Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique

Didier Bourigault*, Cécile Frérot†[1]

** CLLE-ERSS*
CNRS et Université Toulouse Le Mirail
5, allées Antonio-Machado
F-31058 Toulouse cedex 9
didier.bourigault@univ-tlse2.fr
† Université Stendhal Grenoble 3
Domaine universitaire BP 25
F-38040 Grenoble cedex 9
Cecile.Frerot@u-grenoble3.fr

RÉSUMÉ. *Nous présentons une expérience d'utilisation d'informations de sous-catégorisation par un analyseur syntaxique pour la résolution d'ambiguïtés de rattachement prépositionnel. Le lexique de sous-catégorisation est constitué de probabilités associées à des couples (mot, préposition). Il a été construit automatiquement à partir d'un corpus de 200 millions de mots. Pour évaluer ce lexique, nous utilisons quatre corpus de test de genres variés. Nous testons plusieurs stratégies de désambiguïsation, et montrons qu'une stratégie mixte, utilisant à la fois des probabilités de sous-catégorisation spécifiques acquises à partir du corpus en cours de traitement et les probabilités de sous-catégorisation génériques donne les meilleurs résultats : les performances en précision de l'analyseur sur la tâche de désambiguïsation des rattachements prépositionnels varient selon les corpus de 79,4 % à 87,2 %.*

ABSTRACT. *We carry out an experiment aimed at using subcategorization information into a syntactic parser for PP attachment disambiguation. The subcategorization lexicon consists of probabilities between a word (verb, noun, adjective) and a preposition. The lexicon is acquired automatically from a 200 million word corpus, that is partially tagged and parsed. In order to assess the lexicon, we use four different corpora in terms of genre and domain. We assess various methods for PP attachment disambiguation : an exogenous method relies on the sub-categorization lexicon whereas an endogenous method relies on the corpus specific ressource only and an hybrid method makes use of both. The hybrid method proves to be the best and the results vary from 79.4 % to 87.2 %..*

MOTS-CLÉS : *analyse syntaxique, rattachement prépositionnel, sous-catégorisation, évaluation.*

KEYWORDS: *Parsing, pp-attachment, sub-categorization, evaluation.*

1 Ce travail a été réalisé alors que Cécile Frérot était doctorante au sein de l'Équipe de Recherche en Syntaxe et Sémantique de Toulouse, dans le cadre d'une convention CIFRE avec la société Synomia.

1. Introduction

Dans leur grande majorité, les nombreux travaux sur le développement de parseurs statistiques concernent la langue anglaise et tendent à utiliser comme corpus d'apprentissage et comme corpus de test des portions de la section du *Wall Street Journal* du Penn TreeBank (Charniak, 1997). Outre qu'elle permet d'éviter la tâche laborieuse de construction de corpus annotés, cette démarche présente l'avantage de pouvoir comparer les parseurs entre eux (Ratnaparkhi *et al.*, 1994 ; Pantel et Lin, 1998). Cette exploitation monocorpus pose cependant la question de la stabilité des performances en fonction du type de corpus, comme le mentionnent (Kilgarriff et Grefenstette, 2003 p. 341) : « there is little work on assessing how well one language model fares when applied to a text type that is different from that of the training corpus ». Par ailleurs, il est maintenant bien connu que, dans tout corpus, certaines unités lexicales ont des propriétés syntaxiques de sous-catégorisation spécifiques, qui peuvent donc varier d'un domaine à l'autre (Roland, Jurafsky, 1998 ; Basili *et al.*, 1999). Or peu de travaux relatent des expériences sur la variation des performances de l'analyseur en fonction du type de corpus à traiter, sur le problème de la possible variation intercorpus et sur celui de la nécessaire adaptation des règles ou ressources de l'analyseur à un corpus donné. On peut néanmoins citer (Sekine, 1997 ; Gildea, 2001 ; Slocum, 1986).

Dans cet article, nous nous intéressons à l'acquisition et à l'évaluation sur corpus de données de sous-catégorisation syntaxique. Cette étude est menée dans le cadre du développement de l'analyseur syntaxique de corpus Syntex. Dans la section 2, nous présentons de façon succincte les principes qui sont à la base du développement de l'analyseur, et nous décrivons comment l'analyseur exploite des données de sous-catégorisation syntaxique se présentant sous la forme de probabilités de sous-catégorisation (que telle unité lexicale – verbe, nom ou adjectif – se construise avec telle préposition) pour résoudre les ambiguïtés de rattachement prépositionnel. Dans la section 3, nous décrivons comment ces données sont acquises automatiquement à partir d'un corpus de 200 millions de mots, étiqueté et partiellement analysé syntaxiquement. La section 4 est consacrée à l'évaluation de ces données sur 4 corpus de test de genres variés, pour lesquels nous avons annoté à la main plusieurs centaines de cas de rattachements prépositionnels ambigus. Dans la section 5, nous présentons plusieurs stratégies de désambiguïsation : une stratégie de base, une stratégie endogène qui exploite des propriétés de sous-catégorisation spécifiques, acquises à partir du corpus en cours de traitement, une stratégie exogène qui exploite des propriétés de sous-catégorisation génériques, acquises à partir du corpus de 200 millions de mots, et enfin une stratégie mixte qui utilise les deux types de ressources. Dans la section 6, nous dressons un bilan de cette expérience.

