Une approche de la complémentation verbale guidée par les corpus

Cécile Fabre, Josette Rebeyrolle

UNE APPROCHE DE LA COMPLEMENTATION VERBALE GUIDEE PAR LES CORPUS

Cécile FABRE Josette REBEYROLLE

1. Introduction

Le développement de travaux qui se revendiquent de la linguistique de l'usage est en partie concomitant avec la constitution de corpus informatisés et la mise au point de méthodes quantitatives d'investigation de ces vastes ensembles de données textuelles. Parmi ces travaux, on peut citer deux approches principales : la linguistique *corpus-based* (basée sur les corpus) et *corpus-driven* (guidée par les corpus). Cette opposition, théorisée notamment par Tognini-Bonelli (2001), rend compte de la différence entre des travaux qui utilisent les corpus pour étudier le comportement de catégories préétablies, de modélisations préalables (dans la lignée de Geoffrey Leech), et des travaux qui placent le corpus au cœur du processus d'élaboration des catégories et pour lesquels la modélisation est le résultat d'un examen systématique et sans *a priori* des corpus (dans la lignée de John Sinclair). Nous décrivons ici une approche intermédiaire, qui, tout en se situant dans la lignée d'une linguistique *corpus-driven*, s'appuie sur des annotations issues de l'application d'outils de traitements automatiques consistant notamment à segmenter et étiqueter au préalable les unités d'analyse. Notre approche se donne comme point de départ une question linguistique largement débattue, celle de la complémentation verbale indirecte, et exploite un large corpus lemmatisé, catégorisé et analysé syntaxiquement. Plus précisément, nous proposons des mesures statistiques fondées sur les résultats de l'analyse syntaxique, qui permettent d'apprécier la diversité des relations qu'un verbe peut entretenir avec un syntagme prépositionnel (SP), en position de complément. Ces mesures servent à organiser les données issues du corpus. Notre objectif est en effet de classer des constructions de type [V + Prép + SN] en utilisant des indices statistiques évaluant le degré d'autonomie ou de cohésion entre les éléments de la construction.

La fonction des groupes prépositionnels est habituellement évaluée à l'aide de tests linguistiques de grammaticalité permettant de déterminer s'il s'agit d'arguments ou de circonstants. C. Manning a présenté cette question comme une des meilleures illustrations de l'intérêt de passer d'une vision catégorique de la syntaxe à une vision probabiliste :

> « This conception of the argument/adjunct distinction is the best one can do in the categorical 0/1 world of traditional formal grammars: things have to be either selected (as arguments) or not. (...) However, categorical models of selection have always been problematic. The general problem with this kind of model was noticed early on by Sapir (1921:38) who noted that "All grammars leak". In context, language is used more flexibly than such a model suggests. » (2003, p. 298).

* Université de Toulouse, CLLE-ERSS (UMR 5263)

Nous proposons de fait de substituer à cette approche binaire une démarche permettant d'observer des tendances dans le comportement des syntagmes prépositionnels vis-à-vis du verbe, qui se dessinent dans l'espace du corpus considéré. Cette approche inductive de la complémentation verbale, fondée sur l'exploration systématique de grandes collections de textes et le recours à des mesures statistiques permet d'aborder sous l'angle de la quantification l'étude des groupes prépositionnels, de manière à apprécier leur degré d'autonomie vis-à-vis du verbe.

Cette étude nous conduit à envisager à double titre la dimension de l'usage discutée dans ce numéro : d'abord, la caractérisation des SP est fondée exclusivement sur l'analyse de leur comportement dans le corpus. On bénéficie alors d'un dispositif d'observation qui permet de dégager des profils d'usage : profils de SP selon le degré d'autonomie, mais aussi profils de verbes ou de prépositions selon leur tendance à introduire des SP autonomes ou cohésifs. Ensuite, parce que nous avons fait le choix de travailler sur deux corpus différents (un corpus littéraire, sous-ensemble de *Frantext*, et un corpus journalistique extrait des archives du *Monde*), ce qui nous permet d'apprécier la part de stabilité et de diversité dans les comportements des compléments selon les types de textes.

Après avoir présenté notre approche et notamment la méthode de calcul de l'autonomie des SP, nous montrons concrètement comment ces calculs permettent de guider la description linguistique en faisant émerger des types de fonctionnement particuliers. Plusieurs pistes d'observation se dégagent lorsque l'on examine de façon détaillée les résultats obtenus par cette approche (section 2). Nous posons en particulier deux questions : quels types de groupes prépositionnels peut-on placer le long du continuum entre groupes verbaux étroitement cohésifs et groupes prépositionnels fortement autonomes ? (section 3) Quelles informations sont mises au jour concernant le fonctionnement prépositionnel, et peut-on observer des différences d'un corpus à l'autre ? (section 4).

2. Présentation de l'approche

2.1. Des tests de grammaticalité aux mesures sur corpus

Le fait est bien connu : les tests de grammaticalité peinent à départager des compléments structurellement rattachés au verbe – les arguments, ou compléments – et des satellites qui accompagnent le verbe sans être sous-catégorisés – les circonstants, ou ajouts. Parmi l'ensemble des tests habituellement convoqués pour cette tâche, Bonami (1999) parvient à dégager deux critères probants. Le premier concerne le caractère obligatoire du dépendant, exprimé en termes de condition suffisante (s'il est obligatoire, c'est un argument), le second concerne sa position (la position pré-verbe fini ne peut être occupée que par un ajout). On voit que les critères discriminants sont très spécifiques. Ils prouvent cependant, selon Bonami, la validité de l'opposition. Au total, son analyse conclut au maintien de la distinction. Cette thèse tend à être corroborée sur le plan cognitif : dans un article de synthèse sur la question, Tutunjian et Boland (2008) font état d'expériences en psycholinguistique qui démontrent que, dans certaines tâches de compréhension, les lecteurs consacrent moins de temps à la lecture des SP arguments qu'à celle des ajouts, et que cette différence de statut joue un rôle important dans le processus de désambiguïsation. Les auteurs précisent que ces éléments ne sont pas absolument incontestables, et signalent d'autres expériences qui n'étayent pas cette distinction ; elles concluent néanmoins positivement à la question posée dans le titre de leur article (« Do we need a distinction between arguments and adjuncts ? ») : « The evidence is not entirely conclusive on either front [formal linguistics and psycholinguistics], but on

balance, the psycholinguistic evidence supports a formal distinction between arguments and adjuncts. » (*op. cit*., p.645).

