TALN 2004, Fès, 19-21 avril 2004

Ambiguïté de rattachement prépositionnel : introduction de ressources exogènes de sous-catégorisation dans un analyseur syntaxique de corpus endogène

Didier Bourigault et Cécile Frérot

ERSS – Université Toulouse-Le Mirail
Maison de la Recherche
5 allées A. Machado
31058 Toulouse Cedex
{didier.bourigault, cecile.frerot}@univ-tlse2.fr

Résumé – Abstract

Nous présentons les résultats d'expérimentations visant à introduire des ressources lexico-syntaxiques génériques dans un analyseur syntaxique de corpus à base endogène (SYNTEX) pour la résolution d'ambiguïtés de rattachement prépositionnel. Les données de sous-catégorisation verbale sont élaborées à partir du lexique-grammaire et d'une acquisition en corpus (journal *Le Monde*). Nous présentons la stratégie endogène de désambiguïsation, avant d'y intégrer les ressources construites. Ces stratégies sont évaluées sur trois corpus (scientifique, juridique et journalistique). La stratégie mixte augmente le taux de rappel (+15% sur les trois corpus cumulés) sans toutefois modifier le taux de précision (~ 85%). Nous discutons ces performances, notamment à la lumière des résultats obtenus par ailleurs sur la préposition *de*.

We report the results of experiments aimed at integrating general lexico-syntactic resources into a corpus syntactic parser (SYNTEX) based on endogenous learning. We tackle the issue of prepositional phrase attachment. We make use of both French lexico-syntactic resources and automatic acquisition to extract verb subcategorisation data. We describe both the endogenous and hybrid approaches and show how the latter improves the recall rate - +15% in average - but has no impact on the precision rate (~ 85%).

Keywords – Mots Clés

analyse syntaxique automatique, ambiguïté de rattachement prépositionnel, procédures endogènes, ressources exogènes, approche mixte.

automatic parsing, prepositional phrase attachement disambiguation, endogenous learning, exogenous resources, hybrid approach.

1 Introduction

Les études que nous menons sur la levée d'ambiguïtés de rattachement prépositionnel s'inscrivent dans le cadre de la réalisation d'un analyseur syntaxique de corpus à base endogène (SYNTEX, Bourigault, Fabre, 2000, Fabre, Bourigault, 2001). Elles visent à tester si l'introduction de ressources lexico-syntaxiques génériques est compatible avec des procédures non supervisées d'apprentissage sur corpus. En effet, ces procédures, qui n'exploitent aucune autre source d'information que le corpus étiqueté lui-même et s'appuient fortement sur la redondance lexico-syntaxique du corpus, montrent des limites, impliquant de s'interroger sur l'utilisation conjointe de ressources endogènes et exogènes. Notre conviction est que, dans le cadre du développement d'un analyseur de corpus traitant du texte « tout venant », une approche endogène est nécessaire pour résoudre les spécificités du corpus traité, mais qu'une telle approche peut être optimisée par le recours à des ressources exogènes. Ces ressources doivent porter sur des phénomènes linguistiques suffisamment massifs et stables inter-corpus, identifiées comme étant des données verbales transposables d'un corpus à l'autre. En outre, au-delà d'un objectif de performance, nos travaux, qui s'ancrent dans le champ de la linguistique de corpus, entendent affiner notre compréhension des mécanismes syntaxiques à l'œuvre dans les textes, et ainsi alimenter nos réflexions sur les choix d'implémentation de notre analyseur.

Les méthodes utilisées en TAL pour la résolution d'ambiguïtés de rattachement prépositionnel, suivant l'état de l'art proposée par (Gala Pavia, 2003b), concernent principalement la langue anglaise et globalisent en français la problématique du rattachement prépositionnel (traitement indifférencié de la préposition *de*[1] et des autres prépositions). Si la plupart des travaux problématisent à l'heure actuelle l'appel à des données externes, le recours à des ressources lexicales constituées *a priori* privilégie non pas les données syntaxiques (cadres de sous-catégorisation) mais les ressources sémantiques (Basili et al.,1997, Collins, Brooks, 1995, Gaussier, Cancedda, 2001, Nieman, 1998). En outre, c'est l'acquisition de propriétés plutôt que la projection de ressources constituées *a priori* que mettent en avant les méthodes exploitant des données syntaxiques externes (Volk 2001, Volk, 2002, Gala Pavia, 2003a).

