Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf08227

Examen du statut des syntagmes prépositionnels à la lumière de données issues de corpus annotés

Cécile Fabre, Josette Rebeyrolle et Lydia-Mai Ho-Dac

CLLE (UMR 5263) Université de Toulouse & CNRS
{cecile.fabre, josette.rebeyrolle, lydia-mai.ho-dac}@univ-tlse2.fr

1 Introduction

Selon les cadres théoriques, les constituants dotés d'une autonomie syntaxique sont dénommés "éléments périphériques", "adjoints", "ajouts", "circonstants". Comme le note Combettes (2007), si on parle d'ajout c'est parce que la "phrase présente toutes les caractéristiques d'un énoncé complet, syntaxiquement et sémantiquement". En (1), seul le complément direct est syntaxiquement et sémantiquement obligatoire (*écrire quelque chose*), les autres compléments sont des syntagmes prépositionnels (SP) moins centraux : la localisation (*dans*), l'instrument (*avec*) gravitent à la périphérie (pour reprendre une formule de Maurel 1990, p.30) :

(1) Paul écrivait ton nom dans la pierre avec son canif.

On sait que les SP qui se trouvent en position postverbale entretiennent des rapports divers avec le verbe. Entre les deux pôles clairement délimités opposant des syntagmes qui, d'un côté, s'insèrent dans la valence du verbe et, de l'autre, apportent indépendamment du verbe une contribution sémantique propre, se nouent entre le verbe et le SP des relations plus ou moins étroites sur lesquelles il est difficile de statuer en employant les tests grammaticaux usuels. Dans cet article, nous voudrions montrer que le recours à des caractérisations quantitatives sur un gros volume de données permet d'apporter de nouveaux éléments pour guider cette caractérisation. Notre objectif est donc de proposer et tester des indices pouvant être repérés automatiquement et permettant de distinguer différents degrés d'autonomie des SP. Pour atteindre ce but, il faudra expliquer comment s'opère le passage de tests linguistiques à des calculs sur de gros volumes de données issus de corpus syntaxiquement annotés. Nous avons ainsi opté pour une démarche en deux temps : 1) déterminer, parmi les tests existants, lesquels peuvent être automatisés de manière à faire émerger certains comportements syntaxiques des SP à partir d'un gros volume de données textuelles ; 2) au-delà, mettre au jour de nouveaux tests, qui ne peuvent être conçus dans le cadre d'une analyse manuelle guidée par l'introspection, mais que des techniques de linguistique de corpus permettent désormais de mettre en place[1].

Notre présentation ne se situant pas dans un cadre théorique particulier, nous utilisons les termes de "complément" et d' "ajout" pour désigner les deux types d'attachement prototypiques par rapport au verbe. Parmi les constituants qui peuvent remplir la fonction de complément et/ou d'ajout, nous nous intéressons particulièrement au cas des syntagmes prépositionnels qui gouvernent un nom. À ce stade, nous ne nous limitons pas à un sous-ensemble de prépositions, spatiales ou temporelles, mais nous cherchons à examiner la diversité. Toutefois, nous envisageons uniquement les relations qu'entretiennent ces syntagmes avec le verbe et laissons donc de côté les cas où les SP dépendent d'un adjectif, d'un groupe nominal ou d'un adverbe.

Nous commençons en section 2 par rappeler brièvement les indices habituellement convoqués pour distinguer les compléments des ajouts, ce qui nous conduira à retenir le placement comme critère exploitable et à mener une première expérience d'extraction sur corpus de SP situés en position préverbale. En section 3, nous proposons de nouveaux critères issus d'analyses quantitatives mettant en avant quelques propriétés différentielles des compléments et des ajouts. Le critère que nous modélisons dans notre analyse quantitative mesure le degré d'autonomie par rapport au verbe auquel il se rattache.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf08227

2 Adapter les tests usuels pour une exploration en corpus

2.1 Le placement des syntagmes prépositionnels

Pour dégager des propriétés différentielles des compléments et des ajouts, on a habituellement recours à une batterie de tests qui évaluent le degré de cohésion du syntagme prépositionnel avec le verbe. Ces tests sont autant de moyens syntaxiques de faire apparaître le relâchement du lien qui unit le SP circonstanciel au verbe et donc son caractère non essentiel *i.e.* périphérique. Le test de la suppression est le plus fréquemment utilisé. Il vise à estimer le caractère facultatif du SP. Cependant, de nombreux travaux ont montré que ce test n'a pas de véritable valeur discriminante – comme le montre la possibilité d'omettre les compléments d'objet. Bonami (1999) montre en particulier que la seule conclusion positive que l'on peut tirer est la suivante : si le SP est syntaxiquement obligatoire, alors il s'agit d'un argument. Mais cette règle ne résiste pas non plus si l'on considère, à la suite de Goldberg et Ackerman (2001), le cas des « ajouts obligatoires » (*obligatory adjuncts*) requis pour satisfaire certaines contraintes de niveau pragmatique.

Parmi les autres critères utilisés pour évaluer le caractère périphérique d'un SP, rappelons notamment le détachement en position post-verbale, la possibilité de lier le SP ainsi détaché par une coordination, le remplacement du verbe par *le faire* tout en maintenant le complément. L'application de tels tests reste cependant difficile : leurs limites ont été bien démontrées (Borillo, 1990 ; Rémi-Giraud, 1998).

