Vaguer Céline

De l'utilité d'un corpus en syntaxe, mais quel corpus ?

RÉSUMÉ

Après avoir proposé une approche très générale de la notion de corpus *et de son utilité en linguistique, cette contribution se focalise sur le rôle et la place occupés par un corpus en syntaxe dans le cadre d'une étude de la complémentation verbale en* dans *tant du point de vue du recueil des données que de son outillage et son exploitation. Dans ce cadre d'étude, le corpus est loin d'être objectif puisqu'il dépend du point de vue adopté, de la démarche retenue, de l'hypothèse de travail. Notre corpus – banque de données (syntaxique, lexicale, sémantique, diversifiées...) ouvertes qui consiste en un regroupement de phrases isolées les unes des autres (absence de paragraphes) – est présenté à travers les notions de* représentativité, *d'*exhaustivité, *de* clôture.

1. Nécessité du corpus pour une linguistique qui privilégie la description des formes

Notre recherche sur les constructions verbales introduites par la préposition *dans* vise à rendre compte du fonctionnement linguistique (formel et sémantique) en partant des formes. Pour ce faire, nous nous intéressons, d'une part, à l'observation et la découverte de leurs propriétés (par l'étude des distributions et transformations qu'elles permettent ou excluent) et, d'autre part, à l'interprétation des compatibilités et incompatibilités ainsi dégagées. Les formes, puisqu'elles se prêtent à l'observation et à la manipulation, permettent d'accéder au sens en langue : on peut grossièrement distinguer trois grandes sortes de « sens », d'une part le sens intuitivement attribué aux mots, d'autre part le sens tel qu'il est véhiculé par le discours, enfin le sens en langue (à construire par hypothèse) – à quoi on peut ajouter le sens tel que construit par les scientifiques par exemple. Prenons l'exemple du mot *poule* : si l'on demande à un locuteur ce qu'il signifie, il répondra spontanément quelque chose comme « ce sont ces volatiles de basse-cour qu'on élève pour en consommer les oeufs [...] »

(cf. les dictionnaires) – donc sur des bases référentielles ou conceptuelles (ce n'est pas le mot *poule* qui est un volatile). L'insertion dans un discours fait apparaître diverses acceptions selon le cotexte, par exemple *Elever des poules en batterie* vs *Commander une poule au riz au restaurant* vs *Ah ! si vous connaissiez ma poule !* (M. Chevalier), etc. Le sens en langue subsume les acceptions observables en discours (le « signifié » saussurien, le « signifié de puissance » guillaumien, la « forme schématique » culiolienne, etc.). Pour le scientifique, la poule est un oiseau – mais linguistiquement, il n'est pas approprié de parler d'une poule par ce terme (on reprendrait un enfant qui, voyant une poule, la désignerait en disant *Il est joli cet oiseau !*).

Partir des formes suppose une réflexion sur la notion de *corpus* puisqu'on ne peut mener un travail linguistique à bien sans référence à un corpus (une fois que l'on a circonscrit la problématique, la constitution du corpus devient la première étape d'une recherche scientifique, la seconde serait son traitement). En effet, dans le domaine grammatical, personne n'a de définition spontanée permettant de livrer d'emblée l'identité d'une construction, de l'imparfait, de la préposition *par* ou de l'article *le* que toutefois tout le monde emploie quotidiennement sans se poser de question. Le locuteur interrogé peut naturellement avoir en mémoire ce qu'il a appris dans les cours de grammaire à l'école, mais il s'agit alors d'une connaissance extérieure qui vient se superposer à sa pratique de la langue.

Les linguistes eux-mêmes illustrent, par la diversité des théories avancées pour résoudre un problème, que la connaissance et l'utilisation quotidienne de la langue ne fournissent pas automatiquement les moyens d'en décrire le fonctionnement : traditionnellement, l'imparfait est présenté comme un temps (du passé) mais certains lui dénient une valeur temporelle (Touratier 1996, Le Goffic 1995), d'autres le définissent plutôt par ses propriétés aspectuelles (Ducrot 1979, Wilmet 1997), d'autres encore par son caractère anaphorique (G. Kleiber), *etc.* Même si les locuteurs ont une idée, d'ordre épilinguistique, de l'identité des lexèmes ou de la grammaticalité des phrases, elle relève du fragmentaire, de l'intuitif, et ne peut rendre compte de l'ensemble des données de manière globale et cohérente (c'est d'ailleurs la critique essentielle adressée aux grammaires et dictionnaires traditionnels). Autrement dit, en matière de langue, rien n'est jamais donné, n'apparaît de soi-même à la conscience : toute tâche d'ordre métalinguistique suppose un travail, une construction.

