Etude des critères de désambiguïsation sémantique automatique : présentation et premiers résultats sur les cooccurrences

Laurent AUDIBERT

Jeune équipe DELIC – Université de Provence
29 Avenue Robert SCHUMAN
13621 Aix-en-Provence Cedex 1
laurent.audibert@up.univ-aix.fr

Mots-clefs – Keywords

Désambiguïsation sémantique, corpus sémantiquement étiqueté, cooccurrences.

Word sense disambiguation, sense tagged corpora, cooccurrences.

Résumé – Abstract

Nous présentons dans cet article les débuts d'un travail visant à rechercher et à étudier systématiquement les critères de désambiguïsation sémantique automatique. Cette étude utilise un corpus français étiqueté sémantiquement dans le cadre du projet SyntSem. Le critère ici étudié est celui des cooccurrences. Nous présentons une série de résultats sur le pouvoir désambiguïsateur des cooccurrences en fonction de leur catégorie grammaticale et de leur éloignement du mot à désambiguïser.

This paper describes the beginning of a study which aims at researching and systematically analysing criteria for automatic word sense disambiguation. This study uses a French sense tagged corpus developed in the SyntSem project. We analyse here the cooccurrence criterion and report a series of results on the disambiguative power of cooccurrences with respect to their grammatical category and distance from the word to be disambiguated.

1 Introduction

La désambiguïsation automatique est un enjeu important dans la plupart des applications de traitement automatique des langues (T.A.L.). On peut citer comme exemple les applications de recherche d'information, de traduction, de reconnaissance de la parole, de reconnaissance des caractères, de restauration de l'accentuation, etc. (cf. Ide & Véronis, 1998)

Bien que la désambiguïsation du sens des mots soit un thème de recherche important depuis l'origine du T.A.L., les ressources nécessaires pour aborder correctement le problème commencent à peine à être disponibles. Ceci est particulièrement vrai pour le français, langue pour laquelle on commence à disposer d'outils informatiques efficaces d'annotation (lemmatisation[1], étiquetage morpho-syntaxique, dictionnaire de synonymes[2], relation syntaxique simple entre les mots, etc.). Il existe également des applications puissantes et évolutives qui permettent d'écrire des requêtes complexes sur des corpus étiquetés (Intex voir [Silberztein, 2000], (Win/Dos)LoX[3], IMS Corpus Workbench, etc.). Enfin un corpus français désambiguïsé, de taille exploitable, est en cours de constitution (Projet SyntSem[4]).

2 Matériaux de l'étude

La première phase de notre travail fut l'étiquetage de notre corpus avec le logiciel Cordial Analyseur[5], qui offre une lemmatisation et un étiquetage morpho-syntaxique d'une exactitude satisfaisante (Valli, Véronis, 1999). D'autres informations (hyperonymes, relations syntaxiques, classes d'objets …) seront ultérieurement apportées en fonction des besoins du critère de désambiguïsation étudié.

L'une des difficultés majeures de l'étiquetage sémantique automatique réside dans l'inadéquation des dictionnaires traditionnels (Véronis, 2001) ou dédiés (Palmer, 1998) pour cette tâche. Une autre difficulté (Gale, Church, Yarowsky, 1993) provient du manque de corpus sémantiquement étiquetés sur lesquels des méthodes d'apprentissage pourraient être entraînées. Ce manque se transforme même en absence totale pour une langue comme le français : si quelques corpus sémantiquement étiquetés commencent à apparaître pour

---

[1] Selon Y. Choueka, S. Lusignan. (1985), Disambiguation by short contexts, Actes de *Computer and the Humanities*, pp.147-158., il n'existait, en 1985, aucun outil efficace de lemmatisation automatique.

[2] Nous travaillons en collaboration avec J.L. Manguin du CRISCO (CNRS Caen) qui nous fournit un dictionnaire de synonymes, voir J. Francois, B. Victorri, J.-L. Manguin. (1999), Polysémie adjectivale et synonymie, Actes de *Colloque POLYSEMIE*, .

[3] Développés au sein de l'équipe DELIC à l'Université de Provence, WinLoX et DosLoX sont disponibles sur le site *http://laurent.audibert.free.fr/lox.htm*, voir également l'article L. Audibert. (2001), LoX : outil polyvalent pour l'exploration de corpus annotés, Actes de *RECITAL (TALN) 2001*, pp.411-419..