Dans une précédente étude (Fabre et al. 2008), nous avons envisagé le recours à des caractérisations quantitatives sur un gros volume de données comme moyen d'apporter de nouveaux éléments pour préciser cette distinction. Notre proposition consistait alors à exploiter les tests syntaxiques usuels pour départager par exemple les dépendants qui tendent à être obligatoires de ceux qui sont systématiquement optionnels, ou ceux qui tendent à occuper une position fixe de ceux qui occupent des positions variées par rapport au verbe. Finalement, notre étude montrait que l'on ne peut pas se contenter d'adapter les manipulations habituellement utilisées. D'abord, parce qu'elles ne permettent de se prononcer sur le statut de dépendants spécifiques que dans des configurations relativement rares qui ne sont pas facilement observables en contexte réel. Ensuite, parce qu'elles permettent d'observer un nombre limité de types de SP. Enfin, parce qu'elles ne disent rien, négativement, des SP qui n'ont pas été trouvés dans certaines positions – on sait que de la non apparition d'une structure en corpus on ne peut pas conclure à son impossibilité. Appliquer les tests linguistiques existants sur de vastes corpus ne constitue donc pas l'outil descriptif le plus efficace pour déceler les types de liens entre le SP et le verbe. Mais surtout, ces manipulations de type binaire (le groupe est-il obligatoire ou optionnel, déplaçable ou pas, etc.) ne permettent pas de rendre compte de la nature graduelle de la relation de dépendance entre un syntagme prépositionnel et un verbe. C'est également le constat que font, après d'autres (comme par exemple Borillo 1990 ou Miller 1998), Lacheret-Dujour et François (2004) en proposant un modèle modulaire qui articule des informations sémantiques, pragmatiques et prosodiques pour rendre compte du fait que le passage entre compléments intra- et extra-prédicatifs (pour reprendre leur terminologie) opère graduellement.

La piste de travail que nous avons proposée, et que nous reprenons ici, consiste à abandonner les critères de caractérisation usuels pour exploiter de manière systématique la masse d'informations distributionnelles que fournit un grand corpus annoté. Plus précisément, il s'agit de concevoir des tests originaux, plus directement adaptés aux outils qu'offre le travail en corpus. Nous cherchons ainsi à traduire en termes statistiques la tendance à la cohésion ou à l'autonomie par rapport au verbe qui fonde l'opposition entre complément et ajout sur la base des faits suivants : 1) les ajouts sont caractérisés par leur indépendance syntaxique et sémantique par rapport au verbe, dans la mesure où ni leur position ni leur interprétation n'est conditionnée par lui ; et 2), inversement, les arguments sont contraints formellement et sémantiquement par le verbe. Nous montrons comment une telle approche, en fournissant des éléments de quantification du degré d'autonomie du groupe prépositionnel (section 2.2), permet de sonder le continuum entre arguments et ajouts (section 3) et guide la description du fonctionnement de ces groupes dans les corpus considérés (section 4).

2.2. Une mesure de l'autonomie des SP par rapport au verbe

Nous avons mis au point une mesure qui vise à estimer le degré d'autonomie du SP par rapport au verbe auquel il se rattache. L'autonomie est appréciée selon un critère simple : un SP peu autonome a tendance à s'associer dans le corpus à des verbes qui régissent la préposition qui l'introduit. A l'inverse, un SP autonome se combine avec des verbes qui ne présentent pas d'association forte avec cette préposition. On a affaire, dans le cas des SP fortement autonomes, à des syntagmes qui se combinent avec une large gamme de verbes sans qu'une relation de sélection soit en jeu.

L'indice d'autonomie que nous construisons s'appuie sur des calculs de productivité effectués sur l'ensemble d'un vaste corpus, à partir des liens de dépendance posés automatiquement par un analyseur syntaxique, SYNTEX (Bourigault 2007). L'analyse a été effectuée sur deux corpus :

- Corpus LM10 : il s'agit du corpus comprenant 10 ans du journal *Le Monde*, paru entre 1991 et 2000, soit 200 millions de mots.
- Corpus FRANTEXT20 : ce corpus comprend la partie « romans du 20ème siècle » de la base textuelle *Frantext*[1].

La combinaison d'outils de traitement automatique et de méthodes statistiques explique notre choix en faveur de corpus de français écrit, et de taille importante. L'approche comparative que nous esquissons dans la section 4 nous a par ailleurs fait opter pour deux corpus assez nettement distincts par le genre (journalistique *vs* littéraire) – même si les textes de LM10 présentent une grande hétérogénéité au sein de cette vaste catégorie des textes journalistiques.

Ces corpus ont été segmentés et étiquetés, et l'analyseur a posé des liens de dépendance entre les unités de traitement qu'il a identifiées. La figure 1 montre les liens syntaxiques calculés sur une phrase simple. Les flèches relient un gouverneur à son dépendant. Ainsi, le verbe *échapper* gouverne deux adverbes, un sujet, et un syntagme prépositionnel. Le traitement du SP se décompose comme suit : le verbe gouverne la préposition (lien PREP), laquelle gouverne le nom (lien NOMPREP), qui gouverne à son tour un déterminant.

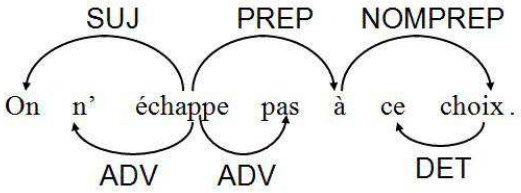

Figure 1 : Extrait d'analyse syntaxique

La méthode que nous présentons s'applique à l'étude des syntagmes prépositionnels en position post-verbale. Nous nous sommes intéressées à cette position dans la mesure où l'analyseur ne rattache au verbe que les SP situés dans cette zone. Partout ailleurs, ils sont laissés flottants, l'hypothèse d'un rattachement à la proposition plutôt qu'au verbe étant systématiquement privilégiée. Sachant que cette position postverbale peut être occupée aussi bien par des compléments du verbe que par des ajouts, on a affaire à une zone d'ambiguïté intéressante à explorer. De fait, SYNTEX rattache un SP à son gouverneur sans déterminer la fonction qui sous-tend cette relation. Dans les deux exemples suivants (extraits tirés de LM10), le SP *à l'audience* est rattaché par SYNTEX au verbe *découvrir* de la même manière que le groupe *à la pugnacité* est rattaché au verbe *tenir* et ce bien que, en [1], on ait affaire à un circonstant temporel et, en [2], à un complément indirect :

[1] M. Alain Schrantz (...) qui paraît découvrir à l'audience que ...
[2] Et la réponse tient sans doute, en partie, à la pugnacité (...)

