# Approche linguistique pour l'analyse syntaxique de corpus

Didier Bourigault & Cécile Fabre*

*Cet article présente les premiers résultats d'un analyseur syntaxique de corpus, SYNTEX, conçu pour effectuer une analyse syntaxique superficielle des textes exploitable pour la constitution de ressources lexicales. Nous décrivons la tâche initiale de résolution du rattachement prépositionnel, effectuée de manière endogène, en exploitant un faisceau d'indices linguistiques. Parmi ceux-ci, nous définissons une mesure spécifique de productivité, qui permet d'évaluer la régularité de l'association entre un mot recteur potentiel et une préposition dans le corpus. Cette mesure, très efficace dans la tâche de désambiguïsation, s'avère également prometteuse pour mettre en évidence différents types de rattachement prépositionnel. Les principes de conception de cet outil sont enfin mis en regard avec quelques hypothèses fondamentales de la théorie syntaxique.*

*In this paper we report the first results of SYNTEX, a corpus-based syntactic analyser, which performs a shallow parsing of texts for the construction of lexical resources. We focus on the first stage of this process, prepositional attachment resolution, which is performed without the use of prior knowledge; among the set of linguistic cues used by the analyser, we define a productivity measure, allowing to assess the degree of regularity of the association between a potential controller and a preposition within the corpus. This measure proves to be very efficient in the disambiguation task, and also helps to differentiate various types of prepositional attachments. Finally, principles that have guided the conception of the analyser are examined from the viewpoint of some syntactic theory's hypothesis.*

* Equipe de Recherche en Syntaxe et sémantique (UMR 5610 CNRS - Université Toulouse-Le Mirail) Opération "Sémantique et corpus".

*Cahiers de Grammaire 25 (2000), « Sémantique et Corpus », pp. 131-151*

## 1. Introduction

L'un des domaines de recherche de l'opération "Sémantique et corpus" de l'ERSS est la construction de ressources lexico-conceptuelles à partir de corpus spécialisés. Les types de ressources envisagés sont très divers : dictionnaire spécialisé, lexique sémantique pour l'extraction d'information, thesaurus pour l'indexation, lexique bilingue pour la traduction, ontologie pour la recherche d'information. Toutes ces ressources ont des fonctions différentes et sont donc constituées selon des objectifs différents. Mais dans tous les cas, d'une part elles doivent être élaborées à partir de l'analyse de corpus du domaine envisagé, et d'autre part, leur structure se rapproche le plus souvent de celle de classes sémantiques et de réseaux lexicaux, basés pour l'essentiel sur les relations classiques d'hyperonymie, de méronymie, de synonymie. Les recherches en linguistique de corpus doivent fournir des éléments théoriques, méthodologiques et logiciels pour cette activité de construction de ressources lexicales. Notre objectif est alors de développer des méthodes d'analyse linguistique permettant de construire de façon aussi réglée et systématique que possible ces ressources à partir de l'analyse linguistique de corpus spécialisés. C'est un axe fort de l'opération "Sémantique et corpus" de l'ERSS, depuis les travaux d'Anne Condamines sur l'étude linguistique de la terminologie (Condamines 1996) et sur la notion de Base de Connaissances terminologiques. Une des volontés qui animent l'opération est de provoquer des échanges et d'établir des passerelles entre des travaux théoriques en linguistique descriptive et des travaux appliqués en sémantique lexicale sur corpus.

Parmi les travaux menés par les linguistes de l'opération dans cette optique, certains visent à la réalisation d'outils de traitement automatique des langues destinés à être utilisés pour la construction de ressources lexico-conceptuelles à partir de corpus. Deux types de recherches sont concernés, correspondant aux deux types de structuration de ces ressources, la structuration en réseau sémantique et la structuration en classes sémantiques. Un premier axe de recherche contribue à la mise au point d'outils de repérage de structures linguistiques en corpus. Les efforts ont porté principalement sur le repérage de marqueurs de la relation d'hyperonymie et sur celui de contextes définitoires (Condamines & Rebeyrolle 2001) (Tanguy & Rebeyrolle 2001). Un second axe vise à outiller une démarche d'analyse distributionnelle "à la Harris", en réalisant un analyseur syntaxique robuste capable de repérer dans un corpus spécialisé les relations de dépendance syntaxique à partir desquelles seront effectués les regroupements distributionnels. C'est aux travaux effectués autour du développement de cet analyseur qu'est consacrée la suite de cet article.

**2. Analyse distributionnelle et analyse syntaxique automatique**

A partir de la notion de sous-langage définie par Harris (1968), a été développée une méthode d'analyse de corpus pour mettre au jour les classes de mots et les patrons syntaxiques caractéristiques d'une langue spécialisée. Pour reprendre les termes de N. Sager : *« Si on applique à un corpus de textes d'un secteur scientifique des méthodes de linguistique descriptive similaires à celles utilisées pour le développement d'une grammaire d'une langue dans son ensemble, on obtient des motifs précis de cooccurrences de mots à partir desquels on peut définir des sous-classes de mots et des séquences de ces sous-classes qui sont caractéristiques* [...]. *Ces catégories lexicales et formules syntaxiques de la grammaire du sous-langage sont étroitement corrélées aux classes d'objets du monde et aux relations qui sont propres à ce domaine. »* (Sager, Friedman & Lyman 1987, cités par Habert et al., 1997, p. 149). Dans les travaux de Harris, l'analyse et la normalisation syntaxiques sont effectuées à la main. Dans la communauté du traitement automatique des langues, un certain nombre de travaux ont été menés pour exploiter le principe de l'analyse distributionnelle, dans lesquels l'analyse syntaxique initiale est réalisée par un analyseur syntaxique (Hirshman & al. 1975) (Sager, Friedman & Lyman, 1987) (Hindle 1990) (Greffenstette 1994).

Dans la communauté française, ce type d'approche connaît un regain d'intérêt depuis le milieu des années 90, principalement dans la communauté de l'ingénierie des connaissances. Citons les travaux H. Assadi (Assadi & Bourigault 1995) et ceux de B. Habert et A. Nazarenko (1996). Dans ces travaux, l'analyseur syntaxique utilisé pour extraire les relations de dépendance est *LEXTER*, un analyseur syntaxique robuste "tout terrain" dédié au repérage des syntagmes nominaux dans les corpus spécialisés (Bourigault 1994).

Les diverses expérimentations réalisées avec *LEXTER* ont mis en évidence la nécessité d'étendre la couverture du logiciel à l'extraction des syntagmes verbaux. Ceci exige une refonte complète du logiciel. Dans cet article, nous présentons les recherches que nous menons pour développer, à partir des principes de base de *LEXTER*, un nouvel analyseur syntaxique à large couverture pour le français, l'analyseur *SYNTEX*. La suite de cet article est composée de trois parties. Dans la section 3, nous précisons quel est le statut méthodologique de l'analyseur et comment sa conception s'inscrit dans un travail de recherche en linguistique. Nous illustrons le travail d'implémentation dans la section 4, en présentant la notion de productivité et les heuristiques basées sur corpus pour l'acquisition des propriétés de complémentation des unités lexicales du corpus. Nous concluons l'article par une tentative de rapprochement entre les hypothèses qui guident la conception de l'analyseur et certains principes de la théorie syntaxique (section 5).

