RÉCITAL 2003, Batz-sur-Mer, 11-14 juin 2003

Procédures d'apprentissage endogène doublées de ressources exogènes : résolution en corpus d'une ambiguïté sur " de "

Cécile Frérot

ERSS – Université Toulouse-Le Mirail
Maison de la Recherche
5 allées A. Machado
31058 Toulouse Cedex
frerot@univ-tlse2.fr

Mots-clefs – Keywords

analyse syntaxique automatique, approche endogène, ressource exogène, approche mixte, ambiguïté catégorielle

automatic parsing, endogenous strategy, exogenous resources, hybrid approach, POS ambiguity

Résumé – Abstract

Dans cette étude, nous nous intéressons à l'apport de ressources exogènes dans un analyseur syntaxique de corpus basé sur des procédures d'apprentissage endogène. Nous menons une expérience en corpus sur un cas d'ambiguïté catégorielle du français (forme *de* en position postverbale, article ou préposition). Après avoir présenté et évalué la stratégie endogène, nous en analysons les limites. Nous discutons ensuite la perspective d'une approche mixte combinant des informations acquises de manière endogène à des informations exogènes (données de sous-catégorisation verbale sur la préposition *de*). Nous montrons alors comment un apport maximal de ressources exogènes améliore les performances de l'analyseur (+8%, +15% sur les deux corpus évalués). Nous présentons les premiers résultats d'une approche mixte avant de conclure sur les orientations futures du travail.

This paper addresses the issue of the contribution of exogenous resources within the framework of a parser, based on endogenous techniques. We discuss how exogenous resources could combine with endogenous techniques in the context of a POS French ambiguity (the word *de,* determiner or preposition). We present and evaluate our endogenous strategy on cases where verbs are adjacent to *de.* We highlight the limits of such a strategy and show how exogenous resources improve the parser output (+8%, +15% on the corpus evaluated). Finally, we present the first results of the combined strategy and conclude on future work.

1 Introduction

Notre cadre de travail est la réalisation de l'outil d'analyse syntaxique de corpus, Syntex (Bourigault, Fabre, 2000), dont l'application principale est la construction de ressources lexicales spécialisées à partir de corpus (terminologies, ontologies, lexiques pour la traduction). La spécificité majeure de l'analyseur est d'être doté de procédures d'apprentissage endogène (Bourigault, 1994) qui lui permettent d'acquérir, sur chaque nouveau corpus, les informations nécessaires à la résolution des ambiguïtés de rattachement syntaxique. Ce choix de ne fournir à l'analyseur aucune information sémantique ni donnée de sous-catégorisation est lié à l'hypothèse, forte, sur l'idiosyncrasie des textes spécialisés. Il s'explique par la conviction, étayée par l'analyse détaillée de textes de domaines distincts, qu'on ne peut pas forcément utiliser des connaissances linguistiques générales dans des textes spécialisés, et que les textes d'un domaine donné peuvent posséder des propriétés de complémentation distinctes de celles d'un autre domaine (Bourigault, Fabre, 2000, Fabre, Bourigault, 2001, Basili et al., 1997, Basili et al., 1999).

Les méthodes basées sur l'acquisition de connaissances à partir des seules données du corpus montrent cependant des limites. Dans notre analyseur, une des limites concerne certains phénomènes rares, par leur nombre d'occurrences dans les corpus, et non moins " généraux " (c'est-à-dire recensés dans des dictionnaires de langue générale) qui échappent à la stratégie endogène, basée sur la notion de productivité (Bourigault, 1994). D'où un nombre d'ambiguïtés de rattachement syntaxique mal résolues, notamment au niveau des prépositions régies par des verbes. C'est le cas en particulier de structures verbales dont on peut penser qu'elles sont décrites dans des ressources lexicales générales et que leur fonctionnement est stable d'un corpus à l'autre. Partant de ce constat, nous faisons l'hypothèse que les performances de l'outil peuvent être améliorées grâce à l'introduction dans l'analyseur de ressources exogènes décrites sous la forme de données de sous-catégorisation verbale. Nous pensons qu'une approche endogène est nécessaire pour résoudre les spécificités du corpus traité, mais qu'une telle approche présente néanmoins des limites et doit être complétée par des ressources exogènes. Les approches mixtes déjà testées concernent (presque) exclusivement l'acquisition de liens sémantiques et exploitent majoritairement des ressources existantes spécialisées (Morin, 1998, Habert et al., 1998), beaucoup plus rarement des ressources générales (Basili et al., 1997, Hamon et al., 1998). C'est ici dans le cadre d'une analyse syntaxique automatique de corpus que nous souhaitons étudier la pertinence d'une approche mixte exploitant des ressources lexicales générales.

