Groupes prépositionnels arguments ou circonstants : vers un repérage automatique en corpus

Cécile Fabre et Cécile Frérot

ERSS, UMR 5610, Université Toulouse-Le Mirail
cfabre@univ-tlse2.fr, cessfrerot@aol.com

Résumé – Abstract

Dans cette étude, menée dans le cadre de la réalisation d'un analyseur syntaxique de corpus spécialisés, nous nous intéressons à la question des arguments et circonstants et à leur repérage automatique en corpus. Nous proposons une mesure simple pour distinguer automatiquement, au sein des groupes prépositionnels rattachés au verbe, des types de compléments différents. Nous réalisons cette distinction sur corpus, en mettant en oeuvre une stratégie endogène, et en utilisant deux mesures de productivité : la productivité du recteur verbal vis à vis de la préposition évalue le degré de cohésion entre le verbe et son groupe prépositionnel (GP), tandis que la productivité du régi vis à vis de la préposition permet d'évaluer le degré de cohésion interne du GP. Cet article présente ces deux mesures, commente les données obtenues, et détermine dans quelle mesure cette partition recouvre la distinction traditionnelle entre arguments et circonstants.

This paper addresses the issue of the automatic distinction between arguments and adjuncts within the framework of a corpus-based parser. We propose a simple measure aimed at distinguishing automatically different types of complements among prepositional phrases (PP) attached to the verb. This corpus-based approach relies on an endogeneous technique as well as two measures of productivity. The productivity of a governor and its preposition determines the degree of cohesion between a verb and its PP. Conversely, the productivity of a governee assesses the degree of autonomy of a PP. After presenting both measures, we comment on the output data and determine whether this division characterises the traditional distinction between arguments and adjuncts.

Mots Clés – Keywords

analyse syntaxique automatique, analyse endogène, productivité, argument, circonstant

automatic parsing, endogeneous analysis, productivity, argument, adjunct

1 Objectif

Dans le domaine de l'analyse syntaxique automatique, les systèmes élucident à des degrés très divers la structure relationnelle des phrases d'un texte. Entre l'analyse syntaxique superficielle, qui effectue un simple parenthésage des composants maximaux, et l'analyse complète qui vise à construire l'arbre syntaxique de la phrase, il est possible d'aller plus ou moins loin dans la résolution des liens syntaxiques, en fonction des informations que l'on souhaite acquérir pour une application donnée : relations de dépendance binaire pour l'analyse distributionnelle et la structuration de terminologie (Nazarenko et al., 2001) (Bourigault, Lame, 2002), cadres de sous-catégorisation pour la constitution de lexiques (Federici et al., 1998) (Briscoe, Carroll, 1997), etc.

Notre cadre de travail est la construction d'un outil d'analyse syntaxique de corpus, Syntex (Bourigault, Fabre, 2000). Dans ce contexte, nous nous intéressons ici à la question du rattachement prépositionnel, et cherchons à déterminer une stratégie qui permette d'aller au-delà de la décision simple de rattachement (à quel recteur se rattache une préposition) pour rendre possible une typologie des compléments. En l'état, l'analyseur détermine à quelle tête se rattache un groupe prépositionnel. On constate alors que les GP ainsi rattachés ont des statuts très divers et manifestent au moins deux grands types de dépendance lexicale avec leur recteur, lorsque celui-ci est un verbe (Fabre, Bourigault, 2001). Le régi peut être de nature argumentale (*ressembler à un cratère*) ou circonstancielle (*disparaître dans le gouffre*). Rappelons, avant de développer ce point en 2.1, que la distinction entre arguments et circonstants oppose des compléments essentiels étroitement liés au verbe du point de vue syntaxique et sémantique, et des compléments satellites caractérisant secondairement les circonstances du procès décrit par le verbe (Miller, 1997).