### 3. Statut de l'analyseur

#### 3.1. L'analyseur comme outil d'aide à l'interprétation de corpus

*« Un analyseur, considéré en dehors du cadre théorique qui préside à sa conception, est une machine célibataire. Un analyseur, considéré à partir du cadre qui préside à sa conception, est le reflet de ses options fondamentales. »* Nous souscrivons à cette affirmation de J.-M. Marandin (1993, p. 31), et tentons dans cette section de préciser quel statut nous donnons à l'analyseur au sein d'un programme de recherche en linguistique de corpus.

La fonction de notre analyseur est d'identifier des relations de dépendance entre mots et d'extraire d'un corpus des syntagmes (verbaux, nominaux, adjectivaux). Le résultat de l'analyse se présente sous la forme d'un réseau de dépendance, dans lequel chaque syntagme extrait est relié à sa tête et à son expansion syntaxiques (figure 1). Ces relations de dépendance permettent d'effectuer automatiquement des regroupements distributionnels : par exemple la liste de tous les compléments de tel verbe ou la liste des adjectifs modifieurs de tel nom, qui constituent des amorces de classes sémantiques. A titre d'illustration, l'analyse du réseau présenté sur la figure 1 suggère un regroupement des noms *alluvion*, *sable* et *lave* qui sont tous les trois arguments des verbes *disparaître sous* et *creuser dans*. Ces regroupements devront être corrigés, affinés, ou rejetés par l'analyste, utilisateur de l'outil, en fonction du type de description lexico-conceptuelle qu'il cherche à construire à partir de l'analyse du corpus.

Notre programme de recherche s'inscrit donc dans le paradigme de l'*aide à l'interprétation de textes*. Notre analyseur n'est pas un module dans un dispositif visant la compréhension automatique, mais un outil dont les résultats doivent aider un analyste humain à construire à partir de l'analyse du texte une interprétation qui se matérialisera sous la forme d'une ressource lexico-conceptuelle. Dans notre tâche de conception de l'analyseur, nous ne faisons donc aucune hypothèse sur les processus cognitifs à l'œuvre lors de la compréhension humaine. Par contre, nous intégrons dans la conception de l'outil les spécifications liées à la tâche d'interprétation au cours de laquelle les résultats qu'il fournit seront exploités par un analyste humain. Cette rétroaction justifie pour une part nos choix d'implémentation.

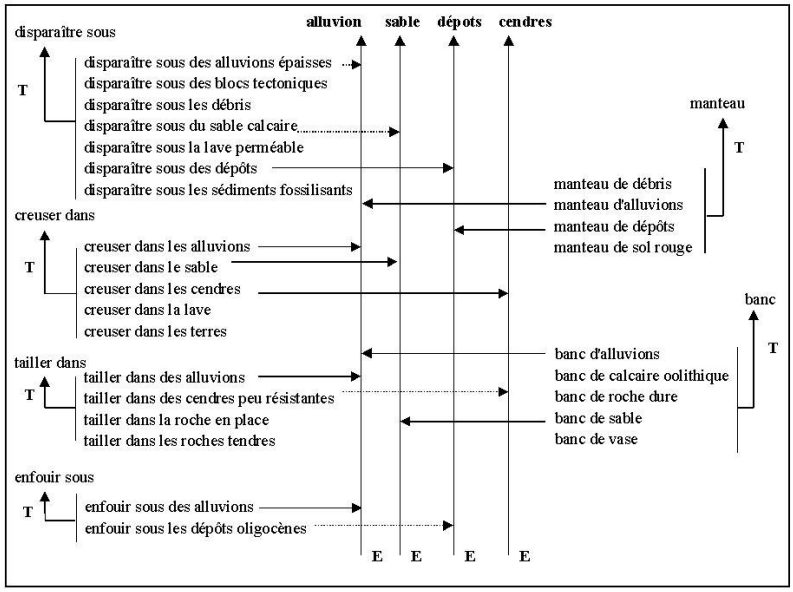


**Figure 1** - Extrait d'un réseau de dépendance construit par l'analyseur *SYNTEX*[1]. Les liens T se lisent "a pour tête", les liens E "a pour expansion". Les liens en traits pleins sont des liens directs ; par exemple, *disparaître sous des dépôts* a pour expansion *dépôts*. Les liens en pointillés sont des liens indirects ; par exemple, *disparaître sous des alluvions épaisses* a pour expansion un groupe qui a pour tête *alluvions*.

### 3.2. Un analyseur de corpus

Par rapport aux analyseurs développés classiquement dans le domaine du traitement automatique des langues (Abeillé 1992), l'analyseur *SYNTEX* présente cette différence fondamentale d'être un analyseur de corpus, et non pas de phrases. C'est un analyseur de corpus parce que (i) le résultat de l'analyse est un réseau de dépendance global construit pour l'ensemble du corpus, (ii) le réseau de dépendance construit constitue un mode d'accès au corpus pour l'analyste en charge de l'interprétation, (iii) le corpus est source d'information pour l'analyseur.

(i) Les représentations construites par l'analyseur ne sont pas de façon cruciale des représentations de phrases, mais un réseau global de dépendance entre mots ou syntagmes construit pour l'ensemble du corpus (figure 1). Dans



[1] Les exemples présentés dans cet article sont tirés d'un corpus de géomorphologie. Nous remercions vivement Danièle Candel, de l'Institut National de la Langue Française, de nous avoir aimablement fourni ce corpus, extrait de la base SCITECH.

la conception de l'analyseur, la phrase joue un rôle en tant que domaine pour la recherche des syntagmes et des relations de dépendance : l'analyseur cherche des relations de dépendance syntaxique à l'intérieur des frontières de la phrase. Mais l'objectif final de l'analyse n'est pas de construire une représentation de la structure globale de la phrase. L'objectif est d'identifier des relations de dépendance locales, d'extraire des syntagmes et de synthétiser ces analyses locales en construisant un réseau global de dépendance entre mots.

(ii) Dans une telle approche, le rôle du corpus est fondamental. La construction d'un réseau global de dépendance n'a de sens que si sont posées des hypothèses de clôture et d'homogénéité thématiques du corpus. Dans notre conception de l'activité de construction de ressources lexico-conceptuelles à partir de corpus, la tâche de construction du corpus est problématisée. Le corpus d'analyse n'est jamais déjà donné. Il doit être constitué en fonction du type de ressources à construire. Il faut à la fois viser une "bonne" couverture, et donc parfois une certaine diversité, et une "bonne" homogénéité. Par ailleurs, le corpus est le point de départ de l'analyse, en tant qu'objet "traité" par l'analyseur. Il en est aussi le point d'aboutissement, en tant que lieu de passage obligé des parcours interprétatifs de l'analyste. C'est pourquoi, dans la visée d'aide à l'interprétation qui est la nôtre, la conception de l'analyseur doit s'accompagner de celle de l'interface qui permettra à l'analyste de manipuler et d'interpréter les résultats de l'analyseur. Cette interface doit en particulier ménager un accès aux textes via les occurrences des syntagmes extraits.