Nous présentons un premier travail sur une mise en œuvre de cette approche. Nous menons une expérience en corpus sur une ambiguïté catégorielle du français concernant la forme *de* en position postverbale (article ou préposition). Après avoir présenté et évalué la stratégie endogène, nous en analysons les limites. Nous montrons ensuite comment l'apport d'une ressource lexicale générale, listant les verbes sous-catégorisant la préposition *de,* améliore les performances de l'analyseur. Pour cette expérience, nous construisons cette ressource à partir des corpus soumis à l'évaluation. Ce choix se justifie par l'objectif même de l'étude qui vise à mesurer l'apport maximal d'une ressource exogène dans notre analyseur. Pour cela, il était donc fondamental de nous mettre en situation d'apport maximal, c'est-à-dire de disposer, pour les deux corpus évalués, de l'ensemble de ces verbes.

2 Etude de cas sur une ambiguïté catégorielle du français

Notre étude porte sur une ambiguïté catégorielle du français concernant la forme *de,* et les ambiguïtés portées par les formes *du* et *des.* Cette ambiguïté est particulièrement " rebelle " à l'étiquetage morpho-syntaxique lorsque *de* se trouve en position postverbale (Leconte, 1998, Silberztein, 2000). Les formes *du, des, de* peuvent alors correspondre à un article (*il a fourni des résultats intéressants*) ou à la préposition *de* (*il doute des conclusions*). Cette ambiguïté est généralement traitée au cours de l'étiquetage morpho-syntaxique *via* l'utilisation de ressources lexicales - c'est le cas en particulier de (Aït-Mokhtar et al., 2002) qui exploitent des données de sous-catégorisation verbale pour lever l'ambiguïté. Un traitement post-étiquetage est rarement envisagé même s'il semble s'imposer, comme le constatent (Giguet, Vergne, 1997) qui utilisent des informations sur les sujets et objets des verbes pour résoudre l'ambiguïté.

Dans le cadre de notre analyseur syntaxique, nous sommes confrontés à cette ambiguïté qui n'est que partiellement levée par l'étiqueteur. Or sa résolution s'avère fondamentale pour notre outil. En effet, comme l'ont montré (Aït-Mokhtar, Chanod, 1997), la propagation des erreurs d'étiquetage affaiblit nettement les performances de reconnaissance des objets directs et prépositionnels d'un analyseur syntaxique. En outre, dans Syntex, c'est à partir des résultats de l'analyse syntaxique que s'opère l'analyse distributionnelle, qui vise à faire émerger des classes sémantiques de mots en fonction des contextes syntaxiques partagés (Bourigault, 2002). Et ces contextes sont précisément déterminés par les relations syntaxiques établies par l'analyseur ; dans le cas présent, il s'agit de la relation syntaxique entre un verbe et *de* (marquée OBJ - complément d'objet direct - ou PREP - complément prépositionnel[1]).

Notre stratégie de résolution ne s'appuie pas sur les cas déjà résolus par l'étiqueteur ; nous " effaçons " ses choix et affectons l'étiquette *Prep*[2] à *de.* C'est au cours de l'analyse syntaxique que l'ambiguïté est résolue ; plus exactement, la reconnaissance d'un complément d'objet direct ou prépositionnel nous permet aussi de résoudre l'ambiguïté catégorielle. Le problème consiste donc à établir de manière automatique la relation syntaxique entre un verbe et *de*, quand pour une même configuration syntaxique donnée, la nature de cette relation est ambiguë (tableau 1).

Configuration syntaxique	**Complément d'objet direct (relation OBJ)**	**Complément prépositionnel (relation PREP)**
Vb[3]*+Prep+Det+NomP*	*créer des climats*	*douter des phénomènes*
Vb+Prep+Det+NomS	*faire de la morphologie*	*provenir d'une source*
Vb+Prep+AdjP+NomP	*manifester de belles intentions*	*profiter de diverses opportunités*

Tableau 1 : Configurations syntaxiques ambiguës

1 Précisons que dans l'analyseur nous ne faisons aucune distinction entre groupes prépositionnels (GP) arguments ou circonstants. Quel que soit le statut du GP, nous cherchons à le rattacher au verbe.