Nous faisons l'hypothèse que ces deux types de relations doivent être pris en compte pour contribuer conjointement à la tâche de construction de classes sémantiques : les mots peuvent en effet être regroupés parce qu'ils partagent des arguments ou sont arguments des mêmes unités, mais des relations non argumentales permettent également de mettre au jour des ensembles homogènes de mots. Par exemple, dans notre corpus de géomorphologie[1], les mots (lemmes) *nappe*, *rigole*, *film* qui entretiennent tous les trois une relation circonstancielle de manière avec le verbe *ruisseler* (*ruisseler en nappe, ~ rigole, ~ film*) présentent une affinité sémantique. De même les mots *bassin*, *gouffre*, *lac*, *puits*, qui entretiennent avec le verbe *disparaître* une relation de type locatif (*disparaître dans un bassin, ~ un gouffre, ~ un lac, ~ un puits*). Par ailleurs, l'identification des relations circonstancielles qui structurent un texte (relations d'ordre locatif par exemple, dans notre corpus, comme on va le voir) fournit des indications précieuses sur le contenu du texte. Loin de négliger ce type de complémentation, nous cherchons donc à l'identifier au même titre que la relation argumentale, tout en les distinguant dans la mesure où leur contribution sémantique est très différente. Ces deux types de relation ne peuvent donc pas être simplement combinés, cumulés. L'apport sémantique de chacun doit être distingué. Est-il possible d'opérer cette distinction de manière endogène (Bourigault, 1994), c'est-à-dire grâce aux seules informations acquises au sein du corpus analysé ?

---

[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, d'avoir mis à la disposition de l'équipe ce corpus, extrait de la base SCITECH.

2 Distinction argument - circonstant. Etat de l'art

L'opposition argument-circonstant, traditionnelle en linguistique, s'avère également cruciale en TAL : la description de la structure argumentale des verbes est une dimension majeure de l'activité de construction de ressources lexicales, car elle guide l'analyse syntaxique, et permet en particulier d'identifier les relations actancielles qui structurent un texte, données fondamentales pour les systèmes d'extraction d'information. Les recherches s'orientent vers l'acquisition automatique de ces informations à partir de textes, dans la mesure où la forte variabilité des cadres de sous-catégorisation d'un type de texte à l'autre rend inefficace le recours à des ressources lexicales générales (Basili et al., 1997). Nous faisons le point sur les critères définis jusqu'à présent pour distinguer ces deux types de compléments.

2.1 Critères linguistiques

La distinction argument-circonstant est classique et a été abondamment étudiée. Elle apparaît, sous divers vocables (compléments de verbe, essentiels, régis, sous-catégorisés *vs* modifieurs, ajouts, satellites, etc.), dans la plupart des théories linguistiques. Pour (Borillo, 1990), il s'agit d'établir une distinction entre « les compléments qui s'attachent [au verbe] dans des combinaisons bien réglées » et font partie de son « schéma de construction », et celles « qui apparaissent dans son environnement mais qui ne dépendent pas directement de ses propriétés structurelles ». (Miller, 1997) oppose selon le même principe des compléments régis par le verbe, dont les « propriétés sont régies de façon étroite par celui-ci », et des circonstants, dont « le statut est indépendant de celui du verbe ». Les synthèses, particulièrement claires, de (Miller, 1997) et (Bonami, 1999) rappellent la diversité des critères linguistiques mis en œuvre pour cerner cette opposition : obligatorité syntaxique et sémantique, déplacement en tête de phrase, itérabilité, passivation, topicalisation, degré de sélection par le verbe, etc. La plupart de ces critères sont difficiles à intégrer dans une procédure automatique, dans la mesure où ils mettent en jeu des jugements sémantiques et de grammaticalité. Parmi ces critères cependant, trois nous semblent facilement testables en corpus :

- le caractère obligatoire ou facultatif (*Il consiste* {*en une recristallisation*| *∅ } vs *Il se dédouble* {*sur la zone* |∅}).

- la détermination exercée par le verbe à l'égard de la préposition : le verbe contraint fortement la préposition de son argument (*Il consiste* {*en*|*∅|*à*|**sur*) *une recristallisation*), il contraint faiblement celle de son circonstant (*Il se dédouble* {*sur*|*dans*|*sous*|*près de*|*autour de*} *la zone*).

- le déplacement du GP en tête de phrase : **En une recristallisation, il consiste* vs *Près de la zone, il se dédouble.*

Les deux auteurs s'accordent cependant sur l'absence de critère universel permettant d'opposer ces deux types de compléments, ainsi que sur les résultats contradictoires de certains critères. Pour Miller, on peut tout au plus définir une classe d'arguments et de circonstants prototypiques, qui sont respectivement les objets directs et les compléments locatifs ou temporels.