(iii) Le rôle du corpus est enfin fondamental dans notre approche parce que le corpus est non seulement l'objet du traitement pour l'analyseur, et le lieu d'interprétation pour l'analyste, mais c'est aussi une source d'information pour l'acquisition par l'analyseur d'informations de complémentation, comme nous allons le montrer dans la section suivante.



### 3.3. Apprentissage endogène sur corpus

Comme tout analyseur syntaxique, l'analyseur *SYNTEX* est mis en difficulté dans les situations d'ambiguïté de rattachement des groupes prépositionnels et des adjectifs ou groupes adjectivaux. Le tableau 1 illustre le problème du rattachement de l'adjectif dans la structure 'Nom$_1$ de Nom$_2$ Adjectif'. Quand les informations de genre et de nombre sont non discriminantes, aucun signe local ne permet de décider duquel des deux noms dépend l'adjectif. De même, le tableau 2 illustre le problème du rattachement du groupe prépositionnel '*en* Nom' dans la structure 'Verbe Déterminant Nom Adjectif *en* Nom'. Le syntagme prépositionnel '*en* Nom' peut dépendre de l'adjectif, du nom ou du verbe qui précèdent. Un analyseur déterministe strictement syntaxique, qui n'exploite d'autre information que la simple succession des étiquettes morpho-syntaxiques, est incapable de distinguer le détail de ces groupes, puisque les séquences de catégories sont identiques (Marandin

1993, p. 28). Or nous voulons que l'analyseur prenne une décision (et une seule) dans ces cas ambigus, pour identifier la structure syntaxique du syntagme et enrichir ainsi le réseau de dépendance.

| **Contexte d'extraction** | |
|---|---|
| (1) La discontinuité est marquée par les *réfractions d'ondes séismiques*. | |
| (2) Il se crée une *vague d'érosion remontante* qui creuse une gorge. | |
| (3) Se constitue ainsi une *plaine de bordure karstique*. | |
| **Analyses concurrentes** | |
| (1) | a- [ réfractions d' [ ondes séismiques ] ] |
| | b- [ [ réfractions d'ondes ] séismiques ] |
| (2) | a- [ vague d' [ érosion remontante ] ] |
| | b- [ [ vague d'érosion ] remontante ] |
| (3) | a- [ plaine de [ bordure karstique ] ] |
| | b- [ [ plaine de bordure ] karstique ] |
| **Indices** | |
| (1') Les *ondes séismiques* s'amortissent dans la partie périphérique | |
| (2') La *vague d'érosion* remontera sur le cours principal. | |
| (3') Le poljé est une *plaine karstique*. | |
| **Choix** | |
| (1) : a- [ réfractions d' [ ondes séismiques] ] | |
| (2) : b- [ [ vague d'érosion ] remontante ] | |
| (3) : b- [ [ plaine de bordure ] karstique] | |

**Tableau 1** - Apprentissage endogène :
cas de la structure '$Nom_1$ Préposition $Nom_2$ Adjectif'



L'analyseur est censé être utilisé dans un contexte de construction de ressources lexicales spécialisées, à partir d'un corpus portant sur un domaine *a priori* quelconque. Il est donc hors de question de lui faire exploiter des informations de type sémantique, puisque les informations sémantiques qui lui seraient utiles pour lever l'ambiguïté sont justement celles que l'analyseur doit contribuer à mettre au jour. Nous poussons cette hypothèse nihiliste plus loin encore en affirmant qu'il en va de même pour les propriétés syntaxiques de complémentation. Nous ne fournissons ainsi à notre analyseur aucune information syntaxique de complémentation, du type « tel verbe se construit avec telle préposition ». Ce parti pris est justifié par le constat maintes fois confirmé que les informations de complémentation des mots dans les corpus spécialisés sont souvent inédites et imprédictibles. On peut ajouter à cela qu'il n'existe pas actuellement pour le français de base lexicale complète et

facilement accessible qui recenserait les propriétés de complémentation des verbes, noms et adjectifs. Mais fondamentalement, l'hypothèse de la table rase, qui est également à la base d'autres travaux comparables, comme ceux de Basili et al. (1999) sur l'italien, n'est pas une position de repli face à la pénurie, c'est une position imposée par l'objectif même de construction de ressources sémantiques à partir de corpus spécialisés.

La stratégie alternative adoptée est celle de l'apprentissage endogène sur corpus (Bourigault 1994). L'analyseur est doté d'heuristiques d'apprentissage qui lui permettent d'acquérir par lui-même, grâce à l'analyse du corpus en cours de traitement, les informations pertinentes pour lever les ambiguïtés de rattachement syntaxique. Le principe est simple. Nous l'illustrons dans les tableaux 1 et 2. Pour reprendre l'exemple (1) (resp. 2, 3) du tableau 1, c'est parce qu'il aura relevé une occurrence non ambiguë du syntagme *onde séismique* (resp. *vague d'érosion*, *plaine karstique*) dans la séquence (1') (resp. 2', 3') que l'analyseur, en l'absence d'autres indices, prend la décision de choisir l'analyse (a) (resp. b, b) qui isole la même relation de dépendance entre *onde* et *séismique* (resp. entre *vague* et *érosion*, entre *plaine* et *karstique*). De même, pour reprendre l'exemple (4) (resp. 5, 6) du tableau 2, c'est parce qu'il aura relevé une *série* d'occurrences non ambiguës du verbe *disséquer* (resp. du nom *charge*, de l'adjectif *pauvre*) suivi de la préposition *en* que l'analyseur acquiert l'information que, sur ce corpus, cette unité lexicale est susceptible de régir la préposition *en*. Il utilise cette information pour prendre la décision, en l'absence d'autres indices, de choisir l'analyse (c) (resp. b, a) qui isole une relation de dépendance entre *disséquer* (resp. *charge*, *pauvre*) et le syntagme prépositionnel en *en*. Ce principe de l'apprentissage endogène, repris de LEXTER, est à la base de la conception de l'analyseur SYNTEX, dans lequel il a été étendu et systématisé.

On le trouve déjà en germe dans les travaux de Fathi Debili (Debili 1982), ainsi que ceux sur l'analyseur *ALSF* de Sophie David (David & Plante 1990), voir (Bourigault 1994, pp. 63-78).



Ainsi équipé de procédures d'apprentissage endogène sur corpus et de stratégies de résolution des ambiguïtés qui exploitent les informations acquises, l'analyseur est capable à la fois d'identifier des groupes syntaxiques et d'analyser le détail de leur constitution. D'une certain façon, l'analyse en largeur rend possible l'analyse en profondeur : pour restituer l'analyse syntaxique profonde d'un syntagme extrait, l'analyseur procède à un examen large de l'intégralité du corpus à la recherche de preuves (configurations attestées non ambiguës) lui permettant de choisir parmi un ensemble d'analyses concurrentes.