Ce travail linguistique inclut nécessairement la prise en compte d'un corpus quelle que soit l'approche retenue (*inductive* vs *déductive*) car même le linguiste le plus chevronné ne peut être sûr qu'il mobilise tous les emplois concernés pour le fait qu'il étudie : en témoigne le statut du contre-exemple, c'est-à-dire de l'énoncé qui dément ou met en cause une description élaborée à partir d'un nombre insuffisant de données au départ. Une des garanties de la bonne consistance de l'hypothèse qu'il propose pour expliquer un phénomène quelconque, c'est donc sa couverture empirique. Ainsi se pose à tout linguiste la question de la définition du corpus puisque c'est ce dernier qui l'amène à pouvoir formuler une hypothèse ou à en éprouver la consistance : selon sa représentativité, le corpus aura plus ou moins de chance de permettre au chercheur d'atteindre son objectif… Or on constate que bien souvent, dans les articles de linguistique, rien n'est dit sur le statut des données par le linguiste :

> pourtant la banalité et la fréquence d'emploi de ce terme dissimulent à peine des conceptions notablement diversifiées, en liaison avec les domaines d'étude et les approches spécifiques des uns et des autres (Dalbera, 2002 : 89).

2. Le point de vue à adopter face aux données…

Une fois la nécessité du corpus établie, il existe différents types de données et diverses attitudes à l'égard des données : le chercheur doit donc se situer et justifier son propre point de vue et ce à partir du constat que l'évolution des supports de recherches d'occurrences (notamment Internet) oblige à réfléchir sur la nature des données récoltées, comme le souligne justement S. Mellet (2002 : 6) :

> dans le champ linguistique, la notion de *corpus* s'est complexifiée au cours des dernières décennies en fonction de la diversité des pratiques et des objectifs assignés à la constitution et à l'exploration des corpus.

B. Habert (1995 : 4) mettait déjà en avant la nécessité d'une réflexion sur la constitution des corpus :

> à l'heure où le "captage" de texte […] devient si facile qu'il peut sembler suffisant de rassembler "du texte" en grande masse pour que les analyses quantitatives soient légitimes et productives, une réflexion approfondie sur la méthodologie de construction de corpus s'impose.

La nécessité de cette réflexion apparaît davantage encore lorsque l'on regarde comment les linguistes l'abordent et le définissent. Pour les uns, il

faut entendre par là un ensemble d'énoncés retenus, écrits ou oraux, qui sera soumis à l'analyse (Dubois 1969, Arrivé *et al.* 1986, Bouix-Leeman 1990, Leeman 2002, Mellet 2002). Mais pour d'autres, le corpus est en fait issu d'un travail préalable, puisque l'ensemble est restreint à ce qui est considéré comme "représentatif" (Riegel *et al.* 1994). Ainsi, il y a, nous semble-t-il, autant de corpus que d'objets d'étude, mais aussi autant de corpus que de points de vue non seulement théoriques et méthodologiques, mais encore selon que l'on est lecteur ou chercheur (Fillmore 1992)[1]. En effet,

> des corpus ont été constitués dans des buts divers, influant à leur tour sur la conception, la taille et la nature du corpus individuel. Quelques corpus actuels destinés à la recherche linguistique ont été conçus dans des buts descriptifs généraux – c'est-à-dire qu'ils ont été conçus afin qu'ils puissent être examinés ou épluchés en vue de recherches linguistiques diverses sur la prosodie, le lexique, la grammaire, l'organisation discursive ou la pragmatique de la langue. D'autres corpus ont été conçus dans des buts plus précis comme par exemple découvrir quels mots et quels sens il faudrait répertorier dans un dictionnaire destiné aux apprenants... (Kennedy, 1998 : 3-4, notre traduction).

3. Rôle et place du corpus en syntaxe

Ce qui précède concerne une approche très générale de la notion de *corpus* et de son utilité en linguistique, ce point de vue vaut-il indifféremment pour tous les sous-domaines des Sciences du langage, et en particulier pour la syntaxe ?