2. Syntex, un analyseur syntaxique de corpus

2.1. *De Lexter à Syntex*

L'analyseur Syntex (Bourigaut et Fabre, 2000) (Bourigault, 2007) a été développé à l'origine pour remplacer le logiciel Lexter, un analyseur syntaxique robuste dédié au repérage des syntagmes nominaux dans les corpus spécialisés et utilisé dans des applications de construction de terminologies ou d'ontologies spécialisées (Bourigault, 1994). Les diverses expérimentations réalisées avec Lexter avaient mis en évidence la nécessité d'étendre la couverture du logiciel à l'extraction des syntagmes verbaux. À partir de ce constat, nous avons décidé d'entreprendre la réalisation d'un nouvel analyseur, avec l'objectif d'en faire un outil opérationnel d'analyse syntaxique de corpus, utilisable dans différents contextes applicatifs, dont la construction de ressources lexicales spécialisées pour des systèmes de traitement de l'information (Bourigault *et al.*, 2004). Syntex traite des corpus de phrases réelles, de taille importante (de quelques centaines de milliers à plusieurs millions de mots). Ceci impose des contraintes d'efficacité (temps de traitement), de robustesse (tolérance aux malformations syntaxiques et aux mots ou structures inconnus, possibilité de rendre des analyses partielles et incomplètes) et d'adaptabilité (prise en compte de certaines propriétés syntaxiques particulières des mots dans des corpus spécialisés). Les principes de base sont les suivants : Syntex analyse des corpus préalablement étiquetés (section 2.2) ; c'est un analyseur en dépendance, organisé sous la forme d'un enchaînement de modules de reconnaissance de relations syntaxiques (section 2.3) ; il exploite de façon combinée des procédures d'apprentissage endogène et des ressources lexico-syntaxiques de sous-catégorisation (section 2.4).

2.2. *Étiquetage préalable*

Syntex analyse des corpus préalablement étiquetés. L'organisation du partage des tâches entre étiquetage morphosyntaxique (attribution d'une étiquette morphosyntaxique aux mots de la phrase) et analyse syntaxique (identification de constituants syntaxiques ou de relations de dépendance syntaxique) est un problème délicat (Voutilaïnen, 1998 ; Prins et Van Noord, 2003). Disposer des étiquettes morphosyntaxiques des mots pour identifier les relations syntaxiques est évidemment très pratique. Mais, dans certains cas, la levée d'ambiguïtés catégorielles exige une analyse syntaxique partielle du contexte large, et la présence d'étiquettes erronées à l'entrée de l'analyseur rend *a priori* impossible une analyse correcte. Essentiellement pour des raisons de disponibilité des forces, nous avons choisi de concentrer nos efforts sur l'analyse syntaxique proprement dite et, comme cela était le cas avec Lexter, de confier la tâche préalable d'étiquetage à un outil extérieur. Syntex prend en entrée les résultats du Treetagger[2], développé à

2 http://www.ims.uni-stuttgart.de

l'Université de Stuttgart. Treetagger est un étiqueteur efficace et robuste. Il présente l'intérêt d'être ouvert, en ce sens qu'il est possible de faire en amont, à sa place, une partie du travail de segmentation en mots et d'étiquetage. Nous avons développé un ensemble de procédures de reconnaissance de mots et d'unités syntaxiques complexes qui viennent poser sur le corpus des étiquettes sur lesquelles le Treetagger s'appuie pour étiqueter les mots environnants. Nous avons aussi introduit dans la chaîne de traitement la possibilité pour l'utilisateur d'intégrer un fichier de règles de segmentation et de préétiquetage, données sous la forme d'expressions régulières, spécifiques au corpus à analyser. Cette fonctionnalité est essentielle quand il s'agit de traiter des corpus comportant des mots inconnus ou des structures « bizarres » (codes de produits, nomenclature d'éléments chimiques, etc.). La frontière entre étiquetage et analyse n'est pas étanche. Dans certains contextes syntaxiques, l'analyseur effectue des retours en arrière sur l'étiquetage en venant modifier des étiquettes attribuées par le Treetagger, par exemple pour la forme *que* (Jacques, 2005).