2 Exemples : *le système donne des résultats satisfaisants*
donne Prep|de Det|le résultats
les spécialistes doutent de la nature liquide du magma
doutent Prep|de Det|la nature

3 *Vb* : verbe, *Det* : déterminant, *NomP* : nom pluriel, *NomS* : nom singulier, *AdjP* : adjectif pluriel.

3 Stratégie endogène : présentation et évaluation

3.1 Illustration de la stratégie

Le principe de l'apprentissage endogène est classiquement utilisé dans l'analyseur pour résoudre des cas d'ambiguïté de rattachement syntaxique (adjectival, prépositionnel). Ici, ce principe est mis en œuvre non pour lever une ambiguïté de rattachement mais pour déterminer la nature de la relation syntaxique (OBJ ou PREP) à établir entre un verbe et *de* en contexte contigu. La procédure est la suivante : l'analyseur acquiert des informations sur la complémentation des verbes grâce au repérage de configurations syntaxiques non ambiguës pour la relation PREP. Il utilise ensuite ces informations pour calculer des indices, basés sur la notion de productivité[4] (Bourigault, Fabre, 2000). Dans la phase de résolution des cas ambigus, ces indices sont exploités conjointement aux indices obtenus pour la relation OBJ (informations acquises grâce au module de recherche des objets directs). Nous illustrons cette stratégie dans le tableau 2.

1	**Cas ambigus**	Vb + de + les + NomP OBJ ou PREP ? *parler des sens* : PREP	*réaliser des expériences* : OBJ
2	**Cas non ambigus PREP (indice PREP)**	Vb + de + Ø + NomPluriel Vb + de + Detindef + Nom *parler de formes* *parler d'une désintégration*	
3	**Cas non ambigus OBJ (indice OBJ)**		Vb + Det + Nom *réaliser une action, réaliser les figures, réaliser un triage*
4	**Indices**	OBJ(parler) = 0 PREP(parler) = 2	OBJ(réaliser) = 3 PREP(réaliser) = 0
5	**Décision**	*parler des sens* → PREP	*réaliser des expériences* → OBJ

Tableau 2 : Stratégie endogène . Les données se lisent ainsi : pour résoudre les cas ambigus (1), l'analyseur s'appuie sur la résolution de cas non ambigus pour PREP et OBJ (2, 3) ; il calcule ensuite des indices en faveur de la relation PREP ou OBJ et les exploite conjointement dans la phase de résolution des cas ambigus (4) pour prendre une décision (5).

3.2 Evaluation

L'évaluation porte sur deux corpus, un ouvrage de géomorphologie (GEO, 243 000 mots) et le journal *Le Monde*[5] (LM, 250 000 mots) ; nous avons évalué 850 cas ambigus dans chaque corpus. L'évaluation exclut des cas entrant dans une configuration étiquetée *Verbe+Prep(de)*

[4] La productivité est déterminée par le nombre de régis différents avec lesquels un couple {recteur, préposition} se combine dans le corpus.

[5] Le corpus (année 1998 du journal *Le Monde*) comportait initialement environ 700 000 mots. Nous avons souhaité travailler sur un nombre de mots équivalent pour les deux corpus et n'avons donc exploité qu'une partie de LM.

mais qui ne relèvent pas de l'ambiguïté, tels les sujets inversés (*dans ces montagnes apparaissent des érosions*) ou les structures causatives (*il a fait chuter des taux)* dont le traitement dépend d'un autre module[6]. Les résultats de l'évaluation apparaissent dans le tableau 3.