Nous sommes donc confrontées à une notion essentielle sur le plan linguistique, mais qui repose sur une caractérisation floue, mal assurée, suggérant plutôt un *continuum* qu'une

rupture franche entre les deux types de compléments. A l'inverse, les options prises en TAL cherchent à maintenir, comme on va le voir, l'hypothèse d'une opposition binaire.

2.2 Critères automatiques

Les recherches orientées vers l'acquisition de ressources lexicales visent à identifier dans les textes les arguments des verbes de manière à les intégrer dans des dictionnaires et à alimenter des analyseurs (Manning, 1993) (Meyers et al., 1994) (Briscoe, Carroll, 1997). Peu de travaux mettent spécifiquement l'accent sur la distinction entre arguments et circonstants, qui reste bien souvent implicite. Ceux qui mentionnent néanmoins ce problème sont basés sur deux types de techniques : le partage entre les deux types de compléments est réalisé à partir de corpus annotés syntaxiquement – typiquement, le Penn Treebank (Buchholz, 1998). En l'absence de données d'apprentissage, c'est la fréquence de l'association entre le verbe et son complément qui est proposée comme critère (Brent, 1993) ou la présence côte à côte du verbe et de son complément (Federici et al., 1998). Considérons tout d'abord le critère de la fréquence : (Basili et al., 1999) estiment à l'inverse de Brent que cette distinction ne peut pas être réalisée de manière automatique, dans la mesure où les compléments adjoints contribuent au même titre que les arguments à la sémantique du verbe et lui sont associés avec la même régularité (« usually accidental (i.e. not argumental), though frequently observed with some verb », (Basili et al., 1997)). Les observations faites sur notre corpus vont dans ce sens : s'il est vrai que les fréquences hautes correspondent à des liens argumentaux, on constate que les fréquences moyennes (entre 3 et 10 occurrences) mêlent indistinctement arguments et circonstants. Le deuxième critère, celui de l'apparition côte à côte du verbe et de son complément, correspond au critère exploité par notre analyseur pour déterminer les contextes d'acquisition d'une relation syntaxique : les contextes ambigus (ex : *creuse des gorges dans la lave*), qui offrent plusieurs recteurs possibles pour une préposition, sont résolus à partir de l'observation de contextes non ambigus (ex : *creusé dans la lave*). Néanmoins, on s'aperçoit que l'apparition du complément dans le voisinage immédiat du verbe est un indice fiable pour déterminer son rattachement à celui-ci, mais pas pour déterminer la nature de ce complément, puisque les circonstants peuvent également figurer dans cette position. Nous proposons donc de nous appuyer plus nettement sur la caractérisation linguistique de ces deux types de compléments de manière à définir une nouvelle méthode pour les distinguer en corpus.

3 Distinguer automatiquement arguments et circonstants

En l'absence d'informations sur la valence du verbe (dictionnaire ou corpus annoté), nous cherchons des moyens d'éprouver le degré de cohésion syntaxique et sémantique entre le verbe et son complément : les critères linguistiques proposés (capacité du circonstant à s'effacer, se déplacer) visent en effet à tester l'autonomie *vs* la dépendance du GP vis à vis du verbe. Les circonstants manifestent une certaine indépendance par rapport au verbe, dans la mesure où ni leur position ni leur interprétation n'est conditionnée par lui. A l'inverse, les arguments sont contraints formellement et sémantiquement par le verbe (Miller, 1997). Nous allons voir que le corpus offre un espace au sein duquel la cohésion du GP peut être mesurée selon des critères nouveaux.

3.1 La productivité comme mesure de cohésion du GP

Le rattachement automatique des groupes prépositionnels est effectué dans notre analyseur de manière endogène, sur la base d'indices recueillis à partir des contextes non ambigus du corpus. Parmi ces indices, une mesure permet d'évaluer le degré de cohésion du GP avec son recteur : la productivité du couple recteur-préposition (désormais prodRecteur). Une prodRecteur élevée est l'indice que le recteur se construit régulièrement avec la préposition, dans la mesure où les mots régis par cette relation sont diversifiés. Par exemple, le couple (*alterner, avec*) est productif, avec une prodRecteur de 7, parce qu'il gouverne sept régis différents : *alterner avec+dét* (*cendre*|*crue*|*lit*|*masse*|*période*|*section*|*surface*).