L'expérience que nous relatons dans cet article vise à exploiter conjointement des ressources endogènes et exogènes, de manière à définir une approche mixte optimale. Nous avons déjà testé cette approche dans le cas du rattachement verbal à distance de la préposition *de* (Frérot et al., 2003) et nous l'étendons ici à la levée d'ambiguïtés de rattachement des prépositions *à*, *dans*, *sur,* dans la configuration syntaxique V SN SP[2]. Nous présentons les modes de constitution des ressources exogènes ainsi que les corpus annotés syntaxiquement pour

---

[1] Or la préposition *de* requiert un traitement différent des autres prépositions car elle fait très peu souvent l'objet d'un rattachement à distance, à un verbe en particulier, étant majoritairement un élément de syntagme nominal.

[2] V SN SP : V:verbe à l'actif, SN : syntagme nominal (incluant éventuellement des adjectifs, participes passés et des Nom de Nom), SP :syntagme prépositionnel (nominal) introduit par une préposition *à*, *dans*, *sur*. Exemples de cas étudiés : *consigner des informations confidentielles dans un ouvrage, reporter une visite à Meudon, demander des informations nécessaires à l'étude*.

l'évaluation. Puis nous décrivons les stratégies de désambiguïsation et discutons les performances, notamment à la lumière des résultats obtenus par ailleurs sur la préposition *de*.

2 Constitution des ressources exogènes

2.1 Exploitation du lexique-grammaire

L'objectif de complétude qui a prévalu à la réalisation du lexique-grammaire (LG) ainsi que le français « standard » qui a servi de référence aux concepteurs de la ressource ont largement motivé notre choix de constituer, à partir du LG, une ressource lexico-syntaxique générique portant sur la sous-catégorisation verbale. Le français « standard » du LG est révélé essentiellement par introspection ; en effet, dans ce lexique, le degré d'appartenance à la langue d'une séquence de mots (*i.e.* sa grammaticalité) est estimé au niveau empirique par le jugement d'acceptabilité. Les exemples de travail sont inventés et soumis au jugement d'acceptabilité porté par un sujet, en l'occurrence le linguiste lui-même le plus souvent (Boons et al., 1976a). C'est précisément ce type de ressources que nous souhaitons opposer à des données acquises sur très gros corpus (section 2.2), et mettre à l'épreuve de la réalité des corpus.

Pour construire la ressource à intégrer dans l'analyseur, nous avons extrait des tables l'information selon laquelle un verbe sous-catégorise les prépositions *à, dans, sur* à distance. Autrement dit, les verbes extraits appartiennent à la structure que nous ramenons ici à la forme canonique N_0 *V* N_1 *Prep(à, dans, sur)* N_2. Nous avons ainsi construit une liste de 1637 couples (verbe, préposition) à partir de plusieurs tables[3] . Cette extraction est le résultat d'un filtrage et d'une « décontextualisation » des données, imposés par leur intégration dans notre analyseur, dans la mesure où les contraintes de type transformationnel et sémantique décrites dans les tables ne peuvent être prises en charge par SYNTEX (Frérot et al., 2003).

2.2 Acquisition en corpus *Le Monde*

Dans cette expérience, nous avons souhaité confronter des ressources exogènes constituées selon une démarche introspective, indépendamment de tout corpus et tout domaine, à des ressources élaborées à partir de corpus réels, et donc censées refléter une certaine réalité langagière. Nous avons choisi comme corpus d'apprentissage un échantillon de 40 millions de mots du journal *Le Monde* (années 1995-1996). Nous ne prétendons aucunement que ce corpus soit représentatif de la « langue générale », mais nous considérons que sa taille et son hétérogénéité thématique en font un corpus référentiellement et linguistiquement peu « marqué », à partir duquel il est possible d'acquérir des données de sous-catégorisation relativement génériques. Ces données sont extraites des résultats de l'analyseur SYNTEX sur ce corpus, à partir uniquement de cas de rattachement *non ambigus*, c'est-à-dire de contextes où