Corpus	GEO			LM		
Syntex / *Correct*	OBJ	PREP	Total	OBJ	PREP	Total
OBJ	618	6	624	547	3	550
PREP	84	142	226	167	133	300
Total	702	148	850	714	136	850
Taux de réussite	GEO			LM		
Syntex	89% (618+142/850)			80% (547+133/850)		
	OBJ 99% (618/624)	PREP 63% (142/226)		OBJ 99% (547/550)	PREP 44% (133/300)	
Stratégie de base (OBJ)	73% (624/850)			65% (550/850)		
Gain endogène	16%			15%		

Tableau 3 : Evaluation de la stratégie endogène

Illustrons la lecture des données sur le corpus GEO : sur 624 cas OBJ, l'analyseur en a reconnu 618 ; sur 226 cas PREP, il en a reconnu 142. Le taux de réussite mesure le rapport entre le nombre de cas correctement analysés par Syntex (618+142) et le nombre de cas validés (624+226). Nous détaillons ce taux pour les deux relations OBJ et PREP et le comparons à une stratégie de base qui choisit la relation OBJ pour l'ensemble des cas : sur 850 cas validés, la stratégie de base prend donc la bonne décision dans 624 cas. Cette comparaison nous permet de mieux appréhender le gain de la stratégie endogène (différence entre le taux de réussite de Syntex et celui de la stratégie de base : 89% - 73% = 16%).

3.3 Analyse des résultats

Globalement, les résultats attestent du gain apporté par l'apprentissage endogène par rapport à une stratégie de base qui choisirait la relation OBJ dans chaque cas. Sur l'ensemble des données, on retiendra le taux de réussite pour la relation PREP, qui semble témoigner des limites de la stratégie endogène (GEO : 63%, LM : 44%). C'est à cette relation que nous allons nous intéresser en cherchant à identifier les causes d'échec.

Ces causes concernent notamment l'absence ou le manque d'indices en corpus[7]. C'est le cas par exemple de verbes qui sous-catégorisent la préposition *de* et sont décrits dans des ressources lexicales générales. Aucune occurrence de ces verbes n'a été trouvée en contexte non ambigu, l'indice endogène pour PREP est donc nul. Citons dans GEO les verbes : *affluer de, dériver de, douter de, profiter de, résulter de, se désolidariser de, se distinguer de* et dans

[6] Lorsque ces cas ne sont pas, ou que partiellement, résolus par l'analyseur, les règles pour la résolution de l'ambiguïté catégorielle sur *de* s'appliquent (à tort).

[7] Syntex choisit par défaut la relation OBJ.

LM, les verbes *discuter de, dépendre de, pâtir de, regorger de, bénéficier de*. C'est ici la fréquence d'apparition des verbes qui est " responsable " de l'échec, et plus précisément l'absence de redondance lexico-syntaxique en contexte non ambigu, sachant qu'un cas ambigu a d'autant plus de chance d'être résolu qu'il a été repéré dans un contexte {verbe, relation syntaxique} non ambigu. Des erreurs d'analyse syntaxique expliquent également les résultats. Ces erreurs se propagent dans le calcul des indices endogènes et les " gonflent " artificiellement ; c'est le cas de certains sujets inversés analysés comme des objets directs qui faussent l'indice OBJ. Des cas de conflit indiciel se présentent aussi : deux indices[8] comportent une même valeur non nulle, donc non discriminante ; ce conflit est à la fois imputable au manque de redondance lexico-syntaxique et aux erreurs d'analyse.

La redondance lexico-syntaxique des deux corpus semble une piste intéressante à explorer pour analyser le taux de réussite (PREP) contrasté entre GEO et LM, taux supérieur d'environ 20% sur GEO (63%) par rapport à LM (44%). Comme nous l'avons déjà précisé, l'apprentissage endogène est basé sur la notion de productivité et sa performance repose sur la redondance des structures lexico-syntaxiques. Si le taux de réussite (PREP) est meilleur dans GEO que dans LM, on peut avancer l'idée que la redondance lexico-syntaxique est plus forte dans GEO que dans LM. Et qu'elle est plus forte en contexte non ambigu. Cette redondance peut s'expliquer par l'homogénéité thématique de GEO (géomorphologie) face à l'hétérogénéité thématique de LM (politique, économie, culture, sciences et techniques...), et par leurs vocabulaires respectifs (lexique verbal plus dense dans LM que dans GEO).

4 Introduction de ressources exogènes dans l'analyseur

4.1 Données de sous-catégorisation verbale

Ce constat sur les limites du " tout endogène " nous amène à envisager une stratégie alternative, *via* l'utilisation dans l'analyseur d'une ressource lexicale générale listant l'ensemble des verbes sous-catégorisant la préposition *de.* A défaut d'avoir achevé la constitution exhaustive de cette ressource, nous constituons les données à partir des corpus utilisés pour l'expérience (GEO et LM). Cette démarche se justifie par l'objectif même de notre étude ; nous cherchons ici à mesurer quel peut être l'apport maximal d'une ressource lexicale générale. Pour cela, il était donc indispensable de nous mettre en situation d'apport maximal et de disposer, pour les deux corpus évalués, de l'ensemble de ces verbes.