On remarque dans l'exemple [2] que la position post-verbale ne se réduit pas aux cas où le SP suit immédiatement le verbe. Des adverbes et des incises peuvent l'en séparer.

[1] Nous remercions Jean-Marie Pierrel d'avoir mis ce corpus à notre disposition dans le cadre d'une convention passé entre l'ATILF et CLLE-ERSS.

Nous nous intéressons donc à toutes les instances de la configuration V + SP dans le corpus pour lesquelles l'analyseur a identifié un lien de dépendance entre le SP et le verbe. Ces configurations sont ramenées à des triplets de la forme (v, p, n), soit, pour les exemples [1] et [2], (*découvrir, à, audience*) et (*tenir, à, pugnacité*). Nous ne détaillons pas ici les procédures de désambiguïsation mises en place par l'analyseur. Elles sont exposées dans Bourigault (2007).

Pour calculer la force d'association entre le verbe et la préposition, nous nous fondons sur une série de mesures de productivité[2]. Souvent utilisée en morphologie (pour calculer la productivité d'un affixe) (Baayen 2009), parfois en syntaxe (s'agissant de la productivité d'une construction) (Barodal 2006), la productivité d'une configuration permet d'en estimer la régularité, en fondant le calcul de fréquence sur les types (*type frequency*) plutôt que sur les occurrences (*token frequency*). Ainsi, l'expression *revoir à la baisse* présente une fréquence élevée d'occurrences, mais elle n'est pas l'indice d'une régularité de la construction *revoir* + $[SP]_{à}$, dans la mesure où peu de noms peuvent se substituer à *baisse* dans cette position.
La productivité de la relation (v, p), notée *prod*(v, p), équivaut au nombre de noms N différents qui apparaissent dans l'ensemble des triplets (v, p, n). Nous affinons ce calcul à l'aide d'une mesure de productivité relative, notée $prod_R$(v, p), qui consiste à diviser la productivité du couple (v, p) par la productivité totale du verbe, notée $prod_T$, qui correspond à la somme de ses productivités pour toutes les prépositions avec lesquelles il se construit.

Cette mesure de productivité relative (dont les valeurs vont de 0 à 1) permet d'atténuer les effets liés à la fréquence d'occurrence de la forme verbale et de mieux apprécier le rôle relatif de chaque préposition dans l'ensemble des SP rattachés au verbe. Le tableau 1 donne quatre exemples illustrant différents degrés d'association d'un verbe avec la préposition *à*.

verbe	prod(v, *à*)	$prod_T$(v)	autres prépositions	$prod_R$(v, *à*)
comparer	21	21	Ø	1
accrocher	116	146	*dans, sur, au-dessus de...*	0,79
arriver	399	911	*jusqu'à, au bord de, à la fin de...*	0,44
tracer	10	56	*dans, sur, avec ...*	0,18

Tableau 1 : Exemples de valeur de productivité relative pour la préposition *à*

On voit que le verbe *comparer* ne gouverne que des SP en *à*, d'où $prod_R$(*comparer, à*) = 1. Les verbes *accrocher* et *arriver* acceptent d'autres types de rattachement tout en privilégiant – plus ou moins fortement – la préposition *à*, ce qui explique des valeurs de $prod_R$ respectivement haute (0,79) et assez haute (0,44) pour ces verbes. Dans le cas du verbe *tracer*, la préposition *à* n'est pas aussi nettement dominante (0,18).
La mesure de productivité relative permet ainsi de classer les verbes d'un corpus en mettant au jour différents profils de verbes : certains sont exclusivement ou majoritairement associés à une préposition unique (*comparer à, hériter de, jongler avec,* etc.). D'autres entretiennent une relation prédominante avec une préposition mais présentent une valeur de productivité relative non négligeable pour d'autres prépositions. Cette situation recouvre plusieurs cas de figure : dans le cas du verbe *accrocher*, la même relation (cible du mouvement) est exprimée de façon majoritaire à l'aide d'une préposition (*accrocher à*) et plus marginalement avec d'autres (*dans, sur*). Dans d'autres cas, une relation prédominante, de nature typiquement argumentale, peut s'accompagner de la présence de dépendants plus périphériques (*témoigner de* / *devant*).

[2] Ces mesures ont été mises au point avec l'aide de Didier Bourigault, et présentées dans une publication antérieure (Fabre et Bourigault 2008), avec quelques changements de notation.

Enfin, certains verbes présentent une concurrence plus marquée entre plusieurs prépositions, par exemple dans le cas de constructions alternatives (*hésiter entre* / *sur*, *ruisseler de* / *sur*).

La valeur de productivité relative peut également être utilisée pour calculer le degré d'autonomie d'un SP. Chaque SP est ramené à un couple (p, n_D) qui en est le représentant, où p désigne la préposition, n le nom régi par la préposition, D indique la présence ou l'absence d'un déterminant[3]. Ainsi, le couple (*sur*, *fond*$_D$) est le représentant normalisé de tous les SP du corpus introduits par la préposition *sur* et dont la tête nominale est le nom *fond*, muni d'un déterminant. Le SP *sur le fond sablonneux* en est une instance spécifique, de même *sur des fonds qui varient de 5 à 10 mètres*. Ce représentant permet ainsi de rassembler des occurrences de différentes variantes d'un même SP.