[3] Les tables utilisées sont : 36DT, 38LD, 38LH, 38LR (Guillet A, Leclère C, 1992), 9, 10, 11, 3, 4, 6, 16 (Gross, 1975).

l'analyseur a identifié un verbe comme *seul* recteur potentiel pour l'une des prépositions *à*, *dans* ou *sur* [4]. L'extraction a produit 215 193 triplets {verbe, préposition, nom}, filtrés selon les critères de fréquence et de productivité : un triplet n'est conservé que s'il a été extrait au moins deux fois du corpus ; le couple de sous-catégorisation (verbe, préposition) n'est conservé que si au moins deux triplets avec des noms différents ont été extraits. L'ensemble de ces couples constitue la liste LM.

Le tableau 1 indique le nombre de couples (verbe, préposition) propres à chaque ressource (LM, LG) ainsi que le nombre commun aux deux (liste EXO).

	LM	LG	Intersection Liste EXO
à	590	837	293
dans	422	488	143
sur	249	312	65
Total	1 261	1 637	501

Tableau 1 – Nombre de verbes des listes LM, LG et EXO

3 Constitution de corpus annotés pour l'évaluation

3.1 Une méthode empirique

L'analyseur SYNTEX est développé dans le cadre d'une démarche empirique qui associe un travail d'analyse sur corpus à une réflexion linguistique, ainsi qu'une référence à des théories et ressources existantes. Le développement d'un analyseur syntaxique supposé être « tout terrain » exige une méthode de travail qui assume la très grande variabilité des corpus sur le plan syntaxique. C'est pourquoi les stratégies et règles des différents modules sont à chaque expérimentation élaborées à partir de tests effectués sur plusieurs corpus, aussi diversifiés que possible, pour limiter les biais d'implémentation que ne manquerait pas d'introduire une approche mono-corpus. En outre, à la variabilité inter-corpus, il faut ajouter la variabilité intra-corpus. 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 cette démarche d'apprentissage manuel, il ne peut y avoir de distinction entre corpus de test et corpus d'apprentissage, comme c'est la règle pour le développement d'outils d'apprentissage automatique. Pour la présente étude, les stratégies ont été élaborées sur un ensemble de 3 000 cas annotés, sur 3 corpus différents. Les corpus annotés syntaxiquement

Les auteurs sont intervenus sur l'ensemble des corpus et ont fixé un certain nombre de règles d'annotation : *(i)* ne pas valider de cas présentant des erreurs d'analyse des modules antérieurs, en particulier des erreurs d'étiquetage (le module de rattachement est évalué uniquement dans des contextes où les informations sur lesquelles il s'appuie sont justes) ; *(ii)*

[4] Les contextes d'acquisition non ambigus sont les suivants : verbe au passif, participe passé épithète, complément d'objet à distance et clitique.

en cas de recteurs (corrects) multiples, annoter l'ensemble des recteurs ; il s'agit notamment des cas ambigus de double structure ou le nom et le verbe peuvent régir la préposition (*apporter une aide à, effectuer un atterrissage sur la piste) ; (iii)* ne pas sur-représenter certains cas trop spécifiques au corpus et récurrents afin de limiter le biais dans l'évaluation. C'est par exemple le cas dans un des corpus de test (CTRA) où les rattachements de certains participes passés à une préposition sont très nombreux (*visées à l'article, mentionnés au paragraphe*). A la fin de ce travail d'annotation, non seulement nous disposons d'une base de cas annotés pour l'évaluation, mais l'analyse de ces nombreux cas alimente nos réflexions sur les règles à implémenter et à faire évoluer. Grâce à cette base d'annotation, nous nous sommes assurés d'une relativement bonne couverture de nos règles, de même que nous avons une idée relativement précise de leurs limites.