Les listes constituées sont les suivantes : *i)* verbes qui sous-catégorisent la préposition *de* et ne sont pas transitifs directs (liste A) ; *ii)* verbes sous-catégorisant la préposition *de* ou transitifs directs (liste B1) ; *iii)* verbes à double complémentation (liste B2). Nous détaillons ces listes (tableaux 4, 5).

GEO	LM
arriver de, dépendre ~, dériver ~, douter ~, émerger ~, faire partie[9] *~, jaillir ~, parler ~, préjuger ~, profiter ~, provenir ~, rendre compte ~, venir ~*	*abuser de, bénéficier ~, découler ~, débattre ~, dépendre ~, démissionner ~, discuter ~, émaner ~, faire état ~, faire figure ~, grouiller ~, jouir ~, mourir ~, pâtir ~*

Tableau 4 : Liste A

[8] Indices opposés, c'est-à-dire l'un en faveur de PREP, l'autre de OBJ.

[9] Les listes comprennent des formes verbales complexes (*avoir besoin, faire partie, rendre compte, ...*). Ce choix est lié à la concaténation de certaines séquences *Verbe+Nom* lors de l'étiquetage morphosyntaxique.

GEO		***LM***	
Verbes			
Transitifs directs	**Sous-catégorisant *de***	**Transitifs directs**	**Sous-catégorisant *de***
dater des phénomènes *disposer des feuillets*	*dater de cette époque* *disposer de données*	*relever des erreurs* *dépasser les limites*	*relever d'une grande qualité* *dépasser de quelques mètres*

Liste B1 : verbes transitifs directs ou sous-catégorisant *de*

Verbes	
A double complémentation	**A double complémentation**
imbiber le premier d'eau *enrober de calcite des matériaux* *protéger du ruissellement des portions de la formation*	*accuser des hommes de corruption* *remplir d'humanité ce monstre froid* *faire de ce fait un dossier d'actualité*

Liste B2 : verbes à double complémentation (directe et prépositionnelle)

Tableau 5 : Listes B1 et B2

La liste A est destinée à être utilisée dans l'analyseur lors de la première passe (résolution de cas détectés comme non ambigus). Ces verbes, même s'ils entrent en corpus dans une configuration syntaxique ambiguë, ne sont pas transitifs directs : la forme *de* correspond donc à la préposition. Les verbes de la liste B1 entrent dans une construction transitive directe ou prépositionnelle, qui influe directement sur leur sens : *de* est dans ce cas une forme ambiguë (*disposer de*$_{(PREP)}$ *données* ≠ *disposer des*$_{(OBJ)}$ *feuillets*). Ce qui est également le cas des verbes de la liste B2 : la forme *de* qui suit le verbe correspond à l'article (*accuser des*$_{(OBJ)}$ *hommes de corruption)* ou à la préposition *(enrober de*$_{(PREP)}$ *calcite des matériaux*). Les verbes de B1 et B2 sont actuellement regroupés dans une seule et même liste et font l'objet d'un traitement indifférencié dans l'analyseur. Ils sont utilisés lors de la deuxième passe (résolution de cas ambigus) et constituent un indice exogène pour PREP, destiné à être combiné aux indices endogènes.

4.2 Résolution de l'ambiguïté

Nous décrivons la procédure de résolution de l'ambiguïté en contexte contigu *verbe+de.*

- Première passe : résolution de cas détectés comme non ambigus. L'analyseur se base par ordre de priorité décroissante sur :

- la liste A. Si le verbe est dans la liste, alors l'analyseur choisit la relation PREP entre le verbe et *de* ;
- une relation syntaxique OBJ déjà établie par l'analyseur. On fait l'hypothèse qu'il s'agit de verbes à double complémentation et que si l'analyseur a rattaché un complément d'objet direct au verbe, alors la forme *de* correspond à la préposition. L'objet direct peut être réalisé sous la forme d'un clitique (*une crise climatique qui [le]*$_{(OBJ)}$ *dégage de*$_{(PREP)}$ *tout manteau d'altération*), d'un pronom relatif (*c'est l'interprétation [qu]*$_{(OBJ)}$*'on a donné des*$_{(PREP)}$ *mandelles*), ou d'un nom à distance du verbe (*exclure de*$_{(PREP)}$ *l'étude des socles [les montées tardives]*$_{(OBJ)}$*, faisant du*$_{(PREP)}$ *Brésil aujourd'hui au septième rang [le quatrième producteur mondial]*$_{(OBJ)}$) (règle 1) ;
- les configurations syntaxiques non ambiguës pour la relation PREP (règle 2).