L'idée consiste à ne pas considérer seulement le nombre de verbes différents avec lesquels le SP s'associe (productivité du SP, notée *prod*(*p*,*n*$_D$)), mais le degré d'association moyen de ces verbes avec la préposition considérée (valeur moyenne de $prod_R$ des verbes associés). En effet, le nombre de verbes différents peut être élevé, mais si ces verbes ont eux-mêmes tendance à sélectionner fortement la préposition ($prod_R$ haute en moyenne), il ne s'agit pas d'un indice d'autonomie, mais au contraire de cohésion. A titre d'exemple, les SP représentés par le motif lexico-syntaxique (*à*, *volonté*$_D$) se combinent presque exclusivement avec des verbes qui sélectionnent fortement la préposition *à* dans le corpus (*soumettre*, *résister*, *attribuer*, *obéir*, etc. *à la volonté de*) ; au contraire, les SP de la forme (*à*, *horizon*$_D$) apparaissent essentiellement avec des verbes qui ne sélectionnent pas la préposition *à* (*naître*, *s'étendre*, *regarder*, *baisser*, …). Le degré d'autonomie du second est donc plus fort que celui du premier.

La mesure de productivité du SP qui tient compte de la force d'association entre chaque verbe et la préposition s'appelle la productivité pondérée. Notée *prod*$_p$(*p*,*n*$_D$), elle se calcule de la manière suivante :

$$prod_P(p,n_D) = \sum_{\{v \,/\, f(v,p,n) >= s\}} prod_R(v,p)$$

Considérons le cas du SP (*à*, *volonté*$_D$). Il se combine avec 20 verbes, qui présentent dans leur grande majorité une association forte avec la préposition *à*. Le tableau 2 en montre un échantillon.

verbe	$prod_R$(v, *à*)
se heurter	0,98
résister	0,96
soumettre	0,92

Tableau 2 : Exemples de verbes se combinant fortement avec la préposition *à*

La somme des valeurs de productivité relative de ces 20 verbes avec la préposition *à* est de 15,83. Par comparaison, le SP (*à*, *horizon*$_D$), se combine avec 19 verbes qui présentent une valeur de productivité relative bien moindre, en moyenne, avec *à*. Le tableau 3 en montre un échantillon.

verbe	$prod_R$(v, *à*)

[3] On conserve l'information relative à la présence ou l'absence du déterminant sachant le rôle, souvent décrit, du déterminant dans le fonctionnement des groupes prépositionnels (cf. l'opposition entre le programme de sous-classification en intension et en extension dont parle Cadiot (1997) pour la préposition *à*).

fixer	0,37
produire	0,22
réaliser	0,14

Tableau 3 : Exemples de verbes se combinant faiblement avec la préposition *à*

La somme des valeurs de productivité relative de ces 20 verbes avec la préposition *à* est, dans ce cas, de 4,79.

Au final, la mesure d'autonomie du SP donne la valeur moyenne d'association entre le SP et les verbes auxquels l'analyseur l'a rattaché. Autrement dit, plus le rapport entre la productivité relative moyenne des verbes et la productivité du SP est bas, plus forte est l'indication qu'on a affaire à un SP autonome.

$$auton(p,n_D) = 1 - \frac{prod_P(p,n_D)}{prod(p,n_D)}$$

Ce qui donne une valeur d'autonomie basse pour le premier exemple de SP :

$$auton(à, volonté_D) = 0,21$$

et une valeur haute pour le second :

$$auton(à, horizon_D) = 0,75$$

Dans la suite de l'article, nous montrons comment ce type de mesures fournit des indicateurs du fonctionnement de la complémentation verbale dans les corpus que nous avons considérés.

3. Vers l'étude d'un continuum argument / circonstant

Devant l'absence de critère fiable permettant de trancher dans les cas douteux, on observe deux attitudes : le maintien de la distinction arguments d'un côté et circonstants de l'autre (cf. Bonami *op. cit.*) ; un passage graduel entre compléments plus ou moins liés au verbe (cf. Lacheret-Dujour & François *op. cit* ; Vater 1978). La proposition de Hobaek Haff (1992) est originale en ce qu'elle tient les deux points de vue ensemble puisque tout en maintenant la distinction, elle intègre la notion de gradation à l'intérieur de chaque catégorie. De l'observation de l'ordre des mots dans les constructions inversées, l'auteur montre en effet que ni actants ni circonstants ne constituent des catégories homogènes puisque « il existe des actants qui ressemblent à des circonstants et vice versa » (p.289).

Comme on vient de le voir, la valeur d'autonomie permet de classer les SP entre deux pôles opposant ceux qui manifestent une indépendance vis-à-vis du verbe (valeur d'autonomie forte : fonction d'adjoint) à ceux qui entretiennent un lien étroit avec le verbe (valeur d'autonomie faible : fonction de complémentation). Nous proposons ici une synthèse du continuum qui se dégage, représenté figure 2.

Dans les valeurs les plus hautes (*auton* > 0.9), on trouve des SP introduits par des prépositions au sémantisme bien déterminé, marquant différents types de circonstances, et souvent composées (*en début de*, *au milieu de*, *via*, *à travers*, *depuis*, etc.). Lorsqu'on a affaire aux prépositions les plus courantes et traditionnellement considérées comme polysémiques (*à*, *de*,

dans, *sur*, *avec*, *en*), les valeurs hautes d'autonomie correspondent également à des SP circonstanciels : *en Grèce* (0.9), *en euros* (0.9), *en septembre* (0.8), *à l'étage* (0.7), *au Congo* (0.7), etc. On remarque également la présence en grand nombre de deux autres types d'unités construites autour de ces prépositions : des locutions adverbiales ou des locutions prépositionnelles. Ainsi, parmi les SP les plus autonomes associés à la préposition *en*, figurent les séquences *en majorité*, *en contrepartie*, *en finale*, *en politique*, *en tête*, *en réseau*, etc. Leur valeur d'autonomie est supérieure à 0,9. Parmi les SP les plus autonomes associés à la préposition *à*, figurent les séquences *à l'égard* (*de*), *à l'échelle* (*de*), *au profit* (*de*), *au matin* (*de*), *à parité* (*de*). Leur valeur d'autonomie se situe autour de 0,8. Certains des SP très autonomes peuvent relever simultanément des deux cas de figure (*en cours* (*de*), *en tête* (*de*)). La mesure d'autonomie peut donc faciliter le recensement de ces unités complexes, que l'analyseur n'a pas segmentées.