2.3. *Architecture modulaire*

Syntex ne s'appuie sur aucune grammaire formelle. C'est un analyseur incrémental organisé en « couches » (Constant, 1991 ; Abney, 1996 ; Aït-Moktar *et al.*, 2002). Nous décomposons le problème de l'analyse syntaxique d'une phrase en sous-problèmes élémentaires du type : soit *m* un mot de catégorie C dans la phrase étiquetée et partiellement analysée S, quel est le recteur syntaxique de *m* dans S ? L'analyse s'effectue par un enchaînement en cascade d'une suite de modules qui prennent en charge chacun une relation syntaxique. Chaque module prend en entrée les sorties du module précédent. Cette organisation séquentielle des traitements impose de choisir un ordre dans l'analyse. On est face à un dilemme du type de celui du partage entre étiquetage et analyse. Par exemple, faut-il reconnaître les relations Sujet avant de chercher à identifier les relations de coordination, ou faire l'inverse, ou répartir le traitement à deux moments de la chaîne ? Le choix de l'ordre est un choix difficile qui a un impact fort sur la programmation des différents modules. À l'intérieur de la chaîne d'analyse, les retours en arrière sont possibles. Il peut arriver qu'un module vienne détruire et remplacer une relation syntaxique posée par un module antérieur. Les modules sont constitués d'un ensemble d'heuristiques de parcours de la chaîne étiquetée et partiellement analysée, qui partent d'un régi (resp. recteur) potentiel à la recherche de son recteur (resp. régi). Ces heuristiques exploitent deux contraintes classiques : contrainte de projectivité (les liens de dépendance ne se croisent pas), contrainte d'unicité du recteur (un régi n'a qu'un seul recteur). Les modules sont développés « à la main » par des linguistes informaticiens (dans le langage Perl), selon une méthode empirique qui met en œuvre le recours à la connaissance grammaticale et la réalisation de tests nombreux sur des corpus diversifiés.

2.4. *Ressources lexicales limitées*

L'analyseur Syntex est peu lexicalisé. Nous avons fait le choix initial de la table rase. Cette approche est possible à partir du moment où l'on a choisi de s'appuyer sur les résultats d'un étiqueteur (on bénéficie indirectement des ressources lexicales exploitées par celui-ci). Des informations lexicales sont intégrées dans l'analyseur au fur et à mesure des besoins : liste de locutions prépositionnelles, liste des verbes transitifs, liste des verbes se construisant avec des compléments en *que*, en *de*, etc. Pour résoudre les ambiguïtés de rattachement prépositionnel, l'analyseur exploite des informations de sous-catégorisation associées aux couples (mot, préposition). Depuis l'origine de nos travaux sur l'analyse syntaxique, ces informations sont acquises de façon endogène sur le corpus en cours de traitement (Bourigault, 1993). Les expériences menées sur de nombreux corpus spécialisés ont montré que ces corpus renferment des spécificités lexicales, en particulier que certains mots, fréquents dans le corpus, manifestent des comportements syntaxiques spécifiques et imprédictibles. C'est pourquoi, nous avons porté nos efforts depuis une dizaine d'années sur le développement de procédures d'apprentissage endogène sur corpus qui permettent à l'analyseur d'acquérir lui-même, par analyse du corpus à traiter, des informations de sous-catégorisation spécifiques à ce corpus. Devant les limites inhérentes à l'exploitation exclusive d'informations de sous-catégorisation endogènes, nous travaillons à l'élaboration de ressources générales, susceptibles d'être exploitées pour tout corpus (Frérot *et al.*, 2003) (Frérot, 2005). Dans la section suivante, nous présentons une expérience d'acquisition d'un lexique de probabilités de sous-catégorisation à partir d'un corpus de 200 millions de mots, utilisé par le module de Syntex en charge du rattachement des prépositions.

Le rattachement des prépositions à leur recteur s'effectue en deux passes sur le corpus : (1) recherche des candidats recteurs, (2) choix d'un recteur. Un premier module (*rechercher-candidats*) traite l'ensemble des phrases du corpus, et recherche, pour chaque préposition, le ou les mots susceptibles de régir cette préposition. Ce module est constitué de règles qui reconnaissent un certain nombre de configurations linéaires de mots et de catégories morphosyntaxiques à gauche de la préposition, au sein desquelles sont identifiés des mots susceptibles de régir la préposition. Ces règles s'appuient sur les relations de dépendance placées par les modules antérieurs, et sont capables d'aller chercher des candidats recteurs dans des configurations relativement complexes, incluant par exemple des structures coordonnées ou des incises. Les configurations d'ambiguïtés, définies comme les suites des catégories grammaticales des candidats recteurs, sont très variées. Sur les quatre corpus de test présentés dans la section 4, la configuration 'V N', où seuls un verbe et un nom sont en compétition – configuration traitée dans beaucoup de travaux dont ceux, fondateurs, de Hindle et Rooth (1993) ou ceux plus récents de (Abney *et al.*, 1999) –, ne représente que 50 % des cas dans le corpus littéraire, environ 35 % dans le corpus journalistique et 15 % dans le corpus juridique et le corpus technique. Au cours de la seconde étape du traitement des ambiguïtés prépositionnelles, le second module (*choisir-candidat*) revient sur chaque cas

ambigu et choisit le recteur de la préposition parmi les candidats. Pour cela, le module exploite des informations associées aux couples (candidat, préposition), présentes dans le lexique de sous-catégorisation.