Jusqu'à présent, nous avons utilisé cette mesure pour déterminer à quel recteur candidat se rattache une préposition (Bourigault, Fabre, 2001). Si deux recteurs potentiels figurent dans la zone de rection du verbe, l'analyseur choisit prioritairement (en combinaison avec d'autres indices) celui dont la productivité avec la préposition est la plus forte. Nous proposons ici d'utiliser cette mesure en la combinant à une seconde mesure de productivité : la productivité des couples préposition-régi (désormais prodRégi). Cette fois, le régi est dit productif pour une préposition donnée si le couple prép-régi est gouverné par une diversité de recteurs. Par exemple, le couple (*à+dét*, *quaternaire*) est productif avec prodRégi = 3 parce qu'il est gouverné par 3 verbes différents dans le corpus : (*englacer*|*évoluer*|*subir*) *au quaternaire*.

L'expérience que nous présentons est simple : elle consiste à extraire automatiquement du corpus l'ensemble des triplets *recteur-relation prépositionnelle-régi* tels que le recteur est une forme verbale (*infinitif, participe, forme conjuguée*), la relation prépositionnelle est une séquence *préposition + dét* (*dét* pouvant être vide), et le régi est un nom ou un infinitif. Ces triplets sont extraits de contextes non ambigus (c'est-à-dire, sans ambiguïté de rattachement) dans lesquels la préposition est contiguë au verbe (*modulo* un adverbe). Le régi est alors la première forme nominale ou infinitive que l'on trouve à droite de la préposition. Ainsi, nous extrayons de la phrase *H Baulig insista sur le fixisme tectonique de vastes régions à des époques récentes* le triplet (*insister, sur_le, fixisme*). Les deux mesures de productivité doivent nous permettre dans un deuxième temps de répartir ces GP en deux groupes : les GP présentant une cohésion forte avec le verbe, les GP présentant une cohésion forte en interne. Cette répartition recoupe-t-elle la distinction argument-circonstant ?

3.2 Premiers résultats

Nous livrons ici un extrait des données brutes correspondant aux couples recteur-prép et prép-régi les plus productifs. Nous commentons ces premières données avant d'établir et d'évaluer une procédure de calcul (3.3) qui vise à déterminer automatiquement la nature argumentale ou circonstancielle de la relation prépositionnelle.

- Productivité forte du recteur

Les couples suivants, numérotés de 1 à 20, correspondent aux recteurs les plus productifs pour une relation prépositionnelle donnée (valeur de ProdRecteur entre 15 et 163) :

> *devoir à+dét* (1), *expliquer par+dét* (2), *traduire par+dét* (3), *tendre à+vinf* (4), *correspondre à+dét* (5), *caractériser par+dét* (6), *séparer par+dét* (7), *reposer sur+dét* (8), *transformer en+dét* (9), *réduire à+dét* (10), *trouver dans+dét* (11), *aboutir à+dét* (12), *donner à+dét* (13), *appliquer à+dét* (14), *renseigner sur+dét*

(15), *attribuer à+dét* (16), *continuer à+vinf* (17), *commencer à+vinf* (18), *passer à+dét* (19), *finir par+vinf* (20)

Si l'on sélectionne à présent au hasard vingt couples correspondant à une valeur plus modeste de ProdRecteur (valeur 4 ou 5), on obtient la liste suivante :

collaborer avec+dét (21), *coïncider avec+dét* (22), *causer par+dét* (23), *atteindre par+dét* (24), *affleurer sur+dét* (25), *obliger à+vinf* (26), *destiner à+vinf* (27), *varier selon+dét* (28), *transporter par+vinf* (29), *tomber à+dét* (30), *tendre à+dét* (31), *suivre par+dét* (32), *subdiviser en+dét* (33), *situer sous+dét* (34), *retrouver sur+dét* (35), *renoncer à+dét* (36), *réaliser par+dét* (37), *procéder par* (38), *préparer par+dét* (39), *prendre pour* (40)

Ces résultats appellent plusieurs commentaires. Notons tout d'abord que la forme lemmatisée annule la distinction forme conjuguée *vs* participe passé. Ainsi, les lemmes *devoir*, *suivre*, *réaliser*, renvoient en fait à des formes participiales (*dû à*, *suivi par*, *réalisé par*). A la lecture de ces deux listes, on constate qu'une proportion importante de ces couples expriment effectivement une relation de nature argumentale :