Les valeurs basses d'autonomie correspondent logiquement à des groupes qui dépendent fortement du verbe, comme en atteste l'impossibilité de leur attribuer un sens autonome. C'est le cas de groupes comme *de la capacité*, *de la qualité*, *au progrès*, *au traitement*, etc, dont la valeur d'autonomie est inférieure à 0.2, et qui sont indissociables des verbes auxquels ils se rattachent (par exemple, respectivement, *se doter*, *se plaindre*, *contribuer*, *soumettre*).

Entre ces deux pôles se répartissent les SP qui obtiennent un score d'autonomie médiane (inférieur à 0.7 et supérieur à 0.3). Ces valeurs médianes concernent la majorité des SP des corpus examinés et illustrent deux cas de figure intermédiaires.

Le premier cas concerne les SP qui se situent dans cette zone parce qu'ils sont alternativement arguments ou circonstants selon les verbes auxquels ils s'associent. C'est par exemple le cas du SP *à l'hôtel* (*auton* = 0,6), qui s'associe à la fois avec des verbes comme *dormir* ou *régner* ($prod_R$ avec à < 0,2) et des verbes comme *appartenir* ou *s'adresser* ($prod_R$ avec à > 0,9). On peut parler dans ce cas de SP « hybrides ».

On relève ensuite des SP qui manifestent véritablement une situation intermédiaire de rattachement avec le verbe. Sans participer à la valence du verbe, ils fournissent une indication circonstancielle qui l'accompagne de façon régulière. C'est le cas de nombreux SP introduits par *sur*, *dans*, etc., comme l'illustre ici le SP *sur le drap* (*auton* = 0.5) :

> [3] « elle aperçut d'abord le revolver, abandonné sur le drap, à la place qu'elle occupait. » (Jouve Pierre-Jean, *Paulina 1880*, 1925, p. 242)

En s'intéressant spécifiquement à ce cas de figure, on peut dégager des cas d'affinités fortes entre des séries d'adjoints et des séries de verbes. C'est le cas dans le corpus FRANTEXT20 de SP comme [*avec* + dét + *attention*] et [*avec* + dét + *curiosité*], dont la valeur d'autonomie relativement peu élevée (*auton* = 0.6) s'explique par le fait qu'ils s'associent de manière régulière à une classe de verbes de perception (*considérer*, *contempler*, *dévisager*, *écouter*, *examiner*, *observer*…) :

> [4] « Et moi aussi, je l'examinais avec attention. » (Modiano Patrick, *Les Boulevards de ceinture*, 1972, p.164)
> [5] « C'était pour une sorbonnarde un endroit assez prestigieux et tout en causant j'examinai avec curiosité ce haut lieu. » (Beauvoir Simone de, *Mémoires d'une jeune fille rangée*, 1958, p.243)

Toujours dans FRANTEXT20, on observe le même type de phénomène autour de groupes comme *de plaisir*, *d'orgueil*, *d'indignation*, *de peur*, etc. (valeur d'autonomie autour de 0.6), qui, sans être sous-catégorisés par les verbes auxquels ils s'associent, présentent néanmoins un score d'association relativement élevé ($prod_R$ autour de 0.4 ou 0.5) avec une série d'entre eux (*frémir*, *frissonner*, *trembler*, *rougir*). Ces adjoints présentent de fortes affinités avec les prédicats concernés. Ils peuvent donc être considérés comme des éléments plus ou moins prédictibles de patrons préconstruits (*patterns*).

La figure 2 fournit une représentation schématique de la manière dont les principaux types de SP se répartissent le long de l'échelle donnée par la valeur d'autonomie.

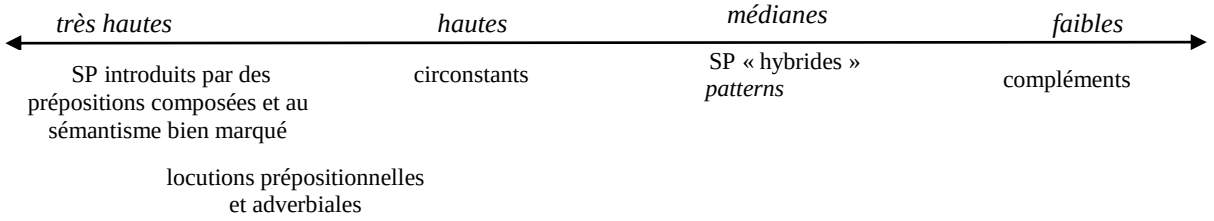

Figure 2 : Représentation du continuum selon les valeurs de la mesure d'autonomie

Un des intérêts de la méthode est donc de mettre au jour une diversité de fonctionnements qui cohabitent dans la zone intermédiaire (valeurs d'autonomie médianes) au sein de laquelle nous avons mis au jour des configurations particulières, ces *patterns* dont nous venons de relever deux exemples :

<verbe de perception> + <complément de manière>
ex : *observer avec attention, examiner avec curiosité*
<verbe de manifestation physique> + <sentiment>
ex : *frissonner de peur, rougir de plaisir*

Bien que nos données soient construites par des méthodes différentes, nous rejoignons ici les travaux qui proposent d'étendre la notion de lexicalisation à des patrons syntaxiques. Ces combinaisons régulières entre lexique et grammaire sont généralement décrites comme des *colligations* (Hunston & Francis 2000, Hoey 2005). En combinant la construction [V + SP] avec des indices statistiques, notre méthode permet de repérer des SP au fonctionnement particulier et, de proche en proche – en examinant la présence de verbes et de valeurs similaires d'association – d'esquisser des configurations.