Une propriété semble cependant permettre de tirer des conclusions plus précises : il s'agit de la mobilité, c'est-à-dire de la capacité des SP à occuper différentes positions dans la phrase. Comme le rappelle Bonami (ibid), le statut de complément ou d'ajout contraint en partie les positions occupées par les SP dans la phrase. Alors que les compléments n'ont pas de position réservée, les ajouts occupent certaines positions qui leur sont propres. Ils sont en effet les seuls à pouvoir se trouver entre le sujet et le verbe fini (2) et à venir se loger entre le verbe fini et le participe passé (3) :

(2) Agnès, avec angoisse, a regardé le sol.

(3) Agnès a, avec angoisse, regardé le sol.

On observe donc de la part des ajouts la capacité à occuper une grande diversité de positions dans la phrase. Outre ces deux positions, la position initiale, parce qu'elle place le constituant hors du champ verbal, est elle aussi généralement réservée aux ajouts (bien que la position en tête de phrase ne soit pas interdite aux compléments à partir du moment où le syntagme est topicalisé comme dans *Au plafond, des guirlandes pendaient*) :

(4) Avec angoisse, Agnès a regardé le sol.

De fait, des travaux récents (Ho-Dac et Péry-Woodley, 2008) montrent que la position initiale peut être considérée comme un marqueur qui confère aux éléments qui y apparaissent un fonctionnement essentiellement discursif, ce qui se traduit par une faible intégration syntaxique (Charolles, 1997). La position initiale des SP peut alors constituer un indice de forte autonomie vis-à-vis du verbe.

Enfin, les SP ajouts peuvent également apparaître en zone postverbale, soit avant les compléments (5), soit après tous les compléments verbaux (6).

(5) Agnès a regardé, avec angoisse, le sol.

(6) Agnès a regardé le sol, avec angoisse.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf08227

Les SP ajouts peuvent au total occuper toutes les positions suivantes :

P1 Sujet **P2** { Aux **P3** Verbe Ppé } **P4** objet **P5**

Les positions ainsi définies peuvent donc être en partie corrélées au type de relation existant entre le SP et le verbe. Nous pouvons alors distinguer trois situations :

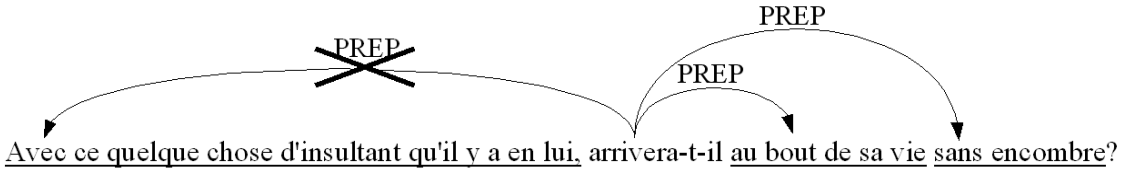
- les SP compléments qui sous-catégorisent le verbe ne peuvent apparaître en P2 et P3 ;
- les SP ajouts qui tombent sous la portée du verbe peuvent apparaître dans toutes les positions P1 à P5 ;
- les SP ajouts qui ne tombent pas sous la portée du verbe et qui étendent leur application à l'ensemble de la proposition - voire la prolongent aux phrases qui suivent - se placent de manière privilégiée en initiale (P1) ou en finale (P5) où ils sont généralement détachés par une virgule du reste de la phrase.

Nous proposons dans ce qui suit d'utiliser ce critère de positionnement pour réaliser une extraction des SP d'un corpus, à l'aide de procédures automatiques. Nous faisons dans un premier temps l'hypothèse qu'il est possible de mettre au jour les SP ajouts d'après leur localisation dans des positions qui leur sont strictement ou généralement réservées.

2.2 Extraction des SP en position préverbale

Précisons tout d'abord la nature du corpus sur lequel nous nous appuyons. Le corpus *Romans XXème* utilisé dans cette expérience est issu de la base *Frantext*[2]. Il comprend 520 romans français du XXème siècle. Sa taille est de 30 millions de mots. Notre démarche consiste à nous appuyer sur les sorties d'instruments d'analyse automatique de manière à tirer parti de traitements réalisés en amont – étiquetage catégoriel, repérage de constituants – pour constituer les observables de notre étude (cf. Habert, 2005). Nous réalisons des extractions et des calculs en sortie d'une annotation syntaxique réalisée automatiquement par l'analyseur SYNTEX (Bourigault, 2007). SYNTEX fournit une analyse robuste, éventuellement partielle, des phrases du corpus. Le module de rattachement prépositionnel, sur les résultats duquel nous nous appuyons, procède de la manière suivante : SYNTEX recherche les recteurs potentiels de chaque préposition en parcourant de droite à gauche la séquence à analyser, jusqu'à rencontrer certaines configurations d'arrêt (présence d'un verbe, d'un pronom relatif…). Lorsque plusieurs candidats recteurs sont identifiés, l'analyse cherche à les départager en exploitant des indices endogènes (fréquence des différentes configurations dans le corpus) ou exogènes (probabilités de sous-catégorisation recensées dans des lexiques).

En revanche, seuls les SP postverbaux (non détachés par une virgule) sont rattachés : les SP situés avant le verbe et pour lesquels aucun recteur nominal ou adjectival n'a été identifié restent "flottants", ce qui évite de rattacher au verbe des groupes qui n'entrent pas dans sa portée. Il en va de même pour les SP postverbaux détachés par une virgule en incise ou en finale.



Comme on le voit dans l'exemple, tout SP post-verbal est rattaché automatiquement au verbe. La section en cours porte uniquement sur les SP non rattachés par l'analyseur, et nous nous focalisons sur ceux situés dans les zones P1 et P2[3] que nous avons définies précédemment, de manière à mettre au jour des SP mobiles, capables de s'émanciper de la position postverbale.