Nous allons voir à présent, à travers la description plus détaillée de la démarche de désambiguïsation mise en œuvre dans SYNTEX, comment ce principe général d'une démarche endogène basée sur le repérage d'informations en corpus se réalise à travers le repérage d'indices linguistiques originaux.

| **Contexte d'extraction** | |
|---|---|
| (4) L'érosion *a disséqué le plateau rocheux en chevrons*. | |
| (5) On *observe une charge excessive en trouble* dans les rivières à méandres | |
| (6) Il faut *distinguer les roches pauvres en magnésium*. | |
| **Analyses concurrentes** | |
| (4) | a- [ disséquer [ le plateau [ rocheux en chevron ] ] ] |
| | b- [ disséquer [ [ le plateau rocheux ] en chevron ] ] |
| | c- [ [ disséquer [ le plateau rocheux ] ] en chevron ] |
| (5) | a- [ observer [ une charge [ excessive en trouble ] ] ] |
| | b- [ observer [ [ une charge excessive ] en trouble ] ] |
| | c- [ [ observer [ une charge excessive ] ] en trouble ] |
| (6) | a- [ distinguer [ les roches [pauvres en magnésium] ] ] |
| | b- [ distinguer [ les roches pauvres ] en magnésium] ] |
| | c- [ [ distinguer [ les roches pauvres] ] en magnésium] |
| **Indices** | |
| (4') des plateaux *disséqués en* bois de renne (…) l'anticlinal, dont la carapace a été *disséquée en* chevrons (…) des crêtes *disséquées en* dents de scie | |
| (5') *charge en* matériaux dissous exceptée (…) chaque couche étant caractérisée par sa *charge en* poussières (…) La *charge en* troubles peut être considérable | |
| (6') il est acide, *pauvre en* biotite, faiblement diaclasé (…) Les roches claires, acides, *pauvres en* fer (…) Les syénites, *pauvres en* plagioclases, sont assez résistantes | |
| **Choix** | |
| (4) : c- [ [ disséquer [ le plateau rocheux ] ] en chevrons ] | |
| (5) : b- [ observer [ [ une charge excessive ] en trouble ] ] | |
| (6) : a- [ distinguer [ les roches [pauvres en magnésium] ] ] | |



**Tableau 2** - Apprentissage endogène :
cas de la structure 'Verbe Déterminant Nom Adjectif *en* Nom'

## 4. Stratégie endogène pour l'acquisition de propriétés de complémentation

L'étape cruciale de désambiguïsation du rattachement prépositionnel consiste à identifier les liens de dépendance qui s'établissent entre des unités rectrices, dotées de propriétés de complémentation (verbes, mais aussi noms et adjectifs), et des unités régies. Nous illustrons dans cette partie la façon dont notre analyseur, basé sur l'observation des cooccurrences entre mots du corpus, met au jour ces relations de dépendance et intègre des stratégies propres à différencier, au sein de ces relations, différents types de dépendance. Nous décrivons au préalable dans ses grandes lignes la stratégie

mise en œuvre par l'analyseur pour résoudre les ambiguïtés du rattachement prépositionnel.

### 4.1. Une notion-clé : la productivité

Comme nous l'avons expliqué dans la section 3.3, le principe de l'apprentissage endogène des propriétés de complémentation s'effectue par le biais du repérage dans le texte de zones de rattachement non ambiguës. Dans ces zones, nous identifions les triplets (*recteur*, *prép*, *régi*), associés par une relation de dépendance syntaxique. Cette relation peut s'instaurer dans le cadre de schémas de sous-catégorisation (comme c'est le cas pour les triplets (*pénétrer, dans, pore*) ou (*aptitude*, *à*, *décrire*)), ou illustrer des liens de dépendance de nature non argumentale (circonstants (*déplacer*, *à, vitesse*), expansions nominales (*côte*, *à*, *fjord*), etc.).

**• Acquisition en contexte non ambigu**

Le module de découpage a pour rôle de circonscrire le domaine de rection d'une préposition, c'est-à-dire la zone textuelle au sein de laquelle la préposition doit trouver son recteur. Il est basé sur une procédure simple, qui consiste à remonter à gauche de la préposition jusqu'à rencontrer un élément frontière[2] (ponctuation, forme verbale, coordonnant, préposition autre que *de*). Le module de découpage livre ensuite à la phase d'acquisition la liste des candidats-recteurs (verbes, noms et adjectifs) rencontrés dans la zone de rattachement de la préposition. Une fois découpée la zone de rection, deux cas de figure se présentent. Dans le premier cas, la zone ne comporte qu'un candidat recteur. Le rattachement est donc considéré comme non ambigu, et le triplet (recteur, préposition, régi) est acquis par le programme. Dans le deuxième cas, plusieurs candidats figurent dans la zone, aucun indice fiable ne peut donc être extrait de cette situation d'ambiguïté. Les cas de rattachement non ambigu fournissent donc l'amorce essentielle du programme.



**• Calcul de la productivité**

Tous les liens acquis en situation non ambiguë ne sont pas exploitables. Un critère intervient pour évaluer leur fiabilité : il s'agit de la productivité de l'association entre un recteur et une préposition. La productivité est déterminée par le nombre de régis différents avec lesquels la paire (recteur, prép) se combine dans le corpus. Un mot est productif vis-à-vis d'une préposition donnée, si la paire se combine avec au moins deux régis

[2] Les éléments frontières que nous avons définis sont en réalité susceptibles d'être franchis, dans bien des cas, par le lien prépositionnel. Cette procédure de base est donc complétée par des règles qui permettent de récupérer, dans certaines configurations (coordination, saut de préposition), des recteurs situés en dehors de la zone.

différents dans le corpus. Par exemple, au vu des contextes non ambigus suivants, on constate que la productivité du mot *disséquer* comme recteur de la préposition *en* a pour valeur 5.

*disséqué parfois en* ***récif***
*disséquer en* ***récif***
*disséqué en* ***dents*** *de scie*
*disséquer en* ***terrasses***
*disséqués en* ***bois*** *de renne*
*disséquée en* ***chevron***
==>*disséquer, en + (récif, dent, terrasse, bois, chevron)*

Cette valeur permet, de façon plus fiable que le simple calcul de la fréquence du couple (recteur, prép), de mesurer la force d'association entre les deux éléments, dans la mesure où une productivité élevée est l'indice que le recteur fonctionne de manière régulière avec cette préposition, donnant lieu à des occurrences diversifiées.

**• Résolution**

Le module de résolution exploite une conjonction d'indices permettant d'évaluer la fiabilité d'un lien recteur/préposition potentiel. Parmi ceux-ci figure l'indice de productivité, que nous venons de définir. L'exemple 4 présenté dans le tableau 2 donne ainsi lieu à la configuration suivante : des trois candidats en lice, seul le verbe *disséquer* possède une indice de productivité non nul (sa productivité avec la préposition *en* est égale à 3).