4. Analyse contrastive : des indicateurs pour l'étude des prépositions

Le fait de travailler sur deux corpus nettement différenciés en termes de genres (littéraire, journalistique), nous offre un deuxième niveau d'observation de la complémentation verbale, en termes cette fois de variation d'un corpus à l'autre (sans prétendre ici en tirer aucune généralité sur une variation systématique au niveau du genre de textes considéré). Nous proposons d'illustrer l'intérêt de cette approche à partir de l'étude du fonctionnement des prépositions. Les calculs que nous avons réalisés nous fournissent en effet des indications quant à la tendance des prépositions à introduire des SP autonomes ou cohésifs. Nous avons ainsi pu calculer le profil des prépositions selon ce critère (valeurs d'autonomie des SP introduits par la préposition) et dégager des observations à deux niveaux. Tout d'abord, ce

profil par préposition est globalement stable d'un corpus à l'autre. Dans les deux corpus, on observe la même répartition pour les prépositions les plus courantes :

- les prépositions *à* et *de* sont celles qui occupent la plus large gamme de valeurs : depuis des compléments très cohésifs (*à l'exigence*, *d'un voile*), jusqu'à des groupes très autonomes (*de front*, *à plat*), en passant par des situations intermédiaires et contrastées (*de bonheur*, *à la maison*).
- les prépositions *en* et *avec* occupent en moyenne les valeurs d'autonomie les plus hautes, formant presque exclusivement des ajouts.
- les SP en *dans* et *sur* occupent principalement des valeurs d'autonomie haute et moyenne.

Dans un deuxième temps, l'examen plus détaillé des SP introduits par les différentes prépositions, et des verbes correspondants, permet de mettre au jour des fonctionnements particuliers qui s'avèrent plus contrastés d'un corpus à l'autre. Nous allons étudier deux types de décalage :

- des différences apparaissent quand on regarde de plus près la répartition des valeurs d'autonomie de certaines prépositions, qui présentent un profil différencié selon les deux corpus, suggérant un usage globalement plus circonstanciel ou plus argumental de la préposition selon le corpus. Par exemple, comme on va le voir, la préposition *comme* introduit des SP en moyenne plus autonomes dans la partie littéraire ; de même, les compléments en *sur* peu autonomes sont plus nombreux dans LM10.
- si le profil est semblable, il s'instancie néanmoins de façon différente d'un corpus à l'autre. C'est le cas des prépositions *avec* ou *dans*, pour lesquelles l'opposition entre SP autonomes et moins autonomes renvoient à un fonctionnement sémantique différent entre les deux corpus.

Nous illustrons ces décalages dans l'usage des SP dans les deux corpus en présentant une brève analyse des quatre prépositions que nous avons mentionnées : *comme*, *sur*, *avec* et *dans*.

La préposition comme

Le tableau 4 montre que *comme*, dans ses empois prépositionnels, s'associe à des verbes qui la sélectionnent plus fortement en moyenne dans le corpus LM10 ($prod_R$ moyenne des verbes avec cette préposition : 0,64) que dans FRANTEXT20 ($prod_R$ moyenne : 0,15).

LM10			FRANTEXT20		
verbe	prod	$prod_R$	verbe	prod	$prod_R$
se considérer	13	1	*considérer*	116	0,64
considérer	529	0,91	*apparaître*	55	0,2
se définir	17	0,81	*devenir*	27	0,19
se affirmer	11	0,73	*agir*	16	0,18
percevoir	79	0,62	*briller*	28	0,16
interpréter	37	0,61	*traiter*	22	0,15
sonner	17	0,59	*aimer*	17	0,13
ressentir	15	0,48	*flotter*	16	0,12
se comporter	12	0,48	*se ouvrir*	13	0,11
désigner	47	0,46	*battre*	13	0,1
apparaître	173	0,39	*éclater*	11	0,09

Tableau 4 : Verbes associés à des SP introduits par la préposition *comme*

Alors que dans FRANTEXT20 la préposition marque presque uniquement la comparaison, et par conséquent une information périphérique pour les verbes concernés (*agir*, *briller*, *flotter*, etc.), dans LM10, on met au jour une série de verbes qui entretiennent une relation nettement plus étroite avec elle. L'examen des SP associés révèle également des différences marquées. On voit émerger de LM10 deux configurations. La préposition *comme* introduit :

- des SP relevant du schéma *comme* + <humain> (*comme directeur*, *comme chef*, *comme candidat*, *comme président*) ;
- des SP présentant une structure locutionnelle *comme* + dét + <abstrait> (de) (*comme le lieu*, *comme une sorte*, *comme un moyen*, *comme un coup*).

Dans FRANTEXT20 dominent les SP relevant du schéma *comme* + dét + <humain, animal, objet> (*comme un chien*, *une fleur*, *une femme*, *un enfant*, *la pierre*). La nette différence de valeur d'autonomie moyenne permet donc de repérer un fonctionnement distinct de la préposition *comme* entre les deux corpus.

La préposition sur

Si la valeur d'autonomie des SP introduits par *sur* est cette fois comparable dans les deux corpus (0,61 dans LM10 et 0,69 dans FRANTEXT20), elle se différencie néanmoins dans les valeurs basses : sur 390 SP introduits par *sur*, un seul présente une valeur d'autonomie inférieure à 0,4 dans *Frantext* (le SP *sur* dét *cas*), alors que cette caractéristique concerne 89 des 498 SP recensés pour le corpus du *Monde*. Cela est lié au fait que plusieurs verbes très productifs sélectionnent fortement la préposition *sur* dans ce corpus. C'est le cas par exemple des verbes *s'appuyer*, *s'interroger*, *déboucher*, *insister*, *reposer* ($prod_R$ > 0,9 et *prod* > 100).

Si l'on examine précisément les SP introduit par *sur* dans FRANTEXT20, on constate que la gamme des valeurs d'autonomie oppose :

- des SP désignant un support[4], un lieu où s'exerce un contact : *sur le poignet*, *sur la rampe*. Ces SP sont introduits par des verbes dont la $prod_R$ est relativement élevée – autour de 0,6 ou 0,7 – comme *poser* ou *appuyer* ;
- des SP désignant les lieux d'une activité (*sur la digue*, *sur la côte*). Les verbes associés n'entretiennent pas de lien privilégié avec la préposition *sur*, et peuvent se combiner avec d'autres prépositions spatiales (ex : *rester*, *mettre*).

Dans LM10, l'opposition concerne :

- des SP non spatiaux, comprenant des noms abstraits (*sur la création*, *sur la réalité*, *sur la capacité*), liés aux verbes que nous avons signalés plus haut ;
- des SP exprimant très majoritairement une information spatiale (*sur la scène*, *sur le continent*) et quelques SP exprimant d'autres circonstances (*sur un ton*, *sur ordre*).