- 22 couples[2] sont mentionnés dans le *Trésor de la Langue Française* comme construction du verbe correspondant,

- 11 autres couples nous semblent relever également d'un rapport argumental (*expliquer par+dét*, *traduire par+dét*, *caractériser par+dét*, *séparer par+dét*, *renseigner sur+dét*, *passer à + dét*, *affleurer sur*, *varier selon*, *situer sous*, *retrouver sur+dét*, *procéder par*). 3 d'entre eux apportent des indications locatives à des procès de type locatif,

- 6 couples renvoient à une relation agentive (complément en *par* d'une forme passive : *atteindre par*, *causer par+dét*, *transporter par+vinf*, *suivre par+dét*, *réaliser par+dét*, *préparer par+dét*).

Au final, il reste donc seulement un couple dans cette liste dont le caractère cohésif est douteux, ou en tout cas moins facilement interprétable : *tomber à+dét* (*centimètre*|*dessous*|*valeur*|*vitesse*). Même dans ce cas, l'observation des contextes nous fait cependant pencher vers une lecture plutôt argumentale (*cette valeur tombe à quelques centimètres*, *on tombe déjà à des valeurs plus faibles*).

A ce stade, nous n'avons pas encore les moyens de caractériser la relation ternaire entre un verbe, une préposition et la tête du GP. La mesure prodRecteur nous permet donc seulement de déterminer globalement une relation verbe + prép, sachant qu'en réalité cette relation peut être hétérogène. Ainsi, si la relation *aboutir à+dét X* renvoie à une relation argumentale pour 17 des 18 valeurs de X dans le corpus, on trouve une relation circonstancielle avec *aboutir à leur tour*. Le couple verbe-prép ne suffit pas toujours à caractériser la nature de la relation. La mesure est donc encore insuffisante pour lever l'ambiguïté.

[2] Couples 1, 4-5, 8-14, 16-18, 20-22, 26-27, 31, 33, 36 et 40.

- Productivité forte du régi

Nous nous intéressons à présent au cas des GP rattachés à une diversité de verbes (au moins 3). Nous devrions avoir affaire à des unités peu cohésives avec les verbes qu'elles modifient, et présentant une certaine autonomie sémantique. Voici la liste des vingt-cinq premiers couples prép-régi extraits par cette méthode :

> *dans la roche, par l'érosion, dans la région, par l'eau, dans la zone, à la surface, sur la pente, sous la forme, sur le fond, par l'action, au dessous, vers l'aval, sur le versant, par l'érosion, à l'époque, dans le cas, dans la vallée, par le courant, au sommet, dans le sens, au niveau, à l'érosion, au contraire, sous le climat, sous l'action*

La cohésion interne de ces groupes est tout d'abord attestée par la présence de locutions adverbiales (*au contraire*) et prépositionnelles (*dans le cas*, *dans le sens*, *au niveau*, *sous la forme*). Certaines séquences s'apparentent à des locutions sans relever tout à fait de cette catégorie. Ainsi, le groupe *à la surface* (c'est vrai aussi des séquences *au sommet*, *sur le fond*) fonctionne dans le corpus à la fois seul - *remonter à la surface, arriver ~, accumuler ~*, et comme introducteur de groupe nominal - *à la surface de la terre, ~ de la glace*. En dehors de ces cas, cette liste regroupe des séquences qui peuvent fonctionner comme circonstants, et expriment massivement une information de type spatial (*dans la roche, sous le climat, sur le versant*, etc.) ou temporel (*à l'époque*). On retrouve dans cette liste des groupes agentifs : *par l'érosion, par l'eau*. Comme nous l'avons vu à l'étape précédente pour les couples recteur-prép, ces séquences prép-régi peuvent également apparaître dans des relations de nature argumentale. Ainsi, le GP *à la surface* entretient une relation circonstancielle avec 11 des 16 recteurs verbaux recensés, mais on trouve aussi les séquences *s'intéresser à la surface, s'attaquer à la surface*, etc.