Nous avons évalué les stratégies de désambiguïsation sur trois corpus :

- VOLC : ce corpus de vulgarisation scientifique dans le domaine de la volcanologie comprend environ 400 000 mots. Le genre de textes qui le compose en fait un corpus hétérogène car il regroupe des articles issus de la presse semi-spécialisée (et encore faut-il distinguer les textes extraits de revues telles que *Science et vie* – écrits par des spécialistes pour des non-spécialistes – des textes issus de revues telles que *Pour la Science* – écrits par des spécialistes pour des initiés), de la presse généraliste (*Le Monde*) mais également des ouvrages pédagogiques et des plaquettes d'exposition.

- CTRA : ce corpus juridique d'environ 200 000 mots est le Code du Travail. C'est un corpus spécialisé, thématiquement homogène, marqué par une terminologie et phraséologie propres au domaine.

- MOND : ce corpus journalistique d'environ 700 000 mots rassemble des textes du journal *Le Monde* (année 98). C'est un corpus très hétérogène sur le plan thématique (rubriques *sports, culture, politique, économique…*).

Le tableau 2 illustre une première disparité syntaxique des trois corpus, selon le point de vue contrastif qui est le nôtre dans cette étude. Pour les prépositions étudiées *à, dans, sur,* le corpus VOLC affiche une plus grande propension au rattachement au verbe que le corpus CTRA, dans lequel le rattachement à un participe passé est massif. Ce point expliquera une partie des résultats de l'évaluation (section 4).

	V	N	A	P	Total
VOLC	710 (64.2%)	280 (25.3%)	32 (2.9%)	84 (7.6%)	1106
CTRA	488 (46.6%)	296 (28.3%)	110 (10.5%)	153 (14.6%)	1047
MOND	591 (58.7%)	329 (32.7%)	28 (2.8%)	59 (5.9%)	1007

Tableau 2. Catégories des recteurs (V:verbe, N:nom, A:adjectif, P:part. passé)

Par ailleurs, une première analyse des cas annotés manuellement montre que la disparité des recteurs impliqués dans les rattachements corrects est beaucoup moins grande dans le corpus CTRA que dans VOLC et MOND. Le millier de cas validés est couvert par 258 couples (recteur ;préposition) différents pour le corpus CTRA, contre respectivement 553 et 501 pour les corpus VOLC et MOND. En guise d'illustration, nous donnons ci-dessous les listes des couples (recteur ; préposition) validés les plus fréquents dans les bases de cas annotés manuellement (fréquence supérieure ou égale à 5). Cette moins grande disparité des recteurs dans le corpus CTRA laisse augurer de meilleures performances de l'apprentissage endogène sur ce corpus, puisque les techniques endogènes d'apprentissage sur corpus, basées sur la redondance lexico-syntaxique (Bourigault, Fabre, 2000), ont d'autant plus de chances d'apprendre des informations sur des recteurs potentiels que leur fréquence est élevée.

- **corpus VOLC** (26 couples, 182 cas) : V;donner;à (18), V;devoir;à (11), V;dévaler;à (11), V;prendre;dans (9), P;devoir;à (9), V;soumettre;à (8), V;faire;à (8), V;mettre;à (8), V;plonger;dans (7), N;information;sur (7), N;accès;à (7), V;projeter;à (7), V;passer;à (6), V;trouver;dans (6), V;provoquer;sur (5), P;situer;à (5), V;publier;dans (5), V;relier;à (5), V;comparer;à (5), V;porter;à (5), V;creuser;dans (5), V;déposer;sur (5), V;faire;dans (5), V;former;dans (5), V;laisser;à (5), V;retrouver;dans (5).

- **corpus CTRA** (46 couples, 455 cas) : V;exercer;dans (31), A;relatif;à (24), P;prévoir;à (23), N;droit;à (21), N;accès;à (20), A;nécessaire;à (18), N;infraction;à (16), V;transmettre;à (16), A;égal;à (15), V;porter;à (14), V;adresser;à (13), V;apporter;à (13), V;notifier;à (11), V;assurer;à (10), V;remettre;à (9), P;viser;à (9), P;mentionner;à (8), N;adaptation;à (8), A;inférieur;à (8), N;exposition;à (8), P;définir;à (8), V;donner;à (8), P;fixer;à (7), V;faire;à (7), N;aptitude;à (7), N;avis;sur (7), V;exposer;à (7), V;envoyer;à (7), V;servir;à (7), V;présenter;à (7), V;mettre;à (6), V;reverser;à (6), N;aide;à (6), A;conforme;à (6).