3. Acquisition de propriétés de sous-catégorisation à partir d'un corpus de 200 millions de mots

Les méthodes d'acquisition de propriétés de sous-catégorisation exploitent classiquement des corpus étiquetés de grande taille (Ushioda *et al.*, 1993 ; Manning, 1993 ; Basili, Vindigni, 1998). Le Web est aussi considéré comme source potentielle d'acquisition (Gala Pavia, 2003 ; Volk, 2001). Dans notre étude, nous utilisons comme base d'apprentissage un corpus de 200 millions de mots, constitué des articles du journal *Le Monde*, des années 1991 à 2000 (corpus LM10[3]). Nous considérons que sa taille et sa diversité thématique en font un corpus référentiellement et linguistiquement peu marqué, à partir duquel il est raisonnable de chercher à acquérir des données de sous-catégorisation qui soient relativement génériques. La procédure d'acquisition est adaptée des méthodes d'apprentissage endogène intégrées dans Syntex. La méthode de calcul des probabilités de sous-catégorisation s'appuie sur un ensemble de triplets (recteur, préposition, régi) extraits d'une analyse syntaxique du corpus LM10 effectuée par Syntex. La procédure d'acquisition se déroule en deux étapes, au cours desquelles la même méthode de calcul de probabilités est lancée successivement sur deux ensembles différents de triplets : une étape d'amorçage et une étape de consolidation.

Au cours de l'étape d'amorçage, le module *rechercher-candidats* traite l'ensemble du corpus LM10, qui a été analysé par les modules antérieurs de Syntex, et construit, à partir des cas non ambigus, c'est-à-dire ceux pour lesquels il n'a identifié qu'un seul candidat recteur pour la préposition, un ensemble de triplets (*w*,*p*,*w'*), où *w* est le recteur de la préposition *p*, et *w'* le mot (nom ou verbe à l'infinitif) régi par la préposition. Le module *rechercher-candidats* compte aussi pour chaque mot *w* le nombre d'occurrences dans le corpus où ce mot n'est candidat d'aucune préposition. À l'issue du traitement de l'ensemble du corpus, on dispose des données de fréquence suivantes :

– F(*w*,0) : nombre d'occurrences non ambiguës où le mot *w* ne régit aucune préposition,

– F(*w*,*p*,*w'*) : nombre d'occurrences non ambiguës où le mot *w* régit la préposition *p*, qui elle-même régit le mot *w'*.

3 Ce corpus a été préparé, à partir de fichiers obtenus auprès de l'agence Elra, à l'aide des programmes de balisage et de nettoyage réalisés par Benoît Habert (LIMSI), qui permettent de transformer les fichiers initiaux en un corpus effectivement « traitable » par des outils de Traitement Automatique des Langues. Nous remercions Benoît Habert et le LIMSI de nous avoir permis de bénéficier de ces programmes.

À partir de ces données, un premier ensemble de probabilités de sous-catégorisation P(*w*,*p*) est calculé, selon la méthode décrite plus loin dans la présente section.

Au cours de l'étape de consolidation, le module *choisir-candidat* exploite ce premier lexique et traite à son tour l'ensemble du corpus LM10, analysé par le module *rechercher-candidats*. Il revient sur les cas ambigus et choisit le candidat recteur dont la probabilité de construction avec la préposition, fournie dans le premier lexique, est la plus élevée. À partir de ces nouvelles annotations, un nouvel ensemble de triplets est constitué, qui inclut le précédent et auquel s'ajoutent les triplets (*w*,*p*,*w'*) issus des cas ambigus résolus. De nouvelles données de fréquence F(*w*,*p*,*w'*) et F(*w*,0) sont alors constituées, à partir desquelles un second ensemble de probabilités de sous-catégorisation est calculé, selon la méthode décrite ci-dessous. C'est le lexique construit à l'issue de cette étape de consolidation qui est utilisé dans Syntex.

La méthode de calcul des probabilités est simple. La probabilité est calculée comme une fréquence relative pondérée[4]. Soit T, l'ensemble des triplets (*w*,*p*,*w'*), obtenu à l'issue de l'étape d'amorçage ou de consolidation. Pour un couple (*w*,*p*), on définit $E_{w,p}$ comme l'ensemble des mots *w'* tels que la fréquence F(*w*,*p*,*w'*) est supérieure à 0. On définit la productivité du couple (*w*,*p*), Prod(*w*,*p*), comme le cardinal de l'ensemble $E_{w,p}$, c'est-à-dire comme le nombre de mots différents que régit la préposition *p* quand elle-même est régie par le mot *w*. Nous utilisons ce coefficient pour pondérer la fréquence totale du couple (*w*,*p*). À fréquence égale, plus le couple (*w*,*p*) a été repéré avec des contextes *w'* différents, plus grande est estimée la propension du mot *w* à régir la préposition *p*. L'expérience montre que, dans des corpus thématiques, la très haute fréquence de certains syntagmes très répétitifs incluant le triplet (*w*,*p*,*w'*) vient biaiser la probabilité d'association lexicale entre *w* et *p*. La pondération proposée ci-dessus vise à limiter une telle surestimation et à accorder un poids non seulement à la fréquence de l'association, mais aussi à sa diversité[5]. La formule de calcul de la probabilité pondérée est donnée dans le tableau 1 : F(*w*,*p*) est la fréquence totale du couple (*w*,*p*), F(*w*) est la fréquence totale du mot *w*, et λ est un coefficient de normalisation, choisi de telle sorte que la somme des probabilités associées à un mot donné soit égale à 1.