Dans le cas de cette préposition, la mesure d'autonomie permet donc d'organiser les SP en identifiant dans chaque corpus les catégories sémantiques associées aux différentes valeurs d'autonomie. On voit que la différence porte ici sur la façon dont les valeurs argumentales des SP en *sur* s'instancient dans les deux corpus.

La préposition avec

Tout comme *sur*, le profil général de la préposition *avec* est similaire dans les deux corpus : avec une valeur d'autonomie moyenne des SP de 0,8 dans les deux cas, on a affaire à une préposition qui introduit généralement des SP très autonomes. Il est intéressant de regarder de plus près les cas de SP manifestant une plus faible autonomie et occupant des valeurs médianes dans les deux corpus : la valeur d'autonomie la plus basse est de 0,4 dans *Le Monde*, et de 0,53 dans *Frantext*. Là encore, cette stabilité recouvre des différences dans le

[4] Nous nous appuyons ici sur les catégories sémantiques employées dans le *Trésor de la Langue Française*.

fonctionnement sémantique de la préposition, que l'on peut observer en considérant la liste des verbes qui, dans les deux corpus, présentent une valeur de productivité relative haute pour *avec* (nous avons relevé pour chaque corpus les 10 premiers verbes classés par ordre décroissant de productivité relative, et dont la productivité brute est supérieure à 15) :
LM10 : *prendre contact, flirter, renouer, coïncider, confondre, rompre, dialoguer, s'entretenir, fusionner, négocier*
FRANTEXT20 : *contraster, dévisager, confondre, se confondre, rompre, contempler, écouter, accueillir, examiner, se dire*
Ces verbes permettent d'expliquer que les valeurs les moins hautes d'autonomie correspondent principalement à l'emploi de *avec* marquant une relation entre individus ou collectifs humains dans le corpus LM10 (*administration, commission, institution, ennemi, Israël*), correspondant à la valeur comitative de la préposition (Cadiot 1997). Dans FRANTEXT20, on a plutôt affaire à des SP exprimant le sentiment ou plus généralement une disposition (*méfiance, insolence, stupeur, attention, curiosité*). On retrouve ici le schéma décrit dans la section précédente (schéma <verbe de perception> + <manière>), lequel n'apparaît pas dans le corpus du *Monde*. Ainsi, le verbe *examiner* a une valeur de $prod_R$ de 0,13 dans LM10, de 0,6 dans FRANTEXT20 ; de même, le SP *avec attention* a une valeur d'autonomie de 0,89 dans LM10, et de 0,63 dans *v*. Si l'on considère cette fois les valeurs hautes, la valeur de manière prédomine dans LM10 (*avec précaution* ou *avec* dét *rigueur*), la valeur instrumentale apparaissant plus nettement dans FRANTEXT20 (*avec* dét *bâton*, *avec* dét *pierre*).

La préposition dans
La préposition *dans* présente elle aussi des valeurs d'autonomie moyenne comparable entre les deux corpus : dans FRANTEXT20, l'autonomie moyenne des SP qu'elle introduit est de 0,69 (valeurs oscillant entre 0,16 et 0,91), elle est de 0,67 dans LM10 (valeurs entre 0,23 et 0,94). La différence de fonctionnement entre les deux corpus tient à nouveau à la façon dont ces fluctuations se traduisent sur le plan sémantique. La variation des valeurs d'autonomie est facile à interpréter dans le cas du corpus *Frantext*. En effet, les cas de SP peu autonomes introduits par la préposition *dans* (valeurs d'autonomie équivalentes ou inférieures à 0,6) marquent principalement l'intériorité dans un espace clos – concret ou figuré (*dans le labyrinthe, le gouffre, le flot, la profondeur, la contemplation*). Les verbes associés qui ont une productivité relative haute avec cette préposition sont par exemple *s'enfoncer, sombrer, plonger, entrer*, etc. On retrouve dans les valeurs d'autonomie basses quelques instances du schéma *dans* + $N_{émotion}$ décrit par Vaguer (2005) comme un cas de complément obligatoire (avec les SP *dans le mutisme, dans la contemplation, dans les larmes*). Ceux qui se situent dans le pôle d'autonomie haute ne marquent généralement pas de localisation spatiale mais expriment la temporalité, la manière (*dans la matinée, dans la jeunesse, dans le style* – valeur d'autonomie $\geq$ 0,85). On trouve entre les deux des noms qui dénotent des lieux, sans que le rapport d'intériorité soit particulièrement marqué. Ces SP (*dans le château* – *auton* = 0,69, *dans le métro* – *auton* = 0,68, etc.) se combinent avec des verbes dont la valeur de $prod_R$ est non négligeable avec *dans* (*travailler, naître, rencontrer, installer*) – autrement dit, ces verbes s'accompagnent assez régulièrement d'une information spatiale. Le mot *nuit* illustre le cas d'un nom dont la sémantique est compatible avec ce double fonctionnement – *auton*(*dans*, $nuit_D$)= 0,74. Ainsi, *dans la nuit* peut donner lieu à la fois à une interprétation spatiale (la nuit est alors considérée comme un espace, un milieu), et temporelle, comme l'ont montré A. Le Draoulec et D. Vigier (*à paraître*). On retrouve de fait l'association de ce SP avec des verbes qui sélectionnent fortement *dans* (*sombrer, plonger*), et des verbes qui se combinent avec une information temporelle (*naître, survenir*).

Dans le corpus LM10, la préposition *dans* a un comportement général similaire, mais il s'instancie dans un lexique assez nettement différent. C'est particulièrement le cas des SP les moins autonomes : si quelques SP marquant l'intériorité apparaissent dans cette zone (*dans les méandres, dans le périmètre*), les plus productifs relèvent d'un autre sémantisme (*dans la création, la crise, la culture, la production*), et sont associés à un tout autre type de vocabulaire verbal (*impliquer, spécialiser, se lancer, investir*, etc.). Si l'on retrouve d'un corpus à l'autre la même opposition graduelle entre des compléments en *dans* cohésifs et autonomes, elle ne s'instancie donc pas dans les mêmes constructions.