CMLF2008
Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf08227

Nous avons ainsi extrait :

- **les SP en position initiale** (P1)

Il s'agit de tous les SP trouvés entre le début de la séquence et le sujet du verbe. Voici quelques exemples (les SP extraits sont soulignés) :

(7) Dans la petite pièce voisine, Yves ne dormait pas.

(8) Le dimanche, en promenade, au café, on ne les voyait jamais l'un sans l'autre.

- **les SP situés en incise entre le sujet et le verbe** (P2)

Les SP que nous extrayons sont entourés d'une virgule. Par exemple :

(9) La route sous nos pas, par ce matin de gel, sonnait dure et légère.

(10) Gaspard, d'un coup de chapeau, avait tué la chandelle.

Nous avons réalisé ces deux séries d'extractions à partir du corpus *Romans XXème*, et abouti à deux listes de SP. Il est important à ce stade de pouvoir rapprocher des occurrences qui sont des instances différentes d'un même type de SP. Nous ramenons ainsi tous les SP à un couple (p, n_{DET}) qui en est le représentant. p désigne la préposition, n_{DET} le nom régi par la préposition assorti d'une information concernant son caractère déterminé ou non déterminé (DET vaut D ou _). Ainsi :

- le couple (*dans, pièce*$_D$) est le représentant normalisé de tous les SP du corpus introduits par la préposition *dans* et dont la tête nominale est le nom *pièce* quand celui-ci est déterminé. Le SP *dans la petite pièce voisine* en est une instance spécifique, de même *dans la pièce*, *dans cette pièce froidie* ou *dans la pièce unique qui occupait tout l'espace entre les quatre murs*.
- le couple (*à*, *coup* _) est le représentant normalisé de tous les SP du corpus introduits par la préposition *à* et dont la tête nominale est le nom *coup* quand celui-ci n'est pas déterminé. Les SP comme *à grands coups de pieds*, *à coups de pioche*, *à coups de mouchoir* en sont des instances spécifiques[4].

Le représentant permet ainsi de rassembler des occurrences de SP qui relèvent d'une même structure. Notons que cette procédure peut cependant rapprocher accidentellement des instances que l'on souhaiterait distinguer. Ainsi, le couple (*à*, *coup* _) est-il également instancié par le groupe adverbial *à coup sûr*.

Après extraction et normalisation, 2 948 couples différents comportant au moins deux occurrences ont été extraits en position initiale, 1 200 couples différents comportant également au moins deux occurrences ont été extraits en incise entre le sujet et le verbe. Le tableau 1 donne des exemples de couples extraits en position initiale.

p	*n*	*dét*	Fréq	Exemple
à	printemps	D	16	Au printemps 44 il était…
de	manière	D	16	De cette manière, il atteignit…
depuis	an	D	16	Est-ce que, depuis un an, je ne mène pas…
dans	village	D	16	Dans les villages que nous traversions, les paysans se tenaient…
dans	noir	D	15	Par-dessus son image, dans le noir du miroir, une autre image venait de se former.

Tableau 1 : Exemples de SP extraits en position initiale (P1)

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf08227

Le tableau 2 donne des exemples de couples extraits en position d'incise entre le sujet et le verbe.

p	*n*	*dét*	Fréq	Exemple
sur	gauche	D	6	Un trou noir discret, sur la gauche, marque…
sous	cheveu	D	6	Derrière le bar, une grosse femme à la figure satisfaite et joviale, sous d'abondants cheveux gris, versait à boire…
sans	mot	-	6	J'avais remercié cette vieille dame qui, sans mot dire et cassée en deux, était repartie…
à	comptoir	D	6	Le patron, à son comptoir, frottait le même verre depuis quelques minutes.
dans	quartier	D	5	Les réseaux sur lesquels s'appuyaient les terroristes, dans les quartiers nord, restaient…

Tableau 2 : Exemples de SP extraits en incise entre le sujet et le verbe (P2)

Le tableau 2 illustre certaines précautions à prendre pour exploiter les résultats produits automatiquement : la méthode peut ramener des locutions plutôt que des SP, comme c'est le cas du couple (*sans*, *mot_*), systématiquement instancié par la locution *sans mot dire*. L'étiquetage réalisé en amont de SYNTEX[5] n'ayant pas permis d'isoler ce groupe, l'analyseur le traite comme une structure à tête nominale. Par ailleurs, le SP apparaissant dans cette position intercalée entre le sujet et le verbe, que SYNTEX laisse "flotter" sans émettre d'hypothèse de rattachement, est parfois en fait un constituant du sujet. C'est le cas de l'exemple indiqué pour le couple (*sous*, *cheveu*$_D$). Nous ne disposons pas pour l'instant d'information permettant de filtrer automatiquement ces cas.