Un cas reste difficilement interprétable, il s'agit de la séquence *à l'érosion*. Il ne semble pas autonome sémantiquement, dans la mesure où, en l'absence de recteur, il n'est pas possible de caractériser en termes de fonction (agentive, locative, temporelle, etc.) l'indication qu'il exprime. Observons la liste des verbes, au nombre de 10, qui gouvernent ce GP :

céder, devoir, exposer, fournir, immuniser, offrir, permettre, présenter, résister, soumettre

On constate qu'il s'agit de verbes qui sous-catégorisent un GP introduit par *à*. On trouve d'ailleurs, pour 5 de ces 10 verbes, une valeur prodRecteur forte avec cette préposition dans le corpus. La valeur prodRégi ne peut donc être utilisée indépendamment de la valeur prodRecteur. Nous introduisons par conséquent un test plus complet, qui croise les deux mesures de productivité que nous venons de présenter.

3.3 Croiser deux mesures de productivité

Nous avons cherché à affiner ce premier traitement et à évaluer les résultats d'une méthode combinant les mesures prodRecteur et prodRégi. Dans ce but, nous avons défini une procédure qui décide, pour un triplet recteur-prép(+dét)-régi donné, si l'on a affaire à une relation argumentale ou circonstancielle :

Si, pour une préposition donnée, la productivité du recteur est forte (> 2) et la productivité du régi est nulle, alors le groupe prépositionnel sera étiqueté comme argument (fichier ARG).

Si, pour une préposition donnée, la productivité du régi est forte (> 2) et la productivité du recteur est nulle, alors le groupe prépositionnel sera étiqueté comme circonstant (fichier CIRC).

Nous avons effectué ce calcul sur tous les triplets du fichier de départ. Les deux fichiers résultats sont composés de 1365 triplets classés comme arguments et 525 triplets classés comme circonstants. A titre d'exemple, le tableau 1 présente un extrait du résultat, correspondant aux 10 premiers verbes trouvés dans les deux séries (par ordre alphabétique).

ProdRecteur > 2, prodRégi = 0	ProdRégi > 2, prodRecteur = 0
aboutir à l'aplanissement	*abandonner dans le sol*
accumuler dans la mouille	*accélérer sur la pente*
adapter à l'ondulation	*accidenter en général*
affleurer sur les berges	*accorder dans le cas*
agir dans le canyon	*acquérir dans la région*
alimenter en débris	*adapter dans le détail*
aller jusqu'à arracher	*affleurer à nu*
alterner avec la cendre	*affouiller dans la roche*
apparaître dans le cratère	*alimenter dans la zone*
apparenter à des gabros	*allonger dans la direction*

Tableau 1 : échantillon des constructions verbales extraites (fichiers ARG et CIRC)

3.4 Evaluation

Les résultats ont été évalués par deux méthodes parallèles (tableaux 2 et 3). 50 triplets ont été sélectionnés aléatoirement dans le fichier CIRC, 50 dans le fichier ARG. Les 100 triplets ainsi retenus ont été triés par ordre alphabétique et soumis dans leur contexte d'origine (limité à une phrase) à une collègue linguiste, qui a caractérisé chaque GP comme argument ou circonstant.

fichier d'origine	GP évalué argument	GP évalué circonstant
fichier ARG	88%	12%
fichier CIRC	28%	72%

Tableau 2 : résultats de l'évaluation manuelle

Ce résultat est corroboré par un deuxième type d'évaluation. Nous avons comparé les séquences testées avec les entrées du *Trésor de la Langue Française*, à la recherche de la mention du GP dans l'entrée du verbe, celle-ci pouvant figurer : *i)* dans une construction verbe+prep ; *ii)* entre crochets ou parenthèses (*la séparation est concrètement matérialisée* : *diviser (qch) en (parties…)*) ; *iii)* dans le cas où seul l'emploi du verbe est spécifié, au sein d'un exemple (*se déposer* : *emploi pronominal. Une broux de noix qui s'est déposée dans les plis de sa paume).*

fichier d'origine	GP décrit dans le *TLF*	GP non décrit
fichier ARG	84%	16%
fichier CIRC	24%	76%

Tableau 3 : comparaison des résultats avec les données du *TLF*

Le recours au dictionnaire ne rend bien entendu pas compte de certaines idiosyncrasies du corpus. En effet, certains sens particuliers de verbes, associés à une construction donnée, sont absents du *TLF* (*s'encaisser dans le lit, être exprimé en mètres, modeler en creux, etc.*).