Le nombre total d'occurrences de triplets (*w*,*p*,*w'*) à partir desquels les probabilités sont calculées est de l'ordre de 6,7 millions à l'issue de l'étape d'amorçage, et de 12 millions à l'issue de l'étape de consolidation (tableau 2). Le nombre total d'occurrences de mots ne régissant pas de préposition est d'environ 87 millions à l'issue de l'étape d'amorçage, et de 95 millions à l'issue de l'étape de

4 Nous n'avons pour le moment pas testé d'autres méthodes de filtrage, comme celle de la distribution polynomiale (Manning, 1993).

5 Par exemple, si le verbe *manger* apparaît cinq fois avec la préposition *avec*, uniquement dans {*manger*,*avec*,*Jean*}, et cinq fois aussi avec la préposition *à*, dans {*manger*,*à*,*restaurant* | *maison* | *self* | *cantine* | *table*}, la probabilité pondérée de {*manger*,*avec*} est de 0,28 et celle de {*manger*,*à*} est de 0,72.

consolidation. Les probabilités ne sont calculées que pour les couples (*w*,*p*) tels que la fréquence totale du mot *w* est supérieure à 20. Un couple n'est retenu dans le lexique de désambiguïsation que si la probabilité dépasse le seuil de 0,01. Le lexique final compte 6 693 verbes différents (chacun pouvant être présent avec plusieurs prépositions), 11 528 noms et 698 adjectifs.

$T = \{ (w,p,w') / F(w,p,w') > 0 \}$, ensemble de triplets
$F(w,p,w')$: nombre d'occurrences où le mot *w* régit la préposition *p*, elle-même régissant le mot *w'*
$F(w,0)$: nombre de cas où *w* ne régit aucune préposition
$E_{W,p,} = \{ w' / F(w,p,w') > 0 \}$, le contexte du couple (*w*,*p*)
$Prod(w,p) = Card(E_{W,p,} p)$, la productivité du couple (*w*,*p*)
$F(w,p) = \sum_{w' \in Ew,p} F(w,p,w')$, le nombre d'occurrences de *w* quand il régit *p*
$F(w)=F(w,0)+ \sum_{p} F(w,p)$, le nombre total d'occurrences de *w*
$P(i,0) = F(w,0)/F(w)$, la probabilité d'occurrence de *w* sans préposition
$P(w,p)=F(w,p) /F(w)*\log(1 + Prod(w,p))/ \lambda$, la probabilité d'occurrence de *w* avec la préposition *p*

Tableau 1. *Méthode de calcul des probabilités de sous-catégorisation*

4. Annotation

De façon générale, le développement d'un analyseur syntaxique robuste exige une méthode de travail qui assume la très grande variabilité des corpus sur le plan syntaxique. Les stratégies et règles des différents modules de Syntex sont à chaque expérimentation élaborées à partir de tests effectués sur plusieurs corpus, aussi diversifiés que possible, pour limiter les biais d'implémentation que pourrait introduire une approche monocorpus. À la variabilité intercorpus, il faut ajouter la variabilité intracorpus. Pour éviter d'élaborer des règles trop dépendantes de telle ou telle configuration syntaxique ou unité lexicale, il faut sur chaque corpus annoter à la main un très grand nombre de cas. Dans le cadre de cette étude, nous avons évalué le lexique de sous-catégorisation sur quatre corpus de test, de genres variés, dans lesquels nous avons validé à la main plusieurs centaines de cas :

- BAL. Le roman *Splendeurs et misères des courtisanes*, d'Honoré de Balzac (199 789 mots) : 672 cas annotés ;

- LMO. Un extrait du journal *Le Monde* (673 187 mots) : 1 238 cas annotés ;

- TRA. Le *Code du travail* de la législation française (509 124 mots) : 1 150 cas annotés ;

- REA. Un corpus de comptes-rendus d'hospitalisation dans le domaine de la réanimation chirurgicale (377 967 mots) : 646 cas annotés.