- **corpus MOND** (23 couples, 153 cas) : V;donner;à (22), V;apporter;à (11), V;faire;à (8), N;passage;à (7), V;demander;à (7), N;rôle;dans (6), N;réponse;à (6), V;mettre;à (6), V;remettre;à (6), V;trouver;dans (6), A;relatif;à (6), V;ouvrir;à (6), N;accès;à (6), N;appel;à (5), V;laisser;à (5), N;soutien;à (5), V;lancer;à (5), V;consacrer;à (5), V;réduire;à (5), V;jouer;dans (5), V;maintenir;dans (5), V;imposer;à (5), V;adresser;à (5).

4 Stratégies de désambiguïsation et évaluation

Nous présentons les résultats concernant l'évaluation de trois stratégies différentes. Précisons que pour l'ensemble des stratégies, si l'analyseur ne dispose pas (ou pas suffisamment) d'indices, il ne prend pas de décision de rattachement par défaut. Chaque stratégie a été testée en confrontant quatre lexiques exogènes différentes : le lexique construit à partir du lexique-grammaire (LG), le lexique acquis à partir du journal *Le Monde* (LM), le lexique constitué des données communes à LG et LM (désormais U) et enfin le lexique regroupant LG et LM (désormais I). Nous désignerons ces quatre lexiques par la liste EXO.

- *Stratégie exogène*. La stratégie exogène est une stratégie de base dans laquelle seule la liste EXO est utilisée : si le verbe candidat appartient à la liste EXO (qui est donc instanciée par LM-LG-U-I), il est choisi comme recteur. Sinon, aucun candidat n'est choisi.

- *Stratégie endogène*. La stratégie endogène n'exploite aucune liste exogène, mais s'appuie sur des indices endogènes, calculés par l'analyseur à partir des cas de rattachement non ambigus. Deux principaux types d'indices sont pris en compte, les indices *arg* et *prod*. Soit P la préposition à rattacher, soit N le nom qu'elle régit et soit C un candidat recteur. L'indice *arg* pour C vaut 1 si le candidat recteur C a été identifié dans un cas non ambigu comme recteur de la préposition P, elle-même régissant le nom N. L'indice *prod* pour C est calculé à partir de la productivité du couple (C ;P), qui est égale au nombre de noms N *différents* régis par la préposition P dans les cas non ambigus où elle-même est régie par le candidat C. L'indice *prod* pour C vaut 1 si la productivité du couple (C ;P) est supérieure à un certain seuil[5]. La stratégie endogène privilégie d'abord les candidats dotés d'un indice *arg* à 1, puis ceux dotés d'un indice *prod* à 1. En cas de concurrence le candidat le plus proche de la préposition est choisi. En l'absence de tout indice (*i.e.* lorsque tous les indices valent 0 pour tous les candidats), aucun recteur n'est choisi.

- *Stratégie mixte*. La stratégie mixte exploite à la fois l'indice *prod* et l'appartenance à la liste EXO. Elle est analogue à la stratégie endogène, avec pour seule modification que, pour les verbes, l'indice *prod* vaut 1 si la productivité du couple (C ;P) est supérieure au seuil *ou* si le couple (C ;P) appartient à la liste EXO.

Le tableau 3 présente les résultats de l'évaluation de ces trois stratégies sur les différents corpus de test.