3.1. Rôle du corpus : en fonction du point de vue, de la démarche, de l'hypothèse retenus

Notre point de vue est celui du chercheur, pour reprendre C. J. Fillmore (1992), qui ne prend pas connaissance d'un certain travail mais qui opère le travail de constitution des données nécessaires à l'identification du fait qu'il étudie. Nous nous situons donc dans le cadre de la linguistique descriptive : notre démarche consiste à rassembler le corpus de compléments verbaux en

1. C. Vaguer (2003) met en évidence l'existence de différentes conceptions de la notion de *corpus*, de différentes attitudes à l'égard des données, des différentes démarches pour élaborer les corpus (les avantages et inconvénients du recours aux données attestées ou construites), des différents jugements que l'on produit sur les données (l'acceptabilité et la grammaticalité), et les motivations qui l'ont conduite à constituer un « corpus informatisé ».

dans à partir des critères retenus par la tradition syntaxique (qui fonctionnent donc comme hypothèses, construites à partir d'un petit nombres d'exemples, le plus souvent forgés) – sinon, le repérage ne pouvait se faire qu'à partir de la présence matérielle de *dans*, ce qui nous mettait face à un nombre trop grand d'énoncés pour que nous puissions les maîtriser.

Ainsi, la constitution du corpus intervient après avoir défini une hypothèse de travail. Autrement dit, les observables ne constituent pas quelque chose de « brut » mais résultent d'une certaine interprétation, d'une certaine décision, d'une prise de conscience préalable de ce qu'il est pertinent d'observer (d'où, sans doute, le fait que certains emplois – tel *dans* au sens « coïncidence » introduisant une apposition – ne soient pas enregistrés par les lexicographes). En somme, on ne peut trouver que si l'on cherche, et pour chercher il faut avoir quelque chose à chercher ! Ainsi construit-on son corpus en fonction de l'objectif de sa recherche. Mais ce point de vue reste celui de la personne qui le constitue. Le corpus n'a de sens que par rapport à ce que l'on va lui faire subir :

> [C'] est un objet heuristique. C'est une construction arbitraire, une composition relative qui n'a de sens, de valeur et de pertinence qu'au regard des questions qu'on va lui poser, des réponses que l'on cherche, des résultats que l'on va trouver [...] C'est l'intention du chercheur qui est importante et lui donne son sens (Mayaffre, 2002 : 55).

Ainsi pour chercher il faut savoir ce que l'on cherche et c'est de cela que va dépendre la mise en œuvre de la démarche de constitution du corpus (quelles données, quelles sources, quel type de recueil, quel outillage...). Projetant de saisir l'identité de la préposition *dans* par l'étude des constructions et des distributions dans lesquelles elle s'insère, principalement dans le cadre de la complémentation verbale, notre corpus a été rassemblé en fonction de cet objectif et à partir du constat que les descriptions disponibles (principalement dans les dictionnaires et les grammaires) étaient insuffisantes. En effet, lorsque l'on quitte les exemplifications fournies pour observer les emplois de la préposition *dans*, on se rend compte qu'elle traduit une « coïncidence interne » dans tous ses emplois. Autrement dit, il n'y a plus lieu de dire, comme le font les grammaires ou dictionnaires, que « *dans* indique le lieu, le temps, la manière d'être et l'état, l'évaluation approximative... » puisque, d'une part, cette liste n'est jamais exhaustive (on ne parle pas en effet des acceptions où

elle serait interprétée comme véhiculant de la cause : *Un couple s'enlise (dans + à cause de la routine)*, ou une cause-durée : *Je m'embrouille (dans + à cause de + au cours de) mes explications*, le moyen : *Il se drape (dans + au moyen de + à l'aide de) un linceul*, ou une identité référentielle : *La principale nouveauté réside dans la création d'un nouveau taux d'imposition* (La création d'un nouveau taux est la principale nouveauté =: La principale nouveauté est la création d'un nouveau taux d'imposition) ; et que, d'autre part, ces descriptions ne fournissent pas l'identité sémantique de la préposition : si *dans* apparaît dans chacun de ces emplois (*temps*, *lieu*...), c'est bien la même forme qui est en jeu, donc elle doit avoir un sens : quel est-il ? Il ressort de notre étude (2004a, 2004b) que la notion de *coïncidence* est à même de caractériser cette préposition.

3.2. Place du corpus dans une recherche sur la complémentation verbale

Après avoir présenté le rôle du corpus pour un chercheur qui se préoccupe de complémentation verbale, nous allons maintenant nous focaliser sur la place occupée par le corpus au sein de la recherche, tant du point de vue du recueil des données que de son outillage et son exploitation.