Cette phase de résolution est en cours de mise au point. Le taux de précision obtenu à l'heure actuelle, à partir d'une dizaine d'indices, est de 86%, le taux de rappel de 60%. Ces résultats sont obtenus en comparant les rattachements effectués de manière automatique aux résultats d'une tâche de rattachement manuel effectuée sur plusieurs milliers d'occurrences. La précision est très satisfaisante. Le taux de rappel doit être amélioré en perfectionnant en particulier le module de segmentation, de manière à rechercher dans certains cas des recteurs potentiels au-delà des éléments frontières que nous avons définis.



### 4.2. La productivité : un outil de repérage des positions syntaxiques ?

Le rattachement prépositionnel, envisagé la plupart du temps comme une procédure préliminaire à la délimitation de syntagmes et à l'analyse syntaxique, permet dans notre cas de déterminer des relations de dépendance exploitables en tant que ressources lexicales pour l'analyse sémantique distributionnelle. Cette expérience pose en chemin une question d'ordre plus théorique : quels types d'information linguistique peut-on tirer de l'observation des associations lexicales, quelle est la nature des relations de dépendance extraites par ce moyen ? En particulier, la force de l'association

dans un texte entre un recteur et un groupe prépositionnel, que nous mesurons en termes de productivité du couple (recteur, prép), dit-elle quelque chose du statut syntaxique de l'expansion prépositionnelle - argument ou circonstant ? Brent (1993) défend l'hypothèse que deux mots seront plus fréquemment associés s'ils sont unis par une relation argumentale ; des heuristiques exprimées en termes de fréquence doivent selon lui permettre d'approcher cette distinction syntaxique fondamentale. Selon Basili et al. (1999) au contraire, cette distinction est impossible à faire de manière automatique, et elle n'est pas souhaitable dans la mesure où les circonstants contribuent également à la sémantique du verbe et manifestent une régularité au même titre que les arguments. Les observations que nous menons sur les corpus montrent en effet à quel point la frontière entre les deux types de compléments est difficile à tracer, même dans les situations de validation manuelle. Nous avons cependant cherché à déterminer s'il était possible d'aller au-delà du simple diagnostic de rattachement et de s'appuyer sur l'indice de productivité pour tenter de repérer des groupes prépositionnels de statut différent. Nous livrons dans ce qui suit nos premières conclusions sur ce point.

#### 4.2.1. Régis productifs *vs* recteurs productifs

Nous avons cherché à comparer deux types de productivité : celle concernant les régis et celle concernant les recteurs. Nous nous intéressons aux noms et nous comparons le statut des informations acquises selon deux critères différents : le recteur est productif vis-à-vis d'une préposition donnée, il s'associe donc par ce biais avec différents noms dans des groupes nominaux dont il est la tête ; ou bien : le régi est productif vis-à-vis de cette préposition, il s'associe donc à différents noms tête. Notre but est de proposer des critères pour établir des propriétés de rection différenciées, en partant de l'étude des noms simples, c'est-à-dire sans lien morphologique avec un verbe. Est-on en mesure de faire un diagnostic différent selon que la force d'association lexicale s'établit entre le recteur et la préposition, ou entre la préposition et le régi ? Notre hypothèse en suivant cette piste était d'observer s'il existait une corrélation entre le type de liens complétifs existant dans un *N prép N* - tels que les définissent par exemple Milner (1989) dans le cas des *N de N*, ou Cadiot (1997) dans le cas des *N à N* - et la force d'association de la préposition avec l'un ou l'autre nom. Nous illustrons cette expérience à partir du cas de la préposition *à*.



#### 4.2.2. L'exemple des structures N à N

L'analyseur extrait de notre corpus 50 noms recteurs et 54 noms régis productifs avec la préposition *à*. Les tableaux 3 et 4 présentent les triplets *N à* (*dét*) *N* présentant une productivité supérieure ou égale à 3. La première ligne du tableau 3 indique par exemple que le mot *carte* a été trouvé dans un

contexte d'acquisition comme seul recteur candidat de la préposition *à*, sans déterminant. Les trois contextes sont *: carte à 1/10 000*, *carte à 1/200 000*, *carte à 1/80 000*. La première ligne du tableau 4 indique que le mot *aval* a été trouvé dans un contexte d'acquisition en situation de régi de la préposition *à*, associé à quatre recteurs différents, avec un déterminant : *ablation à l'aval*, *amont à l'aval*, *plage à l'aval*, *terrasse à l'aval*.

| **nom recteur productif** | **lien** | **régis** |
|---|---|---|
| *carte* | *à* | *1/10 000, 1/200 000, 1/80 000* |
| *cas* | *à* | *Java; Montserrat; Nantasket* |
| *cas* | *à + dét* | *lahar; pays; pays-bas* |
| *craie* | *à* | *bélemnite; micraster; silex* |
| *crête* | *à* | *Porolithon; cheminée; clocheton* |
| *côte* | *à* | *falaise; fjord; plage; ria; skjär; structure* |
| *kilomètre* | *à + dét* | *Nord; dizaine; pôle* |
| *méthode* | *à + dét* | *potassium; strontium; uranium* |
| *roche* | *à* | *diaclase; feldspath; feldspathoïde; grain* |
| *roche* | *à + dét* | *extérieur; minéral; soleil* |
| *région* | *à* | *cuesta; nappe; permafrost; plateaux; saison; sous-sol* |
| *zone* | *à* | *cristal; pergélisol; pluie* |

**Tableau 3** : Recteurs productifs vis-à-vis de la préposition *à* dans le corpus de géomorphologie

| **recteurs** | **lien** | **nom régi productif** |
|---|---|---|
| *ablation; plage; terrasse* | *à + dét* | *aval* |
| *degré; maximum; éprouvette* | *à + dét* | *dessous* |
| *Ouest; actif; haut; rigole* | *à* | *droite* |
| *calcaire; ensemble; granit; granite; grès; leucogranit; roche* | *à* | *grain* |
| *gneiss; granit; micaschiste* | *à* | *mica* |
| *altitude; base; forme* | *à* | *peine* |
| *dépression; glacis; grès; zone* | *à + dét* | *pied* |
| *courant; glacis; meuble; plan* | *à + dét* | *sens* |
| *affaire; ciselure; face* | *à + dét* | *surface* |
| *concavité; côte; côté; pente* | *à + dét* | *vent* |

**Tableau 4** : Régis productifs vis-à-vis de la préposition *à* dans le corpus de géomorphologie

Faisons tout d'abord quelques remarques communes aux deux tableaux : certains triplets présentés sont erronés en raison de plusieurs imperfections du traitement actuel. Certaines erreurs sont dues aux problèmes de découpage. Ainsi, la séquence *Ouest à droite* a été identifiée à tort comme contexte non ambigu. D'autres erreurs sont le fait de mots mal catégorisés par l'étiqueteur dont notre analyseur exploite les résultats (*meuble*, *actif*, sont en fait des adjectifs), ou de noms faisant partie de locutions verbales (*être le cas*) et pouvant difficilement de ce fait être considérés comme des recteurs autonomes. De façon générale, les résultats obtenus sont moins bons que dans le cas des verbes ou des noms processifs, auxquels sont associés des schémas de complémentation plus réguliers.