4 Conclusion et perspectives

La mesure de productivité telle que nous l'avons définie et mise en œuvre sur corpus fournit un nouvel outil pour caractériser l'opposition entre argument et circonstant. Nous nous sommes ainsi dotées d'une mesure permettant d'évaluer la cohésion d'un GP pour un verbe donné. Les deux évaluations que nous avons menées témoignent du bon taux de précision de la méthode et par conséquent de la pertinence de la mesure de productivité pour caractériser le lien de complémentation entre un recteur et son GP. Cette première expérience doit être prolongée à plusieurs niveaux. Pour le moment, nous n'évaluons que des séquences verbe+GP pour lesquelles l'une des deux productivités (recteur ou régi) est nulle, ce qui impose une contrainte forte sur les données susceptibles d'être ainsi caractérisées. Nous projetons à ce titre d'étudier les cas intermédiaires, ce qui pourrait nous engager vers une (meilleure) caractérisation du *continuum* dont font état de nombreux auteurs. Par ailleurs, l'impact du type de corpus sur la méthode doit être évalué afin de déterminer si son efficacité perdure à l'épreuve d'un corpus moins « technique », présentant moins de redondance. Enfin, nous envisageons à court terme d'exploiter d'autres pistes pour améliorer la description de ces GP, et en particulier d'utiliser le critère de détermination de la préposition par le verbe, un des tests linguistiques dont nous avons rendu compte en 2.1. Dès à présent, au vu de nos premières observations, nous constatons en effet une forte relation entre la productivité et l'association d'un recteur à une seule préposition (*aboutir à*, prodRecteur = 11) ou à plusieurs (*développer sur/sous/par/à/dans*, prodRecteur = 3), ce qui pourrait constituer un critère supplémentaire pour mettre en évidence différents types de compléments verbaux en corpus. L'impact de cette différenciation de contextes argumentaux et circonstanciels sur la modélisation du contenu des textes, et en particulier sur l'analyse distributionnelle, doit ensuite être étudié et évalué.

Remerciements

Nous remercions Anne Le Draoulec d'avoir participé à l'évaluation de nos données ainsi que Didier Bourigault pour sa relecture attentive de l'article.

Références

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

Bonami O. (1999), *Les constructions du verbe : le cas des groupes prépositionnels argumentaux*, Thèse de doctorat, Université Paris 7.

Borillo A. (1990), A propos de la localisation spatiale, *Langue française*, Vol.86, pp.75-84.

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

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

Bourigault D., Lame G. (2002), Analyse distributionnelle et structuration de terminologie. Application à la construction d'une ontologie documentaire du Droit. *T.A.L*, Hermès, Vol.43:1, à paraître.

Brent M-R. (1993), From Grammar to Lexicon : Unsupervised Learning of Lexical Syntax, *Computational Linguistics*, Vol.19 : 2, pp.243-262.

Briscoe T., Carroll J. (1997), Automatic Extraction of Subcategorization from Corpora, Actes de *5 th Conference on Applied NLP*, Washington, 356-363.

Buchholz S. (1998), Distinguishing complements from adjuncts using memory-based learning, B. Keller (Ed.), Actes de *ESSLLI-98 Workshop on Automated Acquisition of Syntax and Parsing*, 41-48.

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

Federici S., Montemagni S., Pirrelli V., Calzolari N. (1998), Analogy-based Extraction of Lexical Knowledge from Corpora: the SPARKLE Expérience, Actes de *LREC*, Grenade, 75-82.

Manning C. (1993), Automatic acquisition of a large subcategorization dictionary from corpora, Actes de *31th meeting of the association for Computational Linguistics*, ACL, Morristown, 235-242.

Miller P. (1997), Compléments et circonstants : une distinction syntaxique ou sémantique ?, Actes du *37ème Congrès de la SAES*, édité sous la direction de J.-C. Souesme, 91-103, Presses Universitaires de Nice, 1998.

Meyers A., Macleod C., Grishman R. (1994), *Standardization of the Complement/Adjunct Distinction*, Abstract, Proteus Project Memorandum 64, Computer Science Department, New-York University.

Nazarenko A., Zweigenbaum P., Habert B., Bouaud J. (2001), Corpus-based extension of a terminological semantic lexicon. In Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme (eds), *Recent Advances in Computational Terminology*, chap. 16, pp. 327-351, John Benjamins, Amsterdam.