	Nombre d'occurrences de couples (*w*,*p*) extraits du corpus LM10		Nombre de couples (*w*,*p*) retenus dans le lexique final **f>=20 p>=0.01**	
	Etape d'amorçage	Etape de confirmation		
(verbe,prep+nom)	4 732 143	7 924 140	(verbe,prep+nom)	26 031
(verbe,prep+vinf)	656 083	947 562	(verbe,prep+vinf)	2 501
(nom,prep+nom)	1 207 453	2 671 314	(nom,prep+nom)	15 548
(nom,prep+vinf)	61 957	125 502	(nom,prep+vinf)	294
(adj,prep+nom)	80 641	223 333	(adj,prep+nom)	811
(adj,prep+vinf)	10 900	45 030	(adj,prep+vinf)	40

Tableau 2. *Nombres d'occurrences de couples extraites lors des étapes d'amorçage et de consolidation, et nombres de couples retenus dans le lexique final*

Les règles d'annotation sont les suivantes : (1) ne pas annoter les cas où des erreurs d'étiquetage ont empêché le module de recherche des candidats d'identifier correctement la liste des candidats ; autrement dit on évalue le module de rattachement prépositionnel dans des contextes où les informations sur lesquelles il s'appuie sont justes[6] ; (2) retenir comme valides deux recteurs pour une préposition donnée dans les cas où cela est nécessaire, en particulier pour les constructions à verbe support (*apporter une aide à*) ; (3) valider de manière indifférenciée des groupes prépositionnels arguments ou circonstants. Ce dernier point est important, et peut prêter à controverse, si on ne replace pas la tâche d'annotation dans le contexte de l'évaluation des performances d'un analyseur syntaxique. La distinction argument/circonstant, ou complément essentiel/complément circonstanciel, ne fait pas l'objet d'un consensus dans la communauté linguistique. En dehors des cas triviaux, choisis en général soigneusement pour illustrer cette distinction, la

6 Le nombre de cas ainsi éliminés est très faible (de l'ordre de quelques pour cent). L'examen de ces cas montrent qu'ils n'auraient pas été particulièrement plus difficiles à traiter du point de vue de la résolution de l'ambiguïté de rattachement.

confrontation avec des énoncés réels met à mal la clarté de cette distinction (Fabre, Frérot, 2002 ; Frérot, 2005). Dans ces conditions, la tâche essentielle dévolue à l'analyseur est d'abord de choisir le bon recteur parmi un ensemble de recteurs possibles.

5. Méthode de désambiguïsation

L'algorithme de désambiguïsation mis en œuvre dans le module choisir-candidat est simple. Nous comparons quatre stratégies différentes, selon le type des données de sous-catégorisation qu'elles exploitent.

- Mode *base*. En mode base, le module *choisir-candidat* se contente de choisir comme recteur le premier candidat dans l'ordre linéaire de la phrase, c'est-à-dire le plus éloigné de la préposition[7].

- Mode *exogène*. En mode exogène, le module *choisir-candidat* exploite le lexique de sous-catégorisation construit à partir du corpus LM10 (section 3). Il choisit le candidat dont la probabilité est la plus élevée. On distingue exogène 1 et exogène 2, selon que le lexique utilisé est obtenu après la phase d'amorçage ou après la phase de consolidation.

- Mode *endogène*. En mode endogène, le module *choisir-candidat* exploite le lexique de sous-catégorisation construit à partir du corpus en cours d'analyse[8]. Avant d'exploiter les probabilités de sous-catégorisation, il exploite la liste des fréquences des triplets (*w,p,w'*) construite par le module *rechercher-candidats* : si *p* est la préposition et *w'* le mot qu'elle régit, le module choisit le candidat w_i pour lequel la fréquence F(w_i,*p,w'*) est la plus élevée. Sinon, il choisit le candidat dont la probabilité endogène est la plus élevée.

- Mode *mixte*. Le mode mixte est analogue au mode endogène, à ceci près que le module *choisir-candidat* choisit le candidat qui a la probabilité endogène ou la probabilité exogène la plus élevée.

Dans tous ces modes, la règle par défaut est celle de la stratégie de base, à savoir le choix du premier candidat.

6. Résultats et discussion

Le tableau 3 donne les taux de précision des différentes stratégies de désambiguïsation sur les quatre corpus de test, ainsi que, pour une stratégie donnée, l'écart avec la précision sur le corpus BAL, et, pour un corpus donné, le taux de

7 Globalement - sur l'ensemble des corpus et sur l'ensemble des configurations d'ambiguïté -, cette stratégie est meilleure que celle qui choisirait le candidat le plus proche.
8 Selon la méthode décrite dans la section 3, sans l'étape de consolidation.

réduction de l'erreur par rapport à la stratégie de base. Les principales conclusions que l'on peut tirer de ces résultats sont les suivantes :

- L'apport des ressources génériques est indéniable. Le taux de réduction de l'erreur, qui est de 22,9 % pour le corpus littéraire (BAL), sur lequel la stratégie de base est déjà très performante, s'élève à 54,9 % pour le corpus journalistique (LMO) et à 60,3 % pour le corpus juridique (TRA).

- Il n'y a que sur le corpus médical (REA) que l'apprentissage endogène s'avère être indispensable. La réduction du taux d'erreur est de 45,1 % avec la stratégie endogène, alors qu'elle n'est que de 16 % avec la stratégie exogène 2. Plus que par le domaine couvert, ceci s'explique par le style très particulier utilisé par les médecins pour rédiger les comptes rendus d'hospitalisation, avec un usage abondant de phrases nominales et d'une phraséologie très spécifique.