Signalons également, dans la perspective de l'exploitation de ces informations pour la mise en évidence de classes sémantiques, l'importance de la distinction entre des structures avec ou sans déterminant (conformément aux descriptions de Cadiot sur ce point). Cela est en particulier illustré dans le tableau 3 par l'opposition entre *roche à l'extérieur*, *au soleil* et *roche à diaclase*, *à feldspath*, etc. D'autres exemples, pris dans les cas de productivité égale à 2, illustrent ce même contraste, ainsi pour le mot *zone* s'opposent nettement : *zone à cristal*, *à pergélisol* vs *zone à l'ombre*, *au pied*.

Si l'on considère à présent l'ensemble des triplets de manière à comparer les résultats obtenus selon les deux mesures de productivité, on s'aperçoit que les tableaux 3 et 4 manifestent des relations de dépendance de nature sensiblement différente. Les régis productifs apparaissent dans trois types de structures :

- des locutions prépositives (*à peine*, *au sens*),
- des groupes prépositionnels à valeur circonstancielle (*à l'aval*, *au dessous*, *à droite*, *au pied*, *à la surface*),
- et minoritairement – 2 cas sur 10 – des composés *N à N* (*à mica*, *à grain*).



Les recteurs productifs constituent quant à eux dans 9 cas sur 14 la tête de composés *N à N* (*craie à Bélemnite*, *méthode au potassium*), dans lesquels l'expansion prépositionnelle a valeur de qualification (Cadiot 1997).

Ces résultats confirment l'idée que la forte productivité du recteur vis-à-vis de la préposition peut être utilisée comme indice d'un groupe *N à (dét) N* à forte cohésion. Ils montrent par ailleurs qu'un fort degré d'attachement de la paire (préposition, régi) peut être également exploité en orientant l'analyse vers des groupes au statut particulier, présentant une relative autonomie, et dont la productivité signale une information caractéristique du texte (ici, il s'agit de groupes prépositionnels indiquant une localisation). D'autres résultats concernant les verbes confirment ces premières constatations (Fabre et Bourigault 2001).

## 5. Syntaxe posito-argumentale et conception de l'analyseur

### 5.1. Préambule

Pour terminer cet article, nous tentons d'établir un lien entre recherches théoriques en linguistique et travaux de conception d'outils de traitement automatique des langues. Nous croyons qu'une confrontation entre approches des deux types, menée dans le cadre du programme de recherche de l'aide à l'interprétation de corpus, doivent conduire à une fécondation réciproque. Pourquoi tenter un rapprochement avec une théorie syntaxique ? Les premiers développements de l'analyseur ont été effectués indépendamment de toute théorie syntaxique. La méthode utilisée a été (et reste) une méthode expérimentale basée sur corpus. Les règles d'extraction et d'analyse dépendancielle sont élaborées et testées par l'analyse d'un grand nombre de cas relevés sur des corpus divers. Cependant, la confrontation avec ces données conduit à croiser des problèmes syntaxiques qui ont été traités dans nombre de travaux de linguistique. Par ailleurs, le souci de réaliser des analyses plus fines, pour améliorer la qualité des résultats fournis par l'analyseur, nous conduit naturellement à rechercher un appui du côté des théories de la syntaxe.

Parmi celles-ci, nous avons fait le choix d'analyser la théorie de J.-C. Milner, telle qu'elle est décrite dans son ouvrage « Introduction à une science du langage » (Milner 1989). Ce choix, d'abord subjectif, se justifie par la grande richesse de cette théorie, qui s'appuie sur la tradition de la grammaire générative de Chomsky, tout en présentant une originalité et une rigueur très séduisantes. Nous avons volontairement écarté d'autres théories plus formelles inspirées ou non de la théorie chomskyenne, comme LFG ou HPSG, pour maintenir une distance forte *a priori* entre nos travaux pour l'implémentation d'un analyseur syntaxique et des travaux visant la construction d'une théorie syntaxique.

Bien sûr, la différence est radicale entre l'ambitieux projet de J.-C. Milner, construire une science du langage et développer une théorie syntaxique, et notre objectif plus modeste de fournir des outils méthodologiques et logiciels pour l'analyse sémantique de corpus. Il ne s'agit donc en aucun cas de "comparer" nos approches et hypothèses. Notre analyse de la théorie est guidée par le souci de faire entrer en résonance les problèmes et les choix méthodologiques de la conception de l'analyseur avec certains éléments de la théorie de Milner, notations, hypothèses, principes, analyses de phénomènes. C'est un exercice à la fois stimulant et difficile, dont nous synthétisons les premiers résultats dans la section 5.3. Dans la section 5.2, nous décrivons succinctement la théorie de J.-C. Milner, que nous nommerons désormais "syntaxe posito-argumentale". Celle-ci se présente en trois parties : la théorie restreinte des termes, la théorie des positions et la théorie étendue des termes.

### 5.2. La syntaxe posito-argumentale

#### 5.2.1. La théorie restreinte des termes

La première partie de la syntaxe posito-argumentale est la théorie restreinte des termes. Celle-ci concerne le lexique proprement dit. Elle s'appuie sur l'hypothèse qu'il est possible d'établir les propriétés d'un terme hors emploi[3]. Elle s'intéresse aux propriétés absolues des termes. Ces propriétés sont des propriétés distinctives, et les facteurs d'individuation d'un terme sont :

- son appartenance catégorielle : Milner fait appel aux catégories de la grammaire traditionnelle ;
- sa forme phonologique ;
- sa signification lexicale : la signification lexicale d'un terme est sa référence virtuelle, c'est-à-dire « un ensemble de conditions que doit satisfaire un objet du monde pour pouvoir être désigné, en référence actuelle, par une molécule syntaxique dont (le terme) sera le Nom principal » (Milner op. cit., p. 336).

#### 5.2.2. La théorie des positions

La théorie des positions est la partie strictement syntaxique de la syntaxe posito-argumentale. Cette théorie se fonde sur la distinction entre place et position. Par exemple, en (7a) et en (7b) le syntagme prépositionnel *à mon fils* occupe des places analogues, mais des positions différentes : complément de *faire* en (7a), complément de *donner* en (7b) (Milner op. cit., p. 298).