[4] Le projet SyntSem, financé par l'ELRA/ELDA, vise à produire un corpus étiqueté au niveau morpho-syntaxique avec en plus un marquage syntaxique peu profond et un marquage sémantique de mots sélectionnés.

[5] Cordial 7 Analyseur est développé par la société Synapse Développement (*http://www.synapse-fr.com/*).

l'anglais, notamment dans le cadre de l'action d'évaluation Senseval, cf. (Kilgarriff, 1998), ils sont pour l'instant inexistants pour le français.

Pour ces multiples raisons, Delphine Reymond, avec qui nous travaillons en collaboration, a entrepris la construction d'un dictionnaire distributionnel en se basant sur un ensemble de critères différentiels stricts (Reymond, 2002). Ce dictionnaire comporte pour l'instant la description détaillée de 20 noms communs, 20 verbes et 20 adjectifs. Il lui a permis d'étiqueter manuellement chacune des 53000 occurrences de ces 60 vocables dans le corpus du projet SyntSem (Corpus d'environ 5.5 millions de mots, composé de textes de genres variés). Ce corpus est une ressource de départ pour une étude approfondie des critères de désambiguïsation sémantique automatique puisqu'il permet l'entraînement et l'évaluation des algorithmes.

La dernière pièce manquante était un outil permettant de modéliser les critères que l'on désire étudier. Nous avons donc développé un tel outil, qui permet d'appliquer des requêtes complexes et variées sur des corpus dont le format peut évoluer (Audibert, 2001). Polyvalent, il a servi dans la phase d'étiquetage manuel des corpus. Il est également utilisé par de nombreux étudiants en TAL pour leurs travaux de recherche. Le grand pouvoir expressif de son langage de requête permet d'exprimer les critères dont on désire connaître le potentiel discriminant. Son mécanisme d'abstraction de l'objet corpus nous permettra de le faire évoluer facilement pour qu'il suive l'enrichissement de notre corpus ou l'adjonction d'autres sources d'information (dictionnaires par exemple).

20 noms	barrage, chef, communication, **compagnie**, concentration, constitution, degré, **détention**, économie, **formation**, lancement, **observation**, **organe**, passage, pied, restauration, **solution**, station, suspension, **vol**.
20 adjectifs	biologique, clair, correct, courant, exceptionnel, frais, haut, historique, plein, populaire, régulier, sain, secondaire, sensible, simple, strict, sûr, traditionnel, utile, vaste.
20 verbes	arrêter, comprendre, conclure, conduire, connaître, couvrir, entrer, exercer, importer, mettre, ouvrir, parvenir, passer, porter, poursuivre, présenter, rendre, répondre, tirer, venir.

Tableau 1 : les 60 vocables de l'étude complète

Nom	Sens	Définition	Nb.
Compagnie	Co1	Association de personnes.	327
	Co2	Présence de quelqu'un.	85
Détention	De1	Incarcération, enfermement.	81
	De2	Possession.	31
Formation	Fo1	Instruction	1227
	Fo2	Formation de qqch.	227
	Fo3	Groupement de personnes.	65
Observation	Ob1	Surveillance, étude.	492
	Ob2	Remarques, réflexions.	66
	Ob3	Conformation à.	14
Organe	Or1	Porte-parole d'un groupe.	182
	Or2	Partie d'un corps.	140
	Or3	Partie d'une machine.	44
Solution	So1	Dénouement, réponse.	821
	So2	Mélange liquide.	36
Vol	Vo1	Abréviation de volume.	52
	Vo2	Délit.	65
	Vo3	Déplacement dans l'air.	146

Tableau 2 : les 7 vocables de l'étude préliminaire

L'étude complète portera sur l'ensemble des 60 vocables, énumérés dans le Tableau 1. Ces vocables ont été sélectionnés pour l'importance de leur fréquence et pour leur caractère particulièrement polysémique. L'étude préliminaire présentée dans cet article porte seulement sur 7 noms (en gras dans le Tableau 1), qui représentent 4101 occurrences au total dans le corpus SyntSem. Le Tableau 2 présente le nombre d'occurrences dans le corpus pour chacun des sens de ces 7 vocables. A ce stade de notre étude nous travaillons avec un niveau de sens

grossier (2 à 3 sens par vocable) que nous affinerons par la suite. Le nombre d'occurrences de certains sens peut parfois paraître faible, mais (Gale, et al., 1993) ont montré que très peu d'exemples (de l'ordre de 5) permettent d'obtenir de très bons résultats, et qu'il était rarement utile de dépasser la cinquantaine d'exemples. Il semblerait que ce ne soit pas tant le nombre d'exemples que leur diversité qui compte.