Cette expérience montre comment certains tests syntaxiques peuvent être adaptés pour permettre l'extraction à partir de corpus de configurations exhibant des SP qui présentent par rapport au verbe un fort degré d'autonomie. Le test de mobilité est converti en une estimation sur le corpus de la capacité d'un SP à apparaître en position préverbale. Cependant, ces tests portant sur le positionnement des SP présentent des limites qu'il nous faut dépasser. Tout d'abord, ils portent sur des configurations relativement rares et permettent l'observation d'un nombre limité de types de SP ; ensuite, ils ne disent rien, négativement, des SP qui n'ont pas été trouvés dans ces positions – on sait que de la non apparition d'une structure en corpus on ne peut pas conclure à son impossibilité ; mais surtout, ces tests de nature binaire ne permettent pas de rendre compte de la nature graduelle de la propriété d'autonomie d'un SP. Cette approche ne permet donc pas d'avancer dans la caractérisation des cas intermédiaires sur lesquels les tests usuels n'ont pas de prise. Il s'agit donc dans un deuxième temps de passer à une description quantifiée et probabiliste plutôt que catégorique des phénomènes syntaxiques dans la lignée des propositions de (Manning, 2003). Une première piste consisterait à systématiser cette démarche pour aller dans le sens des hypothèses de Hoey (2005) concernant la propension d'un mot – ou plutôt, dans notre cas, d'une structure ancrée lexicalement – à occuper des positions grammaticales ou textuelles spécifiques, cette propension pouvant être estimée à partir de l'observation cumulée de ses occurrences. Il s'agirait alors de calculer pour un SP donné sa propension à apparaître plutôt dans les positions que nous venons d'explorer, ou plutôt dans les positions postverbales. Le passage à ce type de mesure permettrait de sortir d'une caractérisation binaire de la propriété d'autonomie d'un SP. Dans cette perspective, nous avons opté pour une seconde piste de travail, qui consiste cette fois à s'émanciper des critères de caractérisation usuels pour exploiter de manière plus systématique la masse d'informations distributionnelles que fournit un grand corpus annoté.

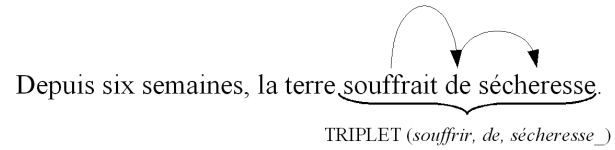
Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf08227

3 Concevoir de nouveaux indices

Dans ce qui précède, nous avons cherché à tester la possibilité d'appliquer les tests existants. Nous montrons à présent que la disponibilité de données linguistiques numérisées qui ont la double caractéristique d'être volumineuses et d'être enrichies d'une couche d'annotation linguistique fournit de nouveaux indices pour caractériser le statut d'un SP.

Nous nous intéressons cette fois aux SP apparaissant en position postverbale, et plus précisément à tous ceux que l'analyseur a rattachés au verbe : nous extrayons tous les triplets (*v,p,n*) apparaissant dans des phrases où le nom *n* est régi par la préposition *p*, et la préposition *p* est régie par le verbe *v*. Par exemple, à partir de la phrase suivante est extrait le triplet (*souffrir*, *de*, *sécheresse*).



Comme dans les expériences précédentes, on ne retient que les triplets dont le nombre d'occurrences dans le corpus est au moins égal à 2. Etant donnée la simplicité de la configuration que nous visons, nous recueillons des données en nombre beaucoup plus important que dans l'approche précédente, ce qui est crucial pour l'approche quantifiée que nous mettons en œuvre. Nous obtenons 94 843 triplets différents (près de 21 000 couples (p, n_{DET}) comparé aux 4 000 couples produits précédemment) à partir du corpus *Romans XXème*. La mesure que nous calculons permet d'ordonner les SP selon leur degré d'autonomie par rapport au verbe auquel ils se rattachent. Pour cela, nous commençons par calculer des informations relatives à l'association entre le verbe et la préposition, avant de nous intéresser à l'association entre la préposition et le nom. C'est en effet en combinant des informations relatives au potentiel combinatoire de ces deux composantes du triplet (*v,p,n*) que nous estimons la propension du SP à l'autonomie vis-à-vis du verbe.

3.1 Degré de sélection d'une préposition par un verbe

Il s'agit tout d'abord de définir la propension d'un verbe à s'associer à une préposition dans le corpus considéré. Pour calculer cette force d'association, nous nous appuyons sur la notion de *productivité*. Celle-ci permet d'apprécier le "rendement de[s] configurations syntaxiques" (Legallois, 2005) et a été intégrée par Didier Bourigault dans les procédures d'analyse syntaxique automatique (Bourigault, 2007).

Nous calculons la productivité du couple (*v,p*) qui correspond au nombre de dépendants nominaux différents qui donnent lieu à un triplet (*v,p,n*).

$$\text{prod}(v,p) = \text{Card}\{n \,/\, \text{f}(v,p,n) >= 2\}$$

C'est un premier indice de la régularité de l'association entre le verbe et la préposition. Ainsi, le couple (*plonger*, *dans*) a une productivité de 185 dans le corpus (il y a 185 noms différents gouvernés par le couple *plonger dans*), alors que le couple (*songer*, *dans*) a une productivité de 6 (seuls 6 noms différents sont recensés en position postverbale du couple *songer dans*).

Le degré de sélection de la préposition par le verbe est ensuite obtenu en calculant la productivité relative du couple (*v,p*) : elle est obtenue en divisant la productivité de (*v,p*) par la productivité *totale* du verbe $\text{prod}_T(v)$, qui est la somme de ses productivités pour toutes les prépositions avec lesquelles il se construit.

$$\text{prod}_T(v) = \sum_p \text{prod}(v,p)$$

$$\text{selec}(v,p) = \text{prod}(v,p)/\text{prod}_T(v)$$

Cela permet de mesurer la contribution de la préposition à l'ensemble des rattachements prépositionnels attribués au verbe. Par exemple, les couples (*insérer*, *dans*) et (*songer*, *dans*) ont une même valeur de

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf08227

productivité (valeur : 6). Mais la productivité totale des deux verbes diffère. Le verbe *insérer* affiche une productivité de 7, *i.e.* le verbe *insérer* apparaît au côté de 7 prépositions différentes (seule la préposition *entre* donne également lieu à un triplet (*insérer,p,n*)), alors que celle du verbe *songer* est de 258 (qui se construit majoritairement avec la préposition *à*). La productivité relative de (*insérer*, *dans*) est donc haute (0,86), celle de (*songer*, *dans*) est très basse (0,02).