Corpus VOLC									
Stratégie	endo	exo				mix			
liste EXO		LG	LM	U	I	LG	LM	U	I
Précision	**81.8**	82.7	81.7	**82.3**	78.8	82.4	82.3	**82.3**	**82.4**
Rappel	**39.1**	43.1	58.4	**61.9**	60.5	53.4	62.1	**65.2**	**53.4**
Corpus CTRA									
Stratégie	endo	exo				mix			
liste EXO		LG	LM	U	I	LG	LM	U	I
Précision	**86.8**	82.7	81.7	75.1	81.6	82.4	82.3	**85.2**	**85.9**
Rappel	**75.2**	43.1	58.4	64.8	51.0	53.4	62.1	**80.8**	**79.3**
Corpus MOND									
Stratégie	endo	exo				mix			
liste EXO		LG	LM	U	I	LG	LM	U	I
Précision	**86.8**	82.7	81.7	78.3	85.1	82.4	82.3	**83.8**	**86.0**
Rappel	**54.0**	43.1	58.4	62.0	45.6	53.4	62.1	**70.9**	**65.6**

Tableau 3. Résultats des stratégies endogène (endo), exogène (exo) et mixte (mix)

Globalement, les ressources exogènes, qu'elles soient issues du lexique-grammaire ou acquises en corpus, augmentent le taux de rappel sans toutefois dégrader le taux de précision, et ce, tous corpus confondus. Dans la stratégie endogène, l'analyseur ne dispose pas (ou pas suffisamment) d'informations pour prendre une décision (cas de rattachement au verbe) alors que dans la stratégie mixte, l'information externe au corpus permet à l'analyseur de se

[5] Le seuil a été fixé à 4 pour cette expérience. Dans le cas d'un verbe ou d'un nom déverbal, la productivité du verbe et celle de son nom déverbal sont cumulées. Cette information sur les noms déverbaux nous est fournie par le lexique Verbaction, élaboré par Nabil Hathout de l'ERSS.

prononcer. Quant au lexique à exploiter (liste EXO, c'est-à-dire qu'il s'agisse des données du LG exclusivement, du LM exclusivement, de l'intersection des deux, ou de l'union), on remarque que son impact est mineur sur les performances de l'analyseur, les différences entre LG-LM-U-I étant extrêmement ténues ; l'union des deux listes reste néanmoins la plus efficace en termes de rappel, l'intersection des deux listes, en termes de précision. Enfin, on note que la stratégie endogène est meilleure sur le corpus CTRA que sur les corpus VOLC et MOND (aucune différence notable sur ces deux derniers corpus) et on peut avancer ici le paramètre de la redondance lexico-sytnaxique pour expliquer ces données ; en effet, comme nous l'avons indiqué à la section 3.2, les techniques endogènes d'apprentissage sur corpus, basées sur la redondance lexico-syntaxique, ont d'autant plus de chances d'apprendre des informations sur des recteurs potentiels que leur fréquence est élevée, ce qui est précisément le cas pour CTRA.

La dégradation des résultats entre la stratégie exogène et la stratégie endogène sur le corpus VOLC appelle quelques commentaires. On peut proposer la conjonction de deux éléments pour expliquer cette dégradation. Tout d'abord, le corpus VOLC est un corpus relativement disparate qui rassemble des textes ayant tous trait à la volcanologie, mais de genre relativement différents : articles scientifiques, articles de vulgarisation, ouvrages pédagogiques, plaquettes d'exposition… La redondance lexicale, sur laquelle s'appuie l'apprentissage endogène, n'est pas aussi forte que sur un corpus plus homogène. Ensuite, les rattachements au verbe sont proportionnellement plus importants dans le corpus VOLC que dans les deux autres corpus (tableau 2). Cette caractéristique désavantage la stratégie endogène, dont une des forces vient de ce qu'elle peut acquérir sur le corpus des informations sur les noms et les adjectifs.

Dans la stratégie mixte, le taux de précision est particulièrement intéressant à contraster avec les résultats obtenus par ailleurs sur le rattachement verbal à distance de la préposition *de* (Frérot et al., 2003). Dans cette étude, la ressource exogène (données issues du lexique-grammaire) a montré un gain en rappel et en précision de 8% et l'analyseur privilégie par défaut le rattachement au nom car la préposition *de* est sous-catégorisée par le verbe dans des proportions nettement moindres que les prépositions *à*, *dans*, *sur* (corpus *Le Monde,* 2% de rattachement pour *de* contre environ 60% pour *à*, *dans*, *sur*). Dans l'expérience que nous avons décrite ici, l'analyseur ne prend aucune décision pour les cas où il ne dispose pas d'indice suffisant. Or le phénomène de sous-catégorisation verbale supplantant (nettement) la sous-catégorisation nominale ou adjectivale, ceci pourrait justifier linguistiquement un rattachement par défaut au verbe.