- Les écarts entre les stratégies exogène 1 et exogène 2 montrent l'intérêt de l'enchaînement des étapes d'amorçage et de consolidation pour acquérir des informations de sous-catégorisation. Par exemple, pour le corpus littéraire (BAL), la taux de réduction de l'erreur passe de 15,9 % à 22,9 % quand on passe du lexique d'amorçage au lexique de consolidation.

- Le résultat le plus remarquable, et le moins attendu, est la très grande homogénéité des taux de précision obtenus avec les stratégies exploitant le lexique générique (stratégies exogène ou mixte) sur les corpus littéraire, journalistique et juridique. Ces taux se tiennent en moins de 2 % (respectivement 86,6 %, 85,9 % et 87.3 % pour ces trois corpus avec la stratégie mixte). L'exploitation de ressources exogènes conduit à un resserrement et à un nivellement par le haut des performances. Chaque corpus trouve dans le lexique générique les informations dont il a besoin pour élever ses performances et ce d'autant plus qu'il a du retard au départ sur les autres corpus. À l'arrivée, le corpus littéraire se fait même doubler par le corpus juridique avec la stratégie mixte.

	BAL		LMO		TRA		REA	
base	83.0 0	0	70.3 0	-12.7	65.5 0	-17.5	59.9 0	-23.1
endogène	83.5 -2.9	0	80.1 -33.0	-3.4	82.3 -48.7	-1.2	78.0 -45.1	-5.5
exogène 1	85.7 -15.9	0	85.5 -51.2	-0.2	85.9 -59.3	-0.2	65.3 -13.5	-20.4
exogène 2	86.9 -22.9	0	86.6 -54.9	-0.3	86.3 -60.3	-0.6	66.3 -16.0	-20.6
mixte	86.6 -21.2	0	85.9 -52.5	-0.6	87.3 -63.2	+0.7	78.3 -45.9	-8.3

Tableau 3. *Taux de précision des différentes stratégies de désambiguïsation sur les quatre corpus de test. À droite, l'écart avec le corpus BAL, au dessous le taux de réduction de l'erreur par rapport à la stratégie de base*

On peut rapprocher ces résultats de ceux, récapitulés dans (Pantel et Lin, 2000), obtenus sur 3 000 cas ambigus extraits de la partie *Wall Street Journal* du Penn TreeBank par différentes méthodes : 81,6 % avec une méthode supervisée utilisant un modèle d'entropie maximale (Ratnaparkhi *et al.*, 1994), 88,1% avec une méthode supervisée utilisant un dictionnaire sémantique (Stetina, Nagao, 1997) et 84,3 % avec une méthode non supervisée utilisant des mots distributionnellement proches (Pantel et Lin, op.cit.). Mais puisque les langues, le type de corpus de test et les conventions d'annotations sont différentes, il est délicat de comparer ces chiffres avec ceux que nous présentons dans le tableau 3.

Les ressources de sous-catégorisation syntaxique construites à partir du corpus LM10 sont exploitées par l'analyseur sans avoir été validées manuellement, et les résultats montrent qu'elles sont performantes pour cette tâche. Il convient de préciser que, sur le plan linguistique, ces propriétés de sous-catégorisation ne sont pas comparables aux descriptions que l'on peut trouver dans des lexiques construits à la main, comme le lexique-grammaire, dans les dictionnaires de langue ou dans les études de psycholinguistique. C'est vrai particulièrement pour les verbes. La probabilité qu'a un verbe de sous-catégoriser telle préposition est calculée à partir de toutes les occurrences (lemmatisées) de ce verbe, sans distinction des différentes acceptions du verbe, alors que l'on sait qu'un même verbe peut avoir des cadres de sous-catégorisation différents selon ses différents sens. Dans le contexte du développement d'un analyseur syntaxique « tout-terrain », l'approximation à laquelle conduit ce lissage des sens est, selon nous, un mal nécessaire.

7. Bibliographie

Abney S., « Partial parsing via finite-state cascades », In: *ESSLLI'96 Workshop on Robust Parsing*, Prague, 1996, p. 8-15

Abney S., Schapire R. et Singer Y., « Boosting applied to tagging and PP attachment », In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (ENMLP 1999), 1999

Aït-Moktar S., Chanod J.-P, Roux C., « Robustness beyond shallowness : incremental deep parsing », *Natural Language Engineering Journal*, 8(2/3), 2002, p. 121-147

Basili R., Pazienza M.-T., Vindigni M., « Adaptative Parsing and Lexical Learning », *Proceedings of VEXTAL'99* , Venise, 1999

Basili R., Vindigni M., « Adapting a Subcategorization Lexicon to a Domain », *Proceedings of the ECML98 Workshop TANLPS*, Chemnitz, Germany, 1998

Bourigault D., « An endogenous Corpus Based Method for Structural Noun Phrase Disambiguation », In *Proceedings of the 6th Conference of the European Chapter of ACL* (EACL), Utrecht, The Netherlands, 1993, p. 81-86,

Bourigault D., *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition*

des connaissances à partir de textes, Thèse de l'Ecole des Hautes Etudes en Sciences Sociales, Paris, 1994

Bourigault D., *Un analyseur syntaxique opérationnel : SYNTEX*, mémoire d'Habilitation à Diriger les Recherches, Université Toulouse-Le Mirail, 2007

Bourigault D., Fabre C., « Approche linguistique pour l'analyse syntaxique de corpus », *Cahiers de Grammaire*, 25, Université Toulouse le Mirail, 2000, p. 131-151.