(7a) *j'ai fait prendre le train à mon fils.*

(7b) *j'ai fait donner une couchette à mon fils.*



Les positions sont l'objet de la syntaxe. Elles sont l'équivalent des fonctions grammaticales de la grammaire. Dans un domaine syntaxique donné, un terme donné occupe une position et une position doit être occupée par un terme. Le jugement linguistique qui joue le rôle d'observation empirique prend donc en syntaxe la forme minimale suivante : *« le terme lexical X qui, pour l'observation immédiate, occupe la place Y, occupe une position syntaxique Z »*. Les places sont perceptibles, elles se déduisent immédiatement de l'ordre linéaire des mots dans la phrase. Les positions ne sont pas perceptibles, elles sont à identifier par l'analyse dans le cadre d'une théorie syntaxique positionnelle. La géométrie des places et la géométrie des positions sont de natures différentes. Dans la théorie, cette distorsion entre place et position est prise en charge par l'application du principe de naturalité, qui est un principe général auquel J.-C. Milner fait régulièrement

[3] Les « termes » sont à considérer ici en tant qu'unités de description identifiées par la théorie du lexique, et non selon l'acception courante en terminologie.

appel dans la théorie. Dans le contexte présent, ce principe requiert que l'on fasse l'hypothèse qu'entre positions et places il y ait autant de coïncidence que possible :

[PN1] *« On prédit donc qu'il y a plus de chances qu'un élément auquel la géométrie syntaxique attribue telle position abstraite occupe effectivement une place correspondante dans le système des places observées. Réciproquement, on prédit qu'une différence de place peut signaler une différence de position. »* (ibid., p. 395).

Les indices qui permettent de restituer les positions à partir de l'observation des places sont :

- la récurrence. Si un élément donné occupe régulièrement une place donnée, on conclura que, sauf forte raison de penser le contraire, cette place correspond à sa position.
- la proximité. Deux positions linguistiquement reliées doivent être géométriquement proches.

Les positions ont des propriétés par elles-mêmes, indépendamment des termes qui les occupent. La propriété absolue d'une position est son étiquette catégorielle attachée. Il faut alors distinguer le nom catégoriel de la position (son étiquette attachée) et l'appartenance catégorielle du terme ou de la molécule lexicale qui occupe cette position. Par exemple, Milner propose pour la phrase (8) l'analyse (9), qu'il faut lire ainsi : la position *sujet* étiquetée N" est occupée par un terme dont l'étiquette catégorielle est S (phrase) (ibid., p. 364).

(8) *mourir n'est rien*

(9) *[$_{N"}$ ($_{S}$ e mourir) ] n'est rien*



Mais là aussi le principe de naturalité s'applique, de la façon suivante :

[PN2] *« Bien que l'appartenance catégorielle d'un terme X et l'étiquette catégorielle de la position Y occupée par X soient indépendantes en droit, il est naturel et normal [qu'elles soient] homonymes. »* (ibid., p. 370).

### 5.2.3. La théorie étendue des termes

La théorie étendue des termes s'intéresse aux propriétés relationnelles des termes. Par exemple, parmi les propriétés sémantiques d'un lexème verbal figurent les types de ses compléments. Cette partie de la théorie se trouve à la jonction du lexique et de la syntaxe. La frontière est marquée par la distinction entre arguments et positions appelées. Un lexème verbal a des arguments (une structure argumentale). En tant qu'opérateur, il impose à ses arguments des propriétés sémantiques (des conditions interprétatives

spécifiques), qui constituent des rôles. Réciproquement, chaque argument d'un lexème verbal reçoit de lui des propriétés sémantiques en tant qu'argument. Par ailleurs, un lexème verbal appelle des positions ; par exemple il peut être transitif, c'est-à-dire appeler une position de groupe nominal. Ici, le principe de naturalité stipule que les arguments d'un lexème verbal se réalisent naturellement dans les positions appelées par ce lexème. Ce qui est dit des verbes peut l'être d'autres catégories (noms, adjectifs). Dans ce contexte, le principe de naturalité s'exprime ainsi :

[PN3] *« Sauf circonstances particulières* […]*, l'argument N'' du verbe est aussi complément proche et non mobile dans le V'' dont le verbe est le noyau. »* (ibid., p. 436).

L'apparition d'une telle théorie des termes rend plus lointain le projet du programme génératif initial : *« Le programme génératif soutenait que des règles formelles* [...] *étaient non seulement nécessaires, mais aussi suffisantes. Or, l'étude empirique semble avoir montré qu'aucun ensemble de règles formelles, définissable a priori, qu'il s'agisse des règles de réécriture seules ou des règles de transformation, n'est suffisant : interviennent également certaines propriétés des termes et notamment leurs propriétés relationnelles. La théorie a du même coup dû accorder un rôle décisif à une information non déductible a priori. En effet, les propriétés des termes, qu'elles soient relationnelles ou absolues, partagent le même caractère : elles ne peuvent qu'être enregistrées une par une, de manière encyclopédique, par une mémoire individuelle. »* (ibid., pp. 456-457).

### 5.3. Syntaxe posito-argumentale et conception de l'analyseur

En écho à la citation ci-dessus, rappelons que l'activité pour laquelle nous concevons l'analyseur est celle de construction, et d'enregistrement, des propriétés des unités lexicales propres à un univers sémantique particulier, à partir de l'analyse d'un corpus. En ce sens, l'hypothèse de base de la théorie *restreinte* des termes, à savoir qu'il est possible d'établir les propriétés d'un terme hors emploi, est en inadéquation avec notre programme de recherche. Ce sont les propriétés relationnelles des termes qui nous intéressent. C'est donc du côté de la théorie *étendue* des termes que nous nous tournons. On y retrouve reformulée dans un programme différent l'hypothèse de l'analyse distributionnelle. Les arguments qui assurent le même rôle vis-à-vis d'un opérateur ont des propriétés sémantiques relationnelles identiques.

On peut alors reformuler les spécifications de l'analyseur dans les termes de la théorie de la syntaxe posito-argumentale. L'analyseur doit aider l'analyste à identifier les structures argumentales des unités lexicales (verbes, noms, adjectifs) du corpus. Selon le principe de naturalité ([PN3], section 5.2.3), les arguments se réalisent dans les positions de complément appelées par l'unité lexicale. L'analyseur devrait donc trouver ces positions de

complément, ou, plus exactement, les unités lexicales ou syntagmes qui occupent ces positions. Par principe, les positions ne sont pas perceptibles par l'analyseur, seules les places peuvent l'être. L'analyseur ne peut donc chercher à repérer que des unités lexicales ou syntagmes occupant certaines *places*. Selon le principe de naturalité ([PN1], section 5.2.2), et si on impose à l'analyseur des contraintes de *proximité* et de *récurrence* dans sa recherche (section 5.2.2), il est susceptible de repérer des places qui correspondent en toute probabilité à des positions syntaxiques. Ces contraintes sont à la base du principe de productivité, dont nous avons montré dans la section 4, comment il est un outil pour le repérage de propriétés de complémentation des unités lexicales et pour la désambiguïsation des rattachements prépositionnels.