3 Critères

Il existe de nombreuses sources d'information pour lever l'ambiguïté du sens des mots. Comme l'ont montré (McRoy, 1992), (Wilks, Stevenson, 1998) ou encore (Ng, Lee, 1996) toutes ces sources peuvent être utilisées simultanément pour aboutir à une meilleure désambiguïsation. Dans cette section nous dressons une liste (non restrictive) des critères de désambiguïsation sémantique automatique que nous projetons d'étudier.

De nombreuses études montrent que les cooccurrences[6] constituent un bon critère pour identifier le sens d'un mot. Mais que doit-on regarder : la forme fléchie du mot ou son lemme ; faut-il ne considérer que les mots pleins ou bien considérer aussi les mots grammaticaux ? Où doit-on regarder : jusqu'à quelle distance les cooccurrences apportent-elles de l'information ; est-il pertinent de sortir du contexte de la phrase ; doit-on différencier le contexte droit du contexte gauche ?

Des cooccurrences ciblées par des relations syntaxiques telles que les relations sujet-verbe, verbe-objet, adjectif-nom, nom-adjectif, nom-nom peuvent être plus pertinentes que des cooccurrences prises au hasard. Aussi est-il certainement important d'accorder une attention particulière à ces cooccurrences, comme l'a fait (Yarowsky, 1993).

Les étiquettes morpho-syntaxiques des mots qui entourent le mot dont on cherche à identifier le sens sont souvent un bon critère de désambiguïsation. Il faudrait donc en tenir compte. Mais une question se pose : quelle est la meilleure granularité des étiquettes morpho-syntaxiques ? Serait-il judicieux de travailler simultanément sur plusieurs niveaux de granularité ?

Selon (Gross, Clas, 1997) les traits syntactico-sémantiques permettent une subdivision des noms en 8 sous-ensembles : *humain*, *animal*, *végétal*, *inanimé concret*, *inanimé abstrait*, *locatif*, *temps*, *événement*. Toujours selon (Gross, Clas, 1997), les classes d'objets permettent une sous-catégorisation des traits. Par exemple, le terme *autobus* se voit attribuer le trait *inanimé concret* et la « classe d'objets » *moyens de transports routiers en commun*. Connaître le trait ou la classe d'objets des mots environnant le mot à désambiguïser permet d'améliorer l'apprentissage sur des petits nombres d'exemples. Il serait intéressant de connaître l'impact de tels regroupements sur la désambiguïsation.

D'autres critères seront également explorés. On peut citer, par exemple, les synonymes des mots en cooccurrence, les *n*-grammes (qui peuvent être constitués de mots, de lemmes, ou encore d'étiquettes morpho-syntaxiques), les informations sur le domaine (le thème) du texte dans lequel on cherche à désambiguïser un mot donné, etc.

---

[6] Nous emploierons, dans cet article, le mot « cooccurrence » dans son acception la plus large, dénuée des notions de fréquence, de figement, de lien syntaxique ou encore de proximité.

4 Mise en œuvre de critères sur les cooccurrences

4.1 Définition d'un critère

Tout d'abord, il faut trouver, puis énoncer clairement un critère susceptible d'être pertinent pour la levée de l'ambiguïté sémantique. Ce critère doit pouvoir être mis en œuvre dans la perspective d'un étiquetage automatique. Les critères les plus appropriés pour l'étiquetage automatique ne peuvent pas être directement calqués sur les critères différentiels présentés dans (Reymond, 2002) destinés à un étiquetage sémantique manuel. En effet, les critères différentiels sont souvent difficiles à mettre en œuvre car ils font intervenir le jugement du linguiste. De plus, les critères qui font intervenir des classes d'objets, ou des informations syntaxiques, imposent de disposer de ces informations. Or, si la lemmatisation ou l'étiquetage morpho-syntaxique sont des techniques qui commencent à arriver à maturité, l'appartenance à une classe d'objets et l'étiquetage syntaxique sont encore du domaine de la recherche. A l'inverse, la puissance de calcul des ordinateurs permet d'utiliser des critères qui seraient difficiles à mettre en œuvre manuellement.