Cet ordre de priorité se fonde sur la performance respective des règles 1 et 2[10]. Les cas résolus par ces règles constituent les contextes d'acquisition à partir desquels sont calculés les indices endogènes utilisés lors de la deuxième passe. Afin de mieux " contrôler " les données d'acquisition erronées, nous envisageons d'affecter une valeur de confiance à chaque règle en fonction de sa fiabilité et de doter également les contextes d'acquisition pour la relation OBJ d'une valeur de confiance.

- Deuxième passe : résolution de cas ambigus

La stratégie de résolution repose sur un calcul de scores. Les listes B1 et B2 (verbes transitifs directs et/ou sous-catégorisant la préposition *de*) constituent un indice exogène en faveur de la relation PREP, exploité conjointement aux indices endogènes. L'idée mise en œuvre est d'exploiter prioritairement les données basées sur les indices endogènes et, lorsque la valeur du score est jugée insuffisante, de faire intervenir l'information exogène.

4.3 Résultats

Nous présentons dans le tableau 6 les résultats de la stratégie mixte.

Corpus	GEO			LM		
Syntex / *Correct*	OBJ	PREP	Total	OBJ	PREP	Total
OBJ	622	2	624	549	1	550
PREP	23	203	226	44	256	300
Total	645	205	850	593	257	850

Taux de réussite Syntex	97% (622+203/850)		95% (549+256/850)	
	OBJ	PREP	OBJ	PREP
	100% (622/624)	90% (203/226)	100% (549/550)	85% (256/300)
Gain exogène[11]	8%		15%	

Tableau 6 : Evaluation de la stratégie mixte

L'utilisation d'informations exogènes améliore les performances de l'analyseur : +8% sur GEO, +15% sur LM. Globalement, ces informations " agissent " sur les faiblesses de l'endogène et permettent à l'analyseur de prendre la bonne décision lorsque la redondance lexico-syntaxique seule ne le permettrait pas. On note que l'écart entre le taux de réussite des deux corpus s'est considérablement réduit : il était de 9% lors de la première évaluation (89% - 80%) ; à l'issue de la deuxième évaluation, il n'est plus que de 2% (97% - 95%). L'introduction de ressources exogènes dans l'analyseur améliore le traitement des deux corpus évalués, mais elle semble d'autant plus l'améliorer que la redondance lexicale du corpus est faible et le nombre de verbes présents dans les ressources exogènes est élevé[12].

[10] Règle 1 : 93% (GEO), 89% (LM) ; règle 2 : 90% (GEO), 87% (LM).

[11] Gain exogène : taux de réussite Syntex (1ère évaluation) – taux de réussite Syntex (2ème évaluation), soit pour GEO : 97% - 89% = 8%, et pour LM : 95% - 80% = 15%.

[12] Liste A : 115 verbes pour LM ; 75 verbes pour GEO.

Le gain exogène est majoritairement apporté par les verbes de la liste A et on peut se demander, au vu des ambiguïtés résiduelles pour PREP, quel gain représenterait un traitement différencié des listes B1 et B2. Les ambiguïtés résiduelles concernent le cas des verbes à double complémentation (complément d'objet direct et complément prépositionnel). Or en l'état, l'analyseur ne traite pas le cas de *de* lorsqu'il se trouve à distance du verbe[13]. Il semble que seule une analyse de la construction verbale dans son ensemble permettra de résoudre correctement les formes ambiguës *de*, contiguës et à distance du verbe. Les ambiguïtés résiduelles concernent aussi des groupes prépositionnels (GP) qui entretiennent avec le verbe une relation de nature circonstancielle et non argumentale. L'utilisation de données de sous-catégorisation listant la nature syntaxique des arguments ne permettra pas de lever l'ambiguïté. On citera l'exemple des compléments de manière (*fonctionner du premier coup, s'apprécier de diverses manières, condamner de la façon la plus ferme*) et groupes adverbiaux (*arriver de surcroît dans un climat tendu, incorporer d'office ce logiciel, attester de fait leur existence*). La résolution de ces cas implique de mettre en œuvre des méthodes de traitement complémentaires. Nous envisageons à ce titre de poursuivre le travail entrepris sur le repérage automatique en corpus de GP arguments ou circonstants (Fabre, Frérot, 2002) pour isoler de manière automatique et retyper ces GP au statut circonstanciel.

