based approaches. In Boguraev, B. and Pustejovsky, J., editors, *Corpus Processing for Lexical Acquisition*. Cambridge, Massachusset : MIT Press, 205-216.

GROSS G. (1994). Classes d'objets et description des verbes. *Langages* 115 , 15-30.

HABERT, B. AND ZWEIGENBAUM, P. (2003). Classer les mots : sémantique à gros grain et méthodologie harrissienne. *Revue de Sémantique et Pragmatique*, (12), 101–119.

HARRIS Z., GOTTFRIED M., RYCKMAN T. (1989). *The Form of Information in Science, Analysis of Immunology Sublanguage*. Kluwer Academic Publisher, Dordrecht, The Netherlands, 1989.

KRAIF O., TUTIN A. (2006). Des corpus bilingues alignés annotés sémantiquement pour l'aide à la rédaction: application aux collocations de la langue scientifique générale. *Aide à la rédaction - Apports du Traitement Automatique des Langues, Journée d'étude l'ATALA*, Paris.

NAZARENKO A., ZWEIGENBAUM P. , HABERT B, BOUAUD J. (2001). Corpus-based Extension of a Terminological Semantic Lexicon. *Recent Advances in Computational Terminology*. Amtserdam : John Benjamins, 327-351.

PECMAN M. (2004). *Phraséologie contrastive anglais-français : analyse et traitement en vue de l'aide à la rédaction scientifique,* Thèse de doctorat, Université de Nice Sophia Antipolis, décembre 2004.

PHAL A. (1971). *Vocabulaire général d'orientation scientifique (V.G.O.S.) - Part du lexique commun dans l'expression scientifique*. Paris : Didier, Crédif.