5. Conclusion

Nous espérons avoir montré ici ce que le recours à de vastes quantités de données peut apporter à la description d'une distinction syntaxique aussi discutée que la distinction complément/ajout. A cette opposition binaire, la méthode proposée permet de substituer une représentation continue des usages observés dans un vaste corpus grâce à des indices statistiques calculés à partir de l'exploitation d'informations syntaxiques produites par un analyseur automatique. Le travail autour de la mesure d'autonomie fournit en effet une mesure graduelle qui permet d'organiser les SP d'un corpus, de les regrouper et d'observer des situations intermédiaires entre arguments et circonstants, comme nous l'avons illustré dans le cas de deux *patterns* particuliers, qui manifestent « un comportement de cooptation, de préférence mesurée statistiquement à partir de corpus, sans (…) qu'il soit redevable à quelque principe structural » (Legallois et François, 2006). La masse d'informations qui se trouve synthétisée à partir de deux corpus permet par ailleurs de dégager des tendances générales et des usages différenciés d'un corpus à l'autre. Elle permet de mettre au jour des associations sémantiques qui peuvent alimenter les travaux descriptifs sur les prépositions, tels que nous avons commencé à les esquisser pour quatre prépositions courantes. Une cartographie plus systématique resterait à faire, qui permettrait par exemple de comparer la façon dont l'information circonstancielle s'exprime d'un corpus à l'autre. Par ailleurs, il faudrait travailler au regroupement sémantique des SP de manière à limiter l'éparpillement des observations sur des cas isolés. Enfin, si nous avons choisi de privilégier une structure et une mesure statistique simples, ce type d'étude gagnerait à combiner d'autres critères d'observation, et en particulier à travailler sur des schémas syntaxiques plus complets ou à disposer d'informations sur le degré d'optionalité des SP. Au total, nous avons présenté dans cet article une approche de l'usage plutôt située dans la lignée de la linguistique *corpus-driven*, même si elle prend en compte des éléments d'annotation préalables. Cependant, d'autres approches se réclament désormais de la linguistique de l'usage, notamment celle défendue par J. Bybee dans le cadre d'une « usage-based theory ». Bien que le contexte théorique dans lequel cette approche se développe soit très différent de celui que nous adoptons dans nos propres travaux, il pourrait être fructueux de comparer nos observations notamment avec la notion de construction, car comme le dit Bybee (2010, 78) : « constructions are particularly appropriate for exemplar models, as they are surface based and can emerge from categorization of experienced utterances. »

NOTES

1. Nous remercions Jean-Marie Pierrel d'avoir mis ce corpus à notre disposition dans le cadre d'une convention passée entre l'ATILF et CLLE-ERSS.

2. Ces mesures ont été mises au point avec l'aide de Didier Bourigault, et présentées dans une publication antérieure (Fabre & Bourigault, 2008), avec quelques changements de notation.

3. On conserve l'information relative à la présence ou l'absence du déterminant sachant le rôle, souvent décrit, du déterminant dans le fonctionnement des groupes prépositionnels (cf.

l’opposition entre le programme de sous-classification en intension et en extension dont parle Cadiot (1997) pour la préposition à).

4. Nous nous appuyons ici sur les catégories sémantiques employées dans le *Trésor de la Langue Française*.

BIBLIOGRAPHIE

Baayen, R. (2009). Corpus linguistics in morphology: Morphological productivity, *Corpus Linguistics. An International Handbook* 2, 899-919.

Barodal, J. (2006). Predicting the productivity of argument structure constructions. In *Proceedings of the 32nd Annual Meeting of the Berkeley Linguistics Society*.

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

Borillo, A. (1990). À propos de la localisation spatiale. *Langue française*, 86 (1) : 75-84.

Bourigault, D. (2007). *Un analyseur syntaxique opérationnel : SYNTEX*. Thèse d'habilitation à diriger des recherches, Université Toulouse 2-Le Mirail.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press.

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

Fabre, C., Rebeyrolle, J. & Ho-Dac, L.-M. (2008). Examen du statut des syntagmes prépositionnels à la lumière de données issues de corpus annotés. In J. Durand, B. Habert & B. Laks (éds.) *Congrès Mondial de Linguistique Française*, Paris, Institut de Linguistique Française.

Fabre, C. & Bourigault, D. (2008). Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants. *Journal of French Language Studies*, 18 (1): 87-102.

Hobaek Haff, M. (1992). Actants circonstanciels et circonstants actanciels – une analyse de ka dichotomie actant/circonstant. *Revue romane*, 27 (2) : 286-290.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Hunston, S & Francis, G. (2000). *Pattern grammar:* a corpus-driven approach to the lexical grammar of English. John Benjamins Publishing Company.

Lacheret-Dujour, A. & François, J. (2004). Circonstance et prédication en français parlé : contraintes sémantico-pragmatiques et filtrage prosodique. *Revue de syntaxe et sémantique : la prédication verbale et averbale dans les langues*, 6 : 35-56.

Le Draoulec, A. & Vigier, D. (*à paraître*). Dans suivi d'un nom de partie de la journée : au croisement de l'espace et du temps. *Revue de Sémantique et Pragmatique*.

Legallois, D. et François, J. (dir). (2006). Autour des grammaires de constructions et de patterns, *Cahier du CRISCO,* n°21.

Manning, C. (2003). *Probabilistic syntax*. In R. Bod et al. (eds) *Probabilistic linguistics*. Cambridge MA : MIT Press, 289-341.

Miller, P. (1998). Compléments et circonstants : une distinction syntaxique ou sémantique ? In J.-C. Souesme (dir.). *Cycos 15, actes du 37e congrès de la SAES (Société des anglicistes de l'enseignement supérieur*), 91-103. Nice : Presses universitaires de Nice.

Tognini-Bonelli, E. (2001). *Corpus linguistic at work*. Amsterdam/Philadelphia: John Benjamins.

Tutunjian, D. & Boland, J. (2008). Do we need a distinction between Arguments and Adjuncts? Evidence from Psycholinguistic Studies of Comprehension. *Language and Linguistics Compass*, 2 (4): 631-646.

Vaguer, C. (2005). Pourquoi sombre-t-on dans le malheur ? Étude de constructions verbales V dans Némotion. *Lidil*. *Revue de linguistique et de didactique des langues* (32).

Vater, H. (1978). On the possibility of distinguishing between complements and adjuncts. In W. Abraham (ed.) *Valence, semantic case and grammatical relations* (21-46). Amsterdam: John Benjamins.