Voici un exemple de critère :

lemme des trois noms, adjectifs, verbes ou adverbes qui suivent le mot détention sans sortir de la phrase.

Il faut ensuite écrire une règle qui modélise notre critère. Une règle est composée d'une requête, par exemple

[*lemme="détention"****] []{0,2}*** *coo:****[****ems~"(^NC|^ADV|^V|^ADJ)"****]*
stop(*ems="PCTFORTE"****),***

qui permet de formaliser le critère, et d'un masque, par exemple :

[P:coo.lemme],

qui permet de spécifier comment seront formatés les résultats de la requête. Ces requêtes sont basées sur des méta-expressions régulières (i.e. qui se situent au niveau des mots) et permettent de décrire de manière formelle des classes de suites de mots, donc des portions de texte. La syntaxe des requêtes et des masques, a été présentée dans (Audibert, 2001).

4.2 Application du critère

Les logiciels WinLoX et DosLoX permettent d'appliquer chaque règle au corpus et de dénombrer les chaînes (nous appellerons ces chaînes des indices) générées par le masque de la règle en fonction du sens du mot cible. On peut ainsi se faire une première idée de la pertinence du critère. Dans le Tableau 3 sont dénombrés des indices générés par l'application du critère « *lemme des trois mots pleins (noms, adjectifs, verbes ou adverbes) qui suivent ou qui précèdent le mot cible sans sortir de la phrase* » en fonction du sens « De1 » (incarcération, enfermement) ou « De2 » (possession) de *détention*.

indices	De1	De2
arme	0	19
contrôle	0	11
provisoire	10	0
acquisition	0	9
condition	8	1
préventif	9	0
faire	6	0
munition	0	5
camp	5	0
durée	5	0
politique	5	0
prisonnier	5	0
torture	5	0
régime	0	4
abusif	4	0
maintenir	4	0
mettre	4	0
placement	4	0

Tableau 3

4.3 Évaluation du critère

indices	De1	De2
arme	0	18
munition	0	4
condition	8	0
procédure	2	0
être	11	8

Tableau 4

Nous évaluons le critère en utilisant une méthode d'évaluation croisée (de type « leave-one-out »). Pour chacune des occurrences du mot étudié on réalise un apprentissage (une application du critère) sur toutes les autres occurrences. Puis on étiquette l'occurrence en question et on vérifie la pertinence de l'étiquetage sur cette occurrence.

L'application de la règle à l'occurrence que l'on désire étiqueter génère une liste d'indices. Nous cherchons ceux qui se trouvent dans le tableau d'apprentissage pour générer un sous-tableau similaire au Tableau 4. On recherche ensuite, dans ce tableau, l'indice dont la dispersion est la plus faible et on étiquette selon le sens le plus probable pour cet indice. Si aucun indice ne se trouve dans ce sous-tableau, on étiquette avec le sens le plus fréquent.

La mesure de dispersion utilise une fonction de la variance comprise entre 0 et 1 qui traduit la dispersion des données[7]. Plus la mesure dispersion est proche de 0, plus la dispersion des données est grande. Notre stratégie d'étiquetage ne combine pas les indices mais se focalise sur le plus fiable, désigné par le plus faible indice de dispersion. Cette stratégie, basée sur une liste de décision, est fort simple à mettre en œuvre et permet de s'affranchir de la condition d'indépendance des indices. Nous comptons cependant évaluer d'autres stratégies comme, par exemple, la quantité d'information pondérée ou les arbres de classification sémantique présentés dans (Loupy, El-Bèze, Marteau, 1998).

5 Premiers résultats sur les cooccurrences

Nous appellerons **précision** le rapport entre le nombre d'étiquetages corrects et le nombre d'étiquetages effectués :

Précision = (Nombre d'étiquetages corrects) / (Nombre d'étiquetages effectués).