5 Conclusions et perspectives

Ces premiers résultats sur l'introduction de ressources exogènes dans l'analyseur sont prometteurs. Cette phase expérimentale en corpus doit à présent s'accompagner de la constitution d'une ressource lexicale générale complète pour les liste A et B, et son apport dans Syntex doit être évalué sur différents corpus. L'étude a porté sur la résolution de l'ambiguïté catégorielle en contexte contigu *verbe+de ;* nous travaillons actuellement à la résolution de l'ambiguïté lorsque *de* se trouve à distance du verbe[14] : la résolution de cette double ambiguïté, catégorielle et structurelle (il s'agit de déterminer, outre la nature de la relation, à quel recteur se rattache *de*) s'inscrit plus globalement dans une stratégie de rattachement prépositionnel visant à " marier " harmonieusement procédures d'apprentissage endogène et ressources exogènes.

Remerciements

Je remercie vivement Didier Bourigault et Cécile Fabre de leurs conseils et remarques au cours de la rédaction de cet article.

Références

Aït-Mokhtar S., Chanod J-P. (1997), Subject and Object Dependency Extraction Using Finite-State Tranducers. Proceedings of *ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid.

Aït-Mokhtar S., Chanod J-P., Roux C. (2002), Robustness beyond shallowness : incremental deep parsing. *Natural Language Engineering,* Vol.8 (2/3), pp. 121-144.

[13] Dans l'exemple : *déduire des$_{(PREP)}$ phénomènes analysés des$_{(OBJ)}$ conclusions intéressantes*, le rattachement de *des$_{(OBJ)}$* au verbe *déduire* n'est pas pris en charge par l'analyseur.

[14] Exemples : *joindre par des segments de droite des$_{(OBJ)}$ points carrés, armer le cours d'eau de$_{(PREP)}$ particules tranchantes.*

Basili R. Pazienza M-T., Vindigni M. (1997), Corpus-driven Unsupervised Learning of Verb Subcategorization Frames, Actes du *5ème congrès AI*IA 97*, M. Lenzerini (ed), Lecture Notes in Artificial Intelligence, 1321, pp. 159-170.

Basili R., Pazienza M-T., Vindigni M. (1999), Adaptative Parsing and Lexical Learning. Actes de *VEXTAL'99*, Venise.

Bourigault D. (1994), Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition de connaissances à partir de textes, Thèse de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.

Bourigault D., Fabre C. (2000), Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de grammaire*, Vol.25, pp.131-151.

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

Fabre C., Bourigault D. (2001), Linguistic clues for corpus-based acquisition of lexical dependencies. Actes de *Corpus Linguistics Conference*, Lancaster, 176-184.

Fabre C., Frérot C. (2002), Groupes prépositionnels arguments ou circonstants : vers un repérage automatique en corpus. Actes de la conférence *TALN*, 215-224.

Giguet E., Vergne J. (1997), From Part of Speech Tagging to Memory-based Deep Syntactic Analysis. Proceedings of the *International Workshop on Parsing Technologies,* MIT, Boston, 77-88.

Habert B., Nazarenko A., Zweigenbaum P. (1998), Extending an Existing Specialized Semantic Lexicon. Proceedings of *Coling-ACL'98*, Granada, 663-668.

Hamon T., Nazarenko A., Gros C. (1998), A step towards the detection of semantic variants of terms in technical documents. Proceedings of *Coling-ACL'98*, 498-504.

Leconte J. (1998), Le catégoriseur Brill14-JL5 / Winbrill-0.3.

Morin E. (1998), Prométhée : un outil d'aide à l'acquisition de relations sémantiques entre termes. Actes de la conférence *TALN*, Paris.

Silberztein M. (2000), Manuel d'utilisation d'Intex 4.12.