En observant les valeurs de sélection calculées sur une série de verbes du corpus on peut repérer quatre profils de verbes :

1) le verbe *v* entretient une relation exclusive avec une préposition *p*, autrement dit prod(*v,p*) = $\text{prod}_T(v)$ (ex : *consacrer à*, *s'intéresser à*, *dépouiller de*, *se transformer en*) ;
2) le verbe admet des rattachements avec plusieurs prépositions, mais l'une d'elle domine très largement (selec(v,p) > 0,7) (ex : *refléter dans*, *témoigner de*, *se méprendre sur*) ;
3) la préséance d'une des prépositions existe mais elle est moins nette (ex : *hésiter entre* (0,38), *répéter avec* (0,38), *sécher sur* (0,47), *se traîner sur* (0,33)) ;
4) il y a concurrence entre plusieurs prépositions que le verbe sélectionne à des degrés similaires (ex : *cracher sur/dans/à*, *ruisseler de/sur*, *sauter de/dans/à, se tromper de/sur*).

Ces éléments fournissent des indications sur la capacité du verbe à s'adjoindre un éventail plus ou moins large de prépositions, et à s'associer plus ou moins exclusivement à l'une d'entre elles. Les cas (1) et (2) renvoient assez clairement à des configurations de sous-catégorisation, les cas (3) et (4), en revanche, qui présentent des situations de concurrence entre plusieurs prépositions pour un même verbe, relèvent plutôt d'associations grammaticales préférentielles.

3.2 Degré d'autonomie d'un SP

Pour calculer le degré d'autonomie du SP par rapport au verbe, on ne se contente pas de l'indice précédent mais on calcule un second indice permettant de tenir compte de l'aptitude d'un même SP, plus précisément d'un constituant ayant pour tête la même préposition et le même nom, le déterminant quant à lui pouvant varier, à s'associer à une grande diversité de verbes. Cette aptitude est, selon nous, un premier signe de l'autonomie du syntagme comme le montre le fait que le couple (*sur, table*$_D$) se construit avec 124 verbes différents, alors que le couple (*sur, sort*$_D$) se rattache seulement à 5 verbes (*fixer, se pencher, pleurer, savoir, tromper*).

Le calcul de cet indice du degré d'autonomie des couples (p, n_{DET}) suit le principe suivant : un couple (p, n_{DET}) sera d'autant plus autonome qu'il se construira avec des verbes qui sélectionnent faiblement la préposition considérée.

Ce critère de productivité n'est cependant pas suffisant car il accorde un poids identique à chaque verbe, que son degré de sélection vis-à-vis de la préposition soit faible ou élevé. Par exemple, pour le couple (*sur, table*$_D$) on accorde la même importance au verbe *s'abattre* et au verbe *remplacer*[6], alors que le premier sélectionne fortement la préposition *sur* (*selec* = 0,76), et le deuxième très faiblement (*selec*=0,03). On veut donc traduire le fait que la présence du verbe *remplacer* dans la liste des recteurs du couple (p, n_{DET}) est un bon indice d'autonomie, ce qui n'est pas le cas du verbe *s'abattre*. Pour affiner la mesure, on définit ainsi une productivité pondérée $\text{prod}_P(p,n)$ comme la somme des degrés de sélection des verbes avec lesquels se construit le couple (p, n_{DET}) :

$$\text{prod}_P(p, n_{\text{DET}}) = \sum_{\{v \,/\, f(v,p,n) >= s\}} \text{selec}(v,p)$$

On fait alors l'hypothèse que plus la proportion de verbes sélectifs se construisant avec le SP est élevée moins on est assuré de l'autonomie du SP. C'est pourquoi on définit la mesure d'autonomie du SP comme l'écart à 1 de ce rapport entre $\text{prod}_P(p,n)$ et prod(p,n) :

$$\text{auton}(p, n_{\text{DET}}) = 1 - (\text{prod}_P(p, n_{\text{DET}}) / \text{prod}(p, n_{\text{DET}}))$$

La valeur de *auton* est comprise entre 0 (autonomie nulle) et 1 (autonomie maximale). Voici une illustration : les couples (*à, question*$_D$) et (*à, dehors*$_D$) ont la même valeur de productivité (ils sont rattachés à 27 verbes différents). Mais le coefficient d'autonomie de (*à, question*$_D$) est bas (0,3), alors que

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf08227

celui de (*à, dehors*$_D$) est élevé (0,8). En effet, tous les verbes auxquels se rattache le couple (*à, dehors*$_D$) ont un degré de sélection faible avec la préposition *à ;* alors que le couple (*à, question*$_D$) se construit avec une grande proportion de verbes qui sélectionnent fortement cette préposition. Le tableau 3 illustre ce contraste en listant les 10 premiers verbes associés à ces deux couples, par ordre décroissant de degré de sélection de la préposition *à*.