Nous appellerons **précision de l'étiquetage naïf** (« base line » en anglais) la précision obtenue en étiquetant toutes les occurrences avec l'étiquette la plus fréquente dans le corpus.

Nous appellerons **gain** l'amélioration de la précision obtenue par rapport à la précision d'un étiquetage naïf :

Gain = (Précision(étiq.) - Précision(étiq. naïf)) / (1 - Précision(étiq. naïf)).

Le **rappel** sera :

Rappel = (Nombre d'étiquetages corrects) / (Nombre d'occurrences à étiqueter).

---

[7] Mesure de dispersion : $1 - \frac{\text{coef. de variation}}{\sqrt{N-1}}$.

5.1 Que faut-il regarder ?

Pour étudier quelles sont les catégories grammaticales les plus utiles, nous avons observé le résultat du critère suivant : *forme fléchie des n mots qui suivent ou qui précèdent le mot cible*. Nous avons fait varier la taille *n* du contexte et nous avons observé la catégorie grammaticale du mot ayant servi à la levée de l'ambiguïté (c'est-à-dire le mot dont l'indice de dispersion était le plus faible) pour nos 7 vocables. Les figures 1 et 2 montrent l'importance relative des catégories grammaticales pour la désambiguïsation en fonction de la taille *n* de la fenêtre. La Figure 1 montre la prédominance que prennent les noms sur toutes les autres catégories grammaticales quand la taille du contexte croît : au-delà d'un contexte de ±400 mots ils ont été sélectionnés pour la levée de l'ambiguïté dans plus de 80% des cas. On peut aussi remarquer que l'importance des adjectifs est à peu près constante : ce sont eux qui ont permis la levée de l'ambiguïté dans 10 à 20% des cas. Pour de petites fenêtres (±1 mot à ±2 mots), la Figure 2 montre l'importance des mots grammaticaux, notamment des déterminants.

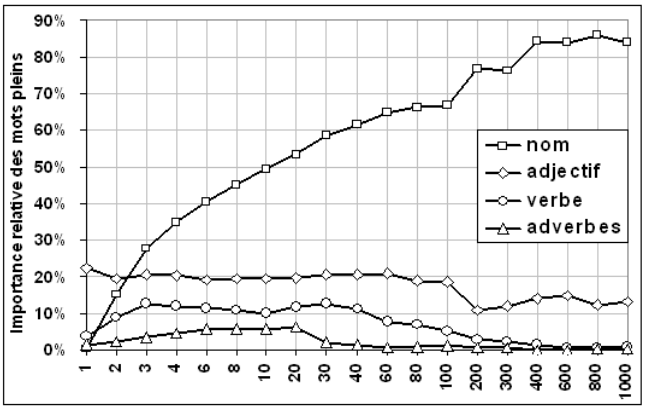

Figure 1

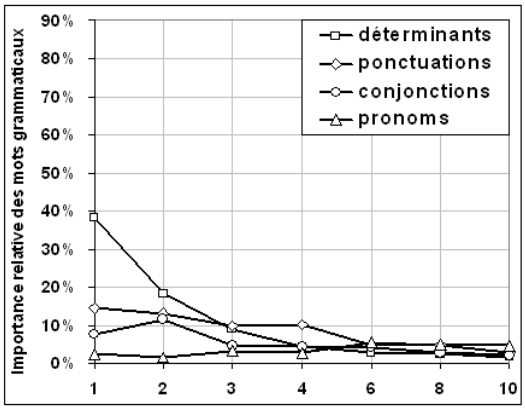

Figure 2

Faut-il regarder le lemme plutôt que la forme fléchie ? Nos expériences sur les 7 mots montrent que la lemmatisation n'apporte rien en terme de précision, mais permet une augmentation du rappel de l'ordre de 10% dans une fenêtre de ±8 mots. La lemmatisation permet, en effet, de regrouper toutes les formes fléchies d'un mot, ce qui a pour effet de diminuer la dispersion des indices et d'améliorer l'apprentissage sur de petits effectifs, d'où une augmentation du rappel. Cette augmentation est cependant réduite à néant pour des fenêtres de plus de ±30 mots. Pour de tels contextes, le nombre d'indices est suffisamment grand pour que l'apprentissage se fasse sur toutes les formes fléchies. D'autant plus que, comme nous venons de le voir, les grands contextes privilégient les noms qui ne possèdent que deux formes fléchies (singulier et pluriel).