Bourigault D., Aussenac-Gilles N., Charlet J., « Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas », *Revue d'Intelligence Artificielle* (RIA) , « Techniques Informatiques et structuration de terminologies », Pierrel J.-M. et Slodzian M. (Ed.) , Paris : Hermès. Vol. 18, n°1/2004, p. 87-110

Charniak E., « Statistical Parsing with a Contexte-Free Grammar and Word Statistics », *Proceedings of the AAAI97 Conference*, Brown University, Rhode Island, 1997, p. 598-603

Constant P., *Analyse syntaxique par couche*, Thèse de l'Ecole Nationale Supérieure des Télécommunications, 1991

Fabre C., Frérot C, « Groupes prépositionnels arguments ou circonstants : vers un repérage automatique en corpus », in *Actes de la Conférence TALN*, 2002, p. 215-224.

Frérot C., *Construction et évaluation en corpus variés de lexiques syntaxiques pour la résolution des ambiguïtés de rattachement prépositionnel*, Thèse en sciences du langage de l'Université Toulouse le Mirail, 2005

Frérot C., Bourigault D., Fabre C., « Marier apprentissage endogène et ressources exogènes dans un analyseur syntaxique de corpus. Le cas du rattachement verbal à distance de la préposition 'de' », in *Traitement Automatique des Langues*, 44-3, 2003

Gala Pavia N., *Un modèle d'analyseur syntaxique robuste basé sur la modularité et la lexicalisation de ses grammaires*, Thèse de l'Université de Paris XI, Orsay, 2003

Gildea D., « Corpus Variation and Parser Performance », in *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, Lee L., Harma D., editors 2001, p. 167-202

Hindle D., Rooth M., « Structural Ambiguity and Lexical Relations », *Computational Linguistics*, 19(1), 1993, p. 103-120

Jacques M.-P., « Que : la valse des étiquettes », in *Actes de la conférence Traitement Automatique des Langues Naturelles*, TALN'05, Dourdan, 2005

Kilgarriff A., Grefenstette G., « Introduction to the special issue of Web as Corpus », *Computational Linguistics*, 29(3), 2003, p. 333-338.

Manning C., « Automatic Acquisition of Large Subcategorization Dictionary from Corpora », *Proceedings of the 31st Meeting of the Association for Computational Linguistics*, Columbus, 1993

Pantel P., Lin D., « An Unsupervised Approach to Prepositional Phrase Attachment using Contextually Similar Words ». In K. VijayShanker and Chang-Ning Huang, editors,

Proceedings of the 38th Meeting of the Association for Computational Linguistics, Hong Kong, 2000, p. 101-108

Prins R., Van Noord G., « Reinforcing parser preferences through tagging », in *Traitement Automatique des Langues*, Volume 44, n° 3/2003, p. 121-140

Ratnaparkhi A., « A Linear Observed Time Statistical Parser Based on Maximum Entropy Models ». *Proceedings of the second conference on Empirical Methods in Natural Language Processing*, 1997

Ratnaparkhi A., Reynar J., Roukos S., « A Maximum Entropy Model for Prepositional Phrase Attachment ». *Proceedings of the ARPA Workshop on Human Language Technology*, Morgan Kaufmann, 1994

Roland D., Jurafsky D., « How verb subcategorization frequencies are affected by corpus choice », *Proceedings of Coling-ACL*, Montréal, Canada, 1998, p. 1122-1128

Sekine S., « The domain dependence of parsing ». Proceedings of the Fifth Conference on Applied Natural Language Processing, 1997, p. 96-102.

Slocum J., « How one might Automatically Identify and Adapt to a Sublanguage: An Initial Exploration », in Grishman R. and Kittredge R., eds., *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Lawrence Erlbaum Associates, Hillsdale, N.J., 1986, pp. 195-210

Stetina J., Nagao M., « Corpus-based PP Attachment Ambiguity Resolution with a Semantic Dictionary ». In J. Zhou and K. Church editors, *Proceedings of the 5th Workshop on Very Large Corpora*, Beijing and Hong Kong, 1997

Ushioda A., Evans D., Gibson T., Waibel A., « The Automatic Acquisition of Frequencies of Verb Subcategorization Frames from Tagged Corpora ». In Boguraev, Pustejovsky editors, *Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text*. Columbus, 1993

Volk M., « Exploiting the WWW as a Corpus to Resolve PP Attachment », Proceedings of Conference on Corpus Linguistics, Lancaster, 2001, pp. 601-606

Voutilainen A., « Does Tagging Help Parsing? A Case Study on Finite State Parsing », *Finite-state Methods in Natural Language Processing*, Ankara, 1998