(*à, question*$_D$) auton=0,3		(*à, dehors*$_D$) auton = 0,8	
v	selec(v,*à*)	v	selec(v,*à*)
soustraire	1	transpirer	0,5
se intéresser	1	tendre	0,4
consacrer	1	manger	0,32
soumettre	0,99	travailler	0,27
renoncer	0,98	retenir	0,25
ressembler	0,97	passer	0,25
se attendre	0,94	se précipiter	0,25
se accrocher	0,91	pousser	0,25

Tableau 3 : Verbes associés à 2 couples (p, n_{DET}) présentant des valeurs d'autonomie différentes

Cette mesure permet donc d'ordonner l'ensemble des SP du corpus en fonction des critères que nous venons de décrire.

3.3 Des indices chiffrés pour statuer sur le degré d'autonomie d'un SP

Nous montrons dans cette dernière partie, à partir d'illustrations portant sur différents cas de figure dégagés, en quoi cet effort de quantification fournit des indications pour aider à déterminer le statut d'un SP (observation des valeurs d'*auton*).

3.3.1 Cas de forte autonomie

Les couples (*p*,*n*) présentant l'autonomie la plus forte comportent des prépositions dites ordinairement "circonstancielles", qu'il s'agisse de locutions (*au bout de* dét *table*, *au bord de* dét *mer*, *à portée de* dét *main*, *au fond de* dét *œil*...) ou de prépositions simples (*pendant* dét *jour*, *dès* dét *arrivée, sans _ émotion*…). Il n'est pas surprenant que ces couples obtiennent un niveau maximal d'autonomie puisqu'ils se rattachent à des verbes pour lesquels la préposition concernée représente une part négligeable des possibilités de rattachement prépositionnel (le degré de sélection est donc systématiquement bas). Ainsi, les prépositions *sans*, *dès*, *pendant*, *après*, etc. introduisent des SP dont l'autonomie est toujours égale ou supérieure à 0,9.

Si l'on s'intéresse aux prépositions plus courantes, seules les prépositions *avec* et *en* introduisent des couples présentant dans leur grande majorité une très forte autonomie. La situation est plus contrastée pour *dans*, *à* et *de*. Le tableau 4 montre des couples situés dans des valeurs d'autonomie hautes pour chacune de ces trois prépositions usuelles.

à	auton	*de*	auton	*dans*	auton
(*à*,*horizon*$_D$)	0,79	(*de*,*ton*$_D$)	0,87	(*dans*,*style*$_D$)	0,91
(*à*,*vitesse*$_D$)	0,77	(*de*,*trait*$_D$)	0,82	(*dans*,*langue*$_D$)	0,89
(*à*,*lueur*$_D$)	0,75	(*de*,*manière*$_D$)	0,82	(*dans*,*an*$_D$)	0,86
(*à*,*crayon*$_D$)	0,75	(*de*,*voix*$_D$)	0,81	(*dans*,*moment*$_D$)	0,85
(*à*,*surface*$_D$)	0,75	(*de*,*façon*$_0$)	0,78	(*dans*,*jeunesse*$_D$)	0,85

Tableau 4 : couples présentant une valeur d'autonomie élevée pour 3 prépositions fréquentes

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf08227

Ces couples renvoient à des séquences syntaxiquement et sémantiquement autonomes, porteuses d'informations circonstancielles relatives au temps (*connaître dans sa jeunesse*), à la manière (*déclarer d'un ton + Adj*), à l'espace (*découvrir à l'horizon*).

3.3.2 Cas de faible autonomie

Seules les prépositions *à* et *de* introduisent des couples dont la valeur d'autonomie est très faible (<0,2), et ces couples sont peu nombreux, comme le montre la figure 1.

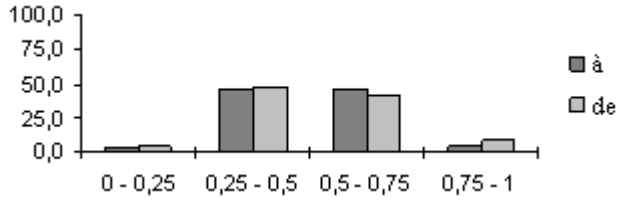


Figure 1 : Répartition des valeurs d'autonomie des couples (p, n_{DET}) introduits par *à* et *de*.

Le tableau 5 montre quelques exemples de couples présentant une cohésion maximale avec le verbe. De fait, on voit, en comparant ces couples à ceux du tableau 4, que la contribution sémantique des groupes prépositionnels correspondants est déterminée par le verbe dont ils dépendent (*renoncer à l'effort*, *tapisser de feuilles*), et que l'on a bien affaire dans cette zone à des couples instanciés par des SP argumentaux.

à	auton	de	auton
(*à*,*effort*$_D$)	0,12	(*de*,*feuille*$_O$)	0,09
(*à*,*volonté*$_D$)	0,14	(*de*,*nuage*$_D$)	0,14
(*à*,*exigence*$_D$)	0,15	(*de*,*voile*$_D$)	0,18
(*à*,*émotion*$_D$)	0,18	(*de*,*droit*$_D$)	0,18
(*à*,*influence*$_D$)	0,19	(*de*,*spectacle*$_D$)	0,19

Tableau 5 : couples présentant une valeur d'autonomie faible pour les prépositions *à* et *de*

3.3.3 Cas intermédiaire

Le contraste entre les deux pôles est net pour une partie des groupes prépositionnels traités (environ ¼ des couples (p, n_{DET}) pour un seuil de productivité de 10). Mais le principe d'une dichotomie marquée cède très vite le pas : la majorité des couples obtiennent un score d'autonomie intermédiaire (inférieur à 0,8 et supérieur à 0,3), et les données s'ordonnent ainsi selon un continuum, que la méthode permet précisément de mettre au jour. Le tableau 6 montre, toutes prépositions confondues, quelques exemples de valeurs intermédiaires.