5 Conclusion et perspectives

Ces expérimentations sur l'introduction de ressources exogènes dans un analyseur de corpus à base endogène nous ont permis de mieux cerner la compatibilité potentielle des deux approches et de définir une approche mixte valide, qui reste bien entendu à affiner. Nous allons d'une part prolonger cette étude par une analyse détaillée de la couverture lexicale respective des deux lexiques LG et LM, et nous allons d'autre part analyser l'impact que peut avoir la variation des seuils de fréquence et de productivité dans l'acquisition automatique de données à partir de corpus.

Nous travaillons actuellement à la mise au point d'une méthode mixte qui intègre des données d'acquisition sur très gros corpus pour les catégories autres que verbale. En effet, de récentes

expérimentations sur l'apport de sous-catégorisation adjectivale dans l'analyseur ont montré l'impact d'un e liste (réduite) d'adjectifs sur les performances de l'analyseur. Ainsi, cibler le type de données de sous-catégorisation à exploiter (adjectif, nom, verbe) s'avère essentiel dans notre approche, ce qui implique d'étendre la constitution de ressources exo gènes aux adjectifs et noms. Par ailleurs, l'introduction d'un indice reflétant la probabilisation des évènements linguistiques en corpus s'impose désormais si l'on souhaite progresser dans l'élaboration de méthodes qui rendent compte du comportement réel des mots en corpus ; nous travaillons donc à la mise au point d'un indice, calculé à partir de la productivité et de la fréquence des mots en corpus, qui permette l'estimation de rattachements pondérée en fonction du corpus (c'est-à-dire la probabilité pour un mot de sous-catégoriser une préposition dans un corpus donné). Cette approche probabiliste en corpus soulève notamment la délicate question de sa compatibilité avec une « ressource catégorique » (Habert, Zweigenbaum, 2002), telle que celle constituée à partir du lexique-grammaire.

Remerciements

Les auteurs remercient vivement Cécile Fabre de sa relecture attentive de l'article.

Références

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

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, 159-170.

COLLINS M., BROOKS J. (1995), Prepositional phrase attachment through a backed-off model. Actes du *Workshop on Very Large Corpora*.

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

FRÉROT C., BOURIGAULT D., FABRE C. (2003), Marier apprentissage endogène et ressources exogènes dans un analyseur syntaxique de corpus. Le cas du rattachement verbal à distance de la préposition « de », *Traitement Automatique des Langues,* 44-3, à paraître.

GALA PAVIA N. (2003a), Une méthode non supervisée d'apprentissage sur le Web pour la résolution d'ambiguïtés structurelles liées au rattachement prépositionnel. Actes de *Conférence sur le Traitement Automatique des Langues Naturelles,* 353.358.

GALA PAVIA N. (2003b), Un modèle d'analyseur syntaxique robuste basé sur la modularité et la lexicalisation de ses grammaires, Thèse de Doctorat, Université Paris XI, Orsay.

GAUSSIER E., CANCEDDA N. (2001), Probabilistic Models for PP-attachment Resolution and NP Analysis, Actes de *Association for Computational Linguistics, Computational Natural Language Learning Workshop*, 45-52.

GROSS M. (1975), *Méthodes en Syntaxe,* Hermann, Paris.

GUILLET A., LECLÈRE C. (1992), *La structure des phrases simples en français – les constructions transitives locatives*, Droz.

HABERT B., ZWEIGENBAUM P. (2002). « Régler les règles », *Traitement automatique des langues,* 43-3, 83-105.

NIEMAN M. (1998), Determining PP attachment through semantic associations and preferences, In D.Estival (ed.) *Abstracts for the ANLP Post Graduate Workshop,* 25-32.

VOLK, M. (2002), The Automatic Resolution of Prepositional Phrase Attachment Ambiguities in German. In *Habilitationsschrift,* Université de Zurich.

VOLK, M. (2001), Exploiting the WWW as a corpus to resolve PP attachment, Actes de *Corpus Linguistics Conference*, 601-606.