5.2 Où doit-on regarder ?

Faut-il accepter de sortir de la phrase ? Nous avons formulé les deux critères suivants :
- *lemme des n mots pleins qui suivent ou qui précèdent le mot cible ;*
- *lemme des n mots pleins qui suivent ou qui précèdent le mot cible sans sortir de la phrase*.

Dans le corpus Syntsem, les phrases ont en moyenne 12,2 mots pleins par phrase. Nous avons donc observé un contexte allant de ±1 mot plein à ±10 mots pleins. La Figure 3 montre le gain et le rappel obtenu pour ces deux critères en fonction de la taille de la fenêtre *n* compté en

nombre de mots pleins. La différence entre les deux critères est sensible. En acceptant les cooccurrences en dehors de la phrase, on augmente le rappel d'environ 5%, mais on diminue la précision de 2% et le gain de plus de 6%.

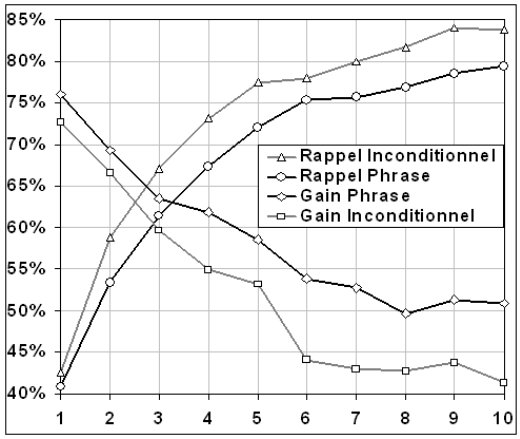

Figure 3

Ainsi, si l'on privilégie le rappel, il faut sortir de la phrase. Mais jusqu'à quelle distance du mot à désambiguïser trouve-t-on de l'information ? Nous avons tenté de désambiguïser chacun de nos 7 vocables en regardant une fenêtre de 5 mots pleins située à une distance de $\pm x$ mots de la cible. La courbe de la Figure 4 représente le gain obtenu en fonction de cette distance de x mots. Bien que nous ne travaillions ni sur la même langue ni sur les même corpus que (Gale, et al., 1993), nous obtenons un résultat analogue : pour désambiguïser un mot, on trouve de l'information dans un contexte s'étendant jusqu'à ±10000 mots.

Figure 4

En résumé, pour l'ensemble de nos 7 vocables, la meilleure précision moyenne obtenue atteint 95% pour une taille de fenêtre très petite, ±1 mot plein, sans sortir de la phrase. Mais en contre-partie, le rappel est assez faible, de l'ordre de 40%. Pour une taille de fenêtre de ±30 mots pleins, en acceptant de sortir de la phrase, le rappel atteint 89% et toutes les occurrences ont été étiquetées, la précision est donc également de 89%.

6 Perspectives et conclusion

Le travail présenté dans cet article sera approfondi et étendu à 60 mots répartis de manière égale dans trois classes grammaticales : les noms communs, les verbes et les adjectifs. Il sera ensuite complété par l'étude d'autres critères de désambiguïsation. A l'issue de cette étude, les critères retenus seront utilisés conjointement pour aboutir à une désambiguïsation automatique plus efficace et plus robuste. Notre méthode d'étiquetage conviendrait également à un étiquetage incrémental de gros corpus dans une perspective d'étiquetage semi-automatique.

On peut remarquer que, parmi les travaux similaires déjà réalisés, très peu l'ont été sur des corpus manuellement étiquetés en raison de leur rareté. Pour pallier ce problème, les chercheurs ont souvent usé de subterfuges pour évaluer leurs algorithmes. Certains fusionnent deux mots quelconques en un seul en gardant l'information du mot d'origine. Un raffinement de cette technique consiste à ne pas choisir les deux mots au hasard mais à en prendre deux qui sont homographes dans une autre langue ou encore qui ne se distinguent que par une seule lettre, cf. (Yarowsky, 1993) par exemple. Ces techniques permettent de réaliser des apprentissages supervisés sur des corpus de grande taille sans avoir à les étiqueter manuellement. Cependant, il est clair que les contextes de tels mots sont très distincts, ce qui facilite leur désambiguïsation et biaise les résultats. Notre étude porte sur de « vrais » mots et s'appuie sur un corpus de taille suffisante manuellement étiqueté.