(*à, scène*$_D$)	0,43
(*dans*, *lecture*$_D$)	0,46
(*à, feu* $_D$)	0,54
(*sur*, *nappe*$_D$)	0,55
(*de*, *nez*$_D$)	0,58
(*avec, curiosité*$_D$)	0,67

Tableau 6 : couples présentant une valeur d'autonomie intermédiaire

Les groupes prépositionnels représentés par les 5 premiers couples ont un comportement contrasté : ils s'associent à la fois à des verbes qui les sous-catégorisent (*assister à la scène, échapper au feu, retirer du nez*), et à des verbes avec lesquels ils entretiennent une relation sémantique plus lâche (*surgir à la scène,*

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf08227

brûler au feu, respirer du nez). Les compléments en *sur* et en *dans* ont tout particulièrement la propriété de se combiner à la fois avec des verbes qui requièrent un complément de destination (*plonger dans le travail*, *étaler sur la nappe*) et avec des verbes pour lesquels cette information est plus périphérique (*aider dans le travail*, *abandonner sur le drap*). Dans le cas du couple (*avec, curiosité*$_D$), on a affaire en revanche à un profil plus homogène : plus de la moitié des verbes auxquels il se rattache a un taux de sélection moyen (compris entre 0,3 et 0,6) et présentent une homogénéité sémantique remarquable (*contempler*, *regarder, examiner, observer, écouter, considérer*…). Ces caractéristiques s'étendant à d'autres compléments en *avec* (*avec stupeur*$_D$, *avec attention*$_D$), on peut dégager une configuration sémantique bien établie dans le corpus : VERBE D'OBSERVATION + COMPLEMENT DE MANIERE. Le calcul d'autonomie permet dans ce cas de mettre en évidence des associations sémantiques récurrentes dans le corpus entre des verbes et des SP qui leur sont régulièrement attachés sans toutefois être habituellement pris en compte dans la structure argumentale de ceux-ci.

3.4 Combiner mobilité et autonomie

Pour finir, nous proposons de recouper les deux informations que nous avons acquises sur les SP à partir du corpus, à savoir leur degré d'autonomie et leur capacité à occuper des positions préverbales, indice de mobilité. On peut s'attendre à ce que tous les SP mobiles soient fortement autonomes. Or, ce n'est pas toujours le cas. Certes, l'autonomie moyenne est élevée (0,76, contre 0,65 pour l'ensemble des SP postverbaux), mais on trouve malgré tout des SP dont la mesure d'autonomie est basse, ce qui semble constituer une anomalie. Considérons successivement les deux cas de figure :

a) Les SP présentent une forte autonomie en position postverbale et sont repérés comme mobiles. C'est par exemple le cas du couple (*dès, début*$_D$). Son autonomie est de 0,99 : le groupe se construit avec 13 verbes différents, dont le degré de sélection pour cette préposition est systématiquement proche de 0. On en trouve des instances à la fois en position initiale et en position d'incise entre le verbe et son sujet :

(11) Dès le début, il s'annonça sévère, et d'autant plus qu'on avait le ventre vide.

(12) L'intrigue, dès le début, éveilla un vieux souvenir.

Le croisement de ces deux méthodes permet de dégager des SP dont l'autonomie est avérée, ce qui fournit une information que l'on pourrait utiliser pour renoncer au rattachement par l'analyseur syntaxique de ces SP en position postverbale.

b) Les SP, bien que mobiles, présentent une autonomie faible en position postverbale.

C'est par exemple le cas des couples (*à, étonnement*$_D$) (auton = 0,14) et (*à, parole* $_D$) (auton = 0,31). Voici pour chacun un exemple d'apparition en position préverbale :

(13) Mais le récit, à son propre étonnement, avait tourné d'une autre manière.

(14) Aux paroles sèches de Naroumof, elle se leva, se dirigea vers Anne.

Cette double information contradictoire s'explique par le fait que, cette fois, lorsque ces couples s'instancient en position postverbale, ils entrent dans la portée du verbe (*il se mêlait à mon étonnement, il avait peine à s'intéresser aux paroles…*). Le décalage entre les deux indices s'explique donc par le fait qu'on n'a pas affaire aux mêmes types d'instances avant et après le verbe. Contrairement au cas du SP *dès le début*, qui s'émancipe systématiquement du verbe de la proposition quelle que soit sa position, les SP représentés par les couples (*à, étonnement*$_D$), (*à, parole* $_D$) n'ont un statut de complément de phrase qu'en position préverbale dans le corpus. Le retour aux occurrences particulières s'avère donc indispensable pour affiner la description obtenue par des approches quantifiées à partir du regroupement des occurrences captées sur l'ensemble du corpus.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf08227