L'un des obstacles de fond à un rapprochement plus étroit entre la théorie syntaxique de J.-C. Milner et notre approche pour concevoir un analyseur syntaxique vient de ce que les unités traitées dans les deux cas sont de types radicalement différents. Les entités lexicales en jeu dans la syntaxe posito-argumentale sont les "termes" ou "molécules lexicales", dont l'information sémantique est portée par la "tête lexicale". L'analyseur, lui, traite en entrée un corpus ayant subi une analyse morpho-syntaxique : à chaque mot du corpus sont associés une (et une seule) catégorie syntaxique et un lemme. Les unités perçues par l'analyseur sont donc des mots, tels que les a découpés l'analyseur morpho-syntaxique, dont les facteurs d'individuation sont : la catégorie morpho-syntaxique, le lemme et la forme graphique. L'analyseur ne perçoit que des séquences de telles unités. Les molécules lexicales (syntagmes) ne sont bien entendu pas connues *a priori* par l'analyseur. Au contraire, la fonction de celui-ci est de les identifier. L'analyseur n'a d'autre possibilité que de travailler sur les têtes lexicales des syntagmes à découvrir. C'est une fois qu'il aura identifié l'ensemble des relations de dépendance syntaxique entre mots que l'analyseur pourra déterminer les syntagmes. Pour reprendre l'exemple (4) du tableau 2, c'est après avoir identifié la relation de dépendance entre la préposition *en* et le verbe *disséquer* que le système est en mesure d'identifier la structure syntagmatique (4c).



## 6. Conclusion

Cette expérience de mise au point d'un analyseur dédié à l'extraction de syntagmes illustre l'objectif général que nous poursuivons : créer un outil destiné à s'intégrer dans une chaîne de traitement des textes pour faciliter la mise au jour de leurs propriétés sémantiques, basé sur des procédures aptes à faire émerger les comportements syntaxiques des mots. Plus spécifiquement, les résultats que nous avons exposés démontrent que le critère de l'association lexicale, exprimée en termes de productivité, fournit un outil puissant pour déterminer les relations de dépendance entre mots. Cette notion, qui trouve des échos dans d'autres domaines de l'analyse linguistique

(cf. en particulier Bayyen et Renouf (1996) en morphologie) s'avère prometteuse à plusieurs titres, puisqu'elle constitue un critère central pour guider la désambiguïsation syntaxique, et fournit également des indices pour aller plus loin dans l'évaluation linguistique des unités ainsi identifiées. Un de nos objectifs consiste donc à systématiser les premières observations que nous avons effectuées concernant la différenciation de types de rattachement prépositionnel, et à intégrer ces résultats dans la phase de résolution.

Dans une perspective plus épistémologique, nous avons affirmé ici la corrélation possible, même distante, entre théorie linguistique et implémentation, à plusieurs niveaux, et en particulier dans la ré-interprétation d'une théorie syntaxique pour l'implémentation d'un analyseur syntaxique. Le paradoxe positionnel est fondateur pour la théorie syntaxique, c'est la source de difficulté essentielle pour l'analyse syntaxique automatique.

**Références bibliographiques**

Abeillé, A. (éd) (1992), *Analyseurs syntaxiques du français*, *TAL*, 32.2 (n° spécial), Paris.

Abeillé, A. (1993), *Les nouvelles syntaxes : grammaires d'unification et analyse du français,* Paris, Armand Colin.

Assadi, H. & Bourigault, D. (1995), « Classification d'adjectifs extraits d'un corpus pour l'aide à la modélisation des connaissances », in *Actes des 3èmes Journées internationales d'analyse des données textuelles (JADT95)*, Rome.

Basili, R., Pazienza, M.T. & Vindigni, M. (1999), « Adaptive Parsing and Lexical Learning », in *Actes de VEXTAL'99*, Venise.

Bayyen, R.H. & Renouf, A. (1996), « Chronicling the times : productive lexical innovations in an english newspaper », in *Language* 72.1, pp. 69-96.



Bourigault, D. (1994), *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.

Brent, M. R. (1993), From Grammar to Lexicon : Unsupervised Learning of Lexical Syntax, in *Computational Linguistics* 19.2, pp. 243-262.

Cadiot, P. (1997), *Les prépositions abstraites en français*, Paris, Armand Colin/Masson.

Condamines, A. (1996), « Aide à l'acquisition des connaissances par l'étude de la terminologie », in N. Aussenac, P. Laublet & C. Reynaud (éds) *Acquisition des connaissances*, Toulouse, Cépaduès-Edition, pp. 247-266.

Condamines, A. & Rebeyrolle, J. (2001), « Searching for and identifying conceptual relationships via a corpus-based approach to Terminological Knowledge Base (CTKB) », in D. Bourigault, C. Jacquemin & M.-C.

L'Homme (eds), *Recent Advances in Computational Terminology*, Amsterdam/Philadelphia, John Benjamins Publishing Company, pp. 128-148.

David, S. & Plante, P. (1990), « De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes », in *ICO*, 2.3.

Debili, F. (1982), *Analyse syntactico-sémantique fondée sur une acquisition automatique de relations lexicales-sémantiques*, Thèse de doctorat d'état, Université de Paris XI, Centre d'Orsay.

Fabre, C. & Bourigault, D. (2001), « Linguistic clues for corpus-based acquisition of lexical dependencies », in *Actes de la conférence Computational Linguistics*, Lancaster.

Grefenstette, G. (1994), *Exploration in Automatic Thesaurus Discovery*, Londres, Kluwer Academic Publishers.

Habert, B. & Nazarenko, A. (1996), « La syntaxe comme marchepied de l'acquisition des connaissances : bilan critique d'une expérience », in *Actes des 7èmes Journées d'acquisition des connaissances (JAC96)*, Sète.

Habert, B., Nazarenko, A. & Salem, A. (1997), *Les linguistiques de corpus*, Paris, Armand Colin.

Harris, Z. (1968), *Mathematical Structures of Language*, New-York, John Wiley & Sons.

Hirshman, L., Grishman, R. & Sager, N. (1975), « Grammatically-based automatic word class formation », in *Information Processing and Management* 11, pp. 39-57.

Hindle, D. (1990), « Noun Classification from Predicate-Argument Structures », in *Proceedings of the Association for Computational Linguistics Conference* (ACL'90), pp. 268-275.

Meyer, I. (2001), « Extracting knowledge-rich contexts for terminography», in D. Bourigault, C. Jacquemin & M.-C. L'Homme (eds), *Recent Advances in Computational Terminology*, Amsterdam/Philadelphia, John Benjamins Publishing Company, pp. 279-302.

Marandin, J.-M. (1993), « Analyseurs syntaxiques, équivoques et problèmes », in *T.A.L.*1, pp. 5-33.

Milner, J.-C. (1989), *Introduction à une science du langage*, Paris, Seuil.

Sager, N., Friedman, C. & Lyman, M. (eds) (1987), *Medical Language Processing : Computer Management of Narrative Data*, Massachusetts, Addison Weslay, Reading.

Tanguy, L. & Rebeyrolle, J. (2001), « Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires », in *Cahiers de grammaire* 25.

Vergne, J. (1998), « Entre arbre de dépendance et ordre linéaire, les deux processus de transformation : linéarisation, puis reconstruction de l'arbre », in *Cahiers de Grammaire* 23, pp. 95-136.