On peut également remarquer que l'une des difficultés de l'étiquetage sémantique automatique réside dans l'inadéquation des dictionnaires traditionnels. Pour cette raison, notre corpus a été étiqueté en utilisant les définitions d'un dictionnaire distributionnel établi sur un ensemble de critères différentiels stricts.

Cette étude préliminaire a porté sur les 7 noms communs suivants : *compagnie*, *détention*, *formation*, *observation*, *organe*, *solution*, *vol*. Le nombre de sens est de 2.6 en moyenne pour ces 7 mots. Les premiers résultats obtenus sont encourageants. La précision moyenne obtenue atteint 95% pour une taille de fenêtre très petite au détriment d'un rappel assez faible. En étiquetant tous les mots la précision moyenne atteint 89% pour une taille de fenêtre de 30 mots. Ces résultats sont comparables à ceux obtenus par d'autres équipes sur des corpus en langue anglaise.

7 Références

L. Audibert. (2001), LoX : outil polyvalent pour l'exploration de corpus annotés, Actes de *RECITAL (TALN) 2001*, pp.411-419.

Y. Choueka, S. Lusignan. (1985), Disambiguation by short contexts, Actes de *Computer and the Humanities*, pp.147-158.

J. Francois, B. Victorri, J.-L. Manguin. (1999), Polysémie adjectivale et synonymie, Actes de *Colloque POLYSEMIE*.

W. A. Gale, K. W. Church, D. Yarowsky. (1993), A method for disambiguating word senses in a large corpus, Actes de *Computers and the Humanities*, pp.415-439.

G. Gross, A. Clas. (1997), Synonymie, Polysémie et Classes D'objets, *Meta*, Presses de l'Université de Montréal, pp.147-155.

N. Ide, J. Véronis. (1998), Word sense disambiguation : the state of the art, *Special Issue on Word Sense Disambiguation*, Presses de l'Université de Montréal, pp.1-40.

A. Kilgarriff. (1998), SENSEVAL : an exercise in evaluating word sense disambiguation programs, Actes de *EURALEX-98*, pp.176-174.

C. d. Loupy, M. El-Bèze, P.-F. Marteau. (1998), WSD Based on Three Short Context Methods, Actes de *SENSEVAL Workshop*.

S. McRoy. (1992), Using multiple knowledge sources for word sense discrimination, Actes de *Computational Linguistics*, pp.1-30.

H. T. Ng, H. B. Lee. (1996), Integrating multiple knowledge sources to disambiguate word sense : an exemplar-based approach, Actes de *34th Annual Meeting of the Society for Computational Linguistics*, pp.40-47.

M. Palmer. (1998), Are WordNet sense distinctions appropriate for computational lexicons ?, Actes de *SIGLEX-98, SENSEVAL*.

D. Reymond. (2002), Méthodologie pour la création d'un dictionnaire distributionnel dans une perspective d'étiquetage lexical semi-automatique, Actes de *RECITAL (TALN) 2002*.

M. Silberztein. (2000), INTEX, Association pour le traitement informatique des langues (ASSTRIL), http://ladl.univ-mlv.fr/INTEX/information.html.

A. Valli, J. Véronis. (1999), Etiquetage grammatical de corpus oraux : problèmes et perpectives, *Revue Française de Linguistique Appliquée*, Association pour le traitement informatique des langues (ASSTRIL), pp.113-133.

J. Véronis. (2001), Sense tagging : does it makes sense ?, Actes de *Corpus Linguistics'2001*, pp.in press.

Y. Wilks, M. Stevenson. (1998), Word sense disembiguation using optimised combinations of knowledge sources, Actes de *COLING-ACL98*.

D. Yarowsky. (1993), One sense per collocation, Actes de *ARPA Human Language Technology Workshop*, pp.266-271.