4 Conclusion

Nous souhaitions montrer ici ce que l'utilisation de vastes quantités de données peut apporter à la description d'une distinction syntaxique aussi discutée que la distinction complément/ajout. Après avoir insisté sur la nécessité d'un certain nombre d'adaptations des méthodes habituellement utilisées en syntaxe et ce quel que soit le cadre théorique, nous avons envisagé de nouveaux critères visant à substituer à des tests binaires une mesure graduelle de l'autonomie des SP vis-à-vis du verbe. L'objectif de l'article était d'illustrer sur cette question le fait que la quantité des données produites par l'application de critères quantitatifs facilite l'accès à une diversité de contextes que les indices permettent de classer. Nous avons d'abord montré que les indices proposés mettent clairement en évidence deux types de fonctionnements : 1) des SP compléments qui dépendent étroitement du verbe isolés sur la base des scores obtenus interprétés comme marquant des dépendances fortes et 2) une diversité de fonctionnements qui cohabitent dans une zone intermédiaire et qu'il faudrait maintenant examiner plus en détail pour la caractériser précisément. En combinant le critère d'autonomie avec un critère distributionnel relatif au placement du constituant, nous avons également mis au jour des SP mobiles mais très peu autonomes lorsqu'ils sont en position postverbale. Cette méthode basée sur la quantification permet de sortir d'une vision strictement dichotomique du fonctionnement des SP pour construire des nouveaux observables, qui permettent de sonder la zone située entre les cas prototypiques habituellement décrits. La démarche que nous proposons, et les données empiriques que nous avons présentées s'inscrivent nettement dans les courants contextualistes dont (Legallois et François, 2006) offrent une synthèse, et notamment dans le courant des *pattern grammars* initié par (Hunston et Francis, 2000). Les informations que nous captons sur l'ensemble du corpus relativement aux associations au sein du triplet (v,p,n) permettent en effet d'observer, à l'interface entre grammaire et lexique, des phénomènes de colligation décrits par D. Legallois comme un « comportement de cooptation, de préférence mesurée statistiquement à partir de corpus, sans […] qu'il soit redevable à quelque principe structural ». Le travail sur corpus basé sur le relevé et la quantification de combinaisons lexico-grammaticales récurrentes offre ainsi la possibilité de s'émanciper de l'opposition binaire complément *vs* ajout pour élargir la description à une palette plus large de fonctionnements. Il reste à présent à mieux caractériser la diversité des comportements que nous avons relevés, qu'il s'agisse de la force d'association entre le verbe et la préposition ou des différents degrés d'autonomie et de mobilité qui caractérisent les entités partiellement lexicalisées (p,n_D) que nous avons examinées.

Références

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

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.

Charolles, M. (1997). L'encadrement du discours : univers, champs, domaines et espaces. *Cahiers de Recherche Linguistique*, 6, 1-73.

Combettes, B. (2007). Les ajouts après le point : aspects syntaxiques et textuels. M. Charolles, N. Fournier, C. Fuchs, F. Lefeuvre (dir). *Parcours de la phrase. Mélanges offerts à Pierre Le Goffic*, 119-131. Paris : Ophrys.

Fabre, C. et 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.

Fabre, C. et Frérot, C. (2002). Groupes prépositionnels arguments ou circonstants : vers un repérage automatique en corpus, *Actes du colloque TALN*, Nancy, 215-224.

Goldberg, A. et Ackermann, F (2001). The Pragmatics of Obligatory Adjuncts Language. *Language*, 77 (4), 798-814.

Habert, B. (2005). *Instruments et ressources électroniques pour le français*. Gap/Paris: Ophrys.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf08227

Ho-Dac, L.-M. et Péry-Woodley, M.-P. (2008). Temporal adverbials and discours segmentation revisited. In W. Ramm et C. Fabricius-Hansen (eds), *Linearisation and Segmentation in Discours (Multidisciplinary Approaches to Discourse 2008)*. Oslo : University of Oslo, 65-77.

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.

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

Legallois, D. (2005). Du bon usage des expressions idiomatiques dans l'argumentation de deux modèles anglo-saxons: la grammaire de construction et la grammaire contextualiste, *Les Cahiers de l'Institut de Linguistique de Louvain (CILL)*, 31, 2-4, 109-127.

Manning, C. D. (2003). Probabilistic Syntax. In R. Bod, J. Hay & S. Jannedy (eds). *Probabilistic Linguistics*. Cambridge : MIT Press.

Maurel, J.-P. (1990). Examen circumstantiarum. *Langue française*, *86 (1)*, 30-36.

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.

Rémi-Giraud, S. (1998). Le complément circonstanciel : problèmes de définition. In S. Rémi-Giraud & A. Roman (dir.). *Autour du circonstant*, 65-113. Lyon : Presses Universitaires de Lyon.

---

[1] Didier Bourigault a largement participé à la mise au point de cette méthode, qui a fait l'objet d'une première présentation dans (Fabre et Bourigault, 2008). Nous le remercions pour l'aide qu'il nous a apportée à tous les stades de ce travail. Nous remercions également Cécile Frérot, qui a contribué à faire avancer cette réflexion et a participé à une première expérience de travail sur la distinction entre arguments et circonstants, retracée dans (Fabre et Frérot 2002).

[2] Nous remercions Jean-Marie Pierrel, directeur de l'ATILF, de nous avoir fourni ce corpus de textes, dont certains sont encore sous droits, à des fins de recherche.

[3] Nous avons laissé de côté la position P3 (entre l'auxiliaire et le participe passé), qui correspond à des configurations difficiles à repérer pour l'analyseur.

[4] Ces SP sont en fait introduits par la locution prépositionnelle *à coup(s) de*. Cependant, cette locution n'étant pas identifiée par l'analyseur comme une unité, le nom *coup* est considéré comme la tête nominale du groupe.

[5] L'étiquetage est réalisé par l'outil d'annotation TREETAGGER développé à l'Université de Stuttgart.

[6] Le triplet (*remplacer, sur, table*$_D$) est par exemple extrait du passage suivant : *Mais le lendemain, je m'aperçus que Céler avait remplacé sur la table de travail à portée de mon lit un style de métal par un calame de roseau*.