3.2.1. Place du corpus et recueil des données

Du point de vue du recueil des données, après avoir défini une hypothèse de travail et nous être munie de la liste de verbes fournie par A. Dugas & H. Manseau (1996) qui spécifie pour chacun des 8 652 verbes retenus les prépositions susceptibles d'être sous-catégorisées, nous avons recherché les énoncés attestés (1 200 énoncés), extraits de sources diverses (essentiellement écrites), susceptibles d'étayer l'inventaire de l'index de A. Dugas & H. Manseau et qui présentent l'avantage de ne pas avoir été produits en fonction de l'hypothèse de travail retenue.

3.2.2. Quel outillage pour le corpus ?

Du point de vue de l'outillage, l'ensemble des énoncés attestés et les données qui y sont associées (identification de la source, propriétés syntaxiques du GP, analyse distributionnelle...) a été regroupé dans une

base de données *Access* qui présente de multiples avantages tant pour son concepteur que pour sa réutilisation[2].

3.2.3. Représentativité, exhaustivité et clôture du corpus

Une fois la machine de recherche d'occurrences lancée (au moyen de requêtes automatiques dans Glossanet ou Frantext, ou munie d'un crayon au fil de nos lectures), se posent à tout linguiste les questions de *représentativité*, d'*exhaustivité*, et de *clôture*. En effet, chacun espère que l'ensemble des énoncés qu'il a rassemblés sera représentatif d'une réalité plus large – par exemple, pour nous, que les attestations contenues dans notre base de données seront représentatives de l'ensemble des emplois de la préposition *dans* dans le cadre de la complémentation verbale et pourraient donc servir de « corpus de référence ». Parler de *représentativité* et d'*exhaustivité* pour un corpus paraît illusoire – or, c'est ce qui définit la notion de *corpus* dans certaines réflexions linguistiques :

> le corpus doit être *représentatif*, c'est-à-dire qu'il doit illustrer toute la gamme des caractéristiques structurelles. On pourrait penser que les difficultés sont levées si un corpus est *exhaustif*, c'est-à-dire qu'il réunit tous les textes produits (Dubois *et al.*, 1999 : 124).

– car qu'est-ce qu'un corpus exhaustif, quels sont les critères qui permettent de l'identifier ?

L'**exhaustivité** est en fait impossible à atteindre, parce que cela supposerait que l'on est capable d'embrasser la totalité du dicible ; *a priori*, la chose est concevable[3] mais dans la réalité concrète du travail linguistique, le corpus ainsi obtenu serait difficilement manipulable – du moins par un humain (il pourrait le devenir grâce aux traitements automatiques) : il faudrait, par exemple pour tester toutes les phrases possibles de type GN_1 V *dans* GN_2, multiplier dans chaque GN tous les déterminants par tous les noms du français et de même toutes les combinaisons de GN_1 avec tous les V et toutes les combinaisons de GN_2 ; mais dans la liste ainsi obtenue, la séparation des séquences acceptables et inacceptables reposerait de toute

2. Je renvoie ici à mes contributions (Vaguer 2003, 2004a & 2004b, 2005) pour une présentation de la base de données, des motivations qui ont présidé à sa constitution et des avantages qu'elle offre.
3. M. Gross (1975 : 18) montre que l'on peut calculer le nombre (infini, d'un point de vue mathématique) de phrases de tel ou tel type. Mais il est un fait que la possibilité de calculer un nombre infini ne veut pas dire que l'on dispose des phrases en question...

façon sur l'intuition du linguiste : à supposer qu'elle soit possible, l'exhaustivité ne garantirait donc pas l'objectivité du corpus obtenu.

Notre étude s'approche donc d'une certaine exhaustivité des emplois « V *dans* GN » mais s'en éloigne aussi du fait qu'il nous manque quelques attestations concernant les verbes signalés par A. Dugas et H. Manseau.

Quant à la **représentativité**, là encore on n'est jamais à l'abri d'une nouvelle découverte (comme le montre l'emploi « coïncidence » de *dans*) : on ne peut qu'espérer que l'ensemble des énoncés relevés est représentatif d'une certaine chose à découvrir – en l'occurrence, dans le cadre de notre recherche, de la complémentation verbale en *dans*. Le fait d'être partie de l'index d'A. Dugas et H. Manseau nous a permis d'avoir une idée de départ de l'étendue des emplois verbaux de la préposition *dans*, et la recherche automatique d'occurrences nous a permis de ne pas nous limiter à cette liste de départ. Mais, tant qu'on ne sait pas en quoi consiste la totalité des emplois (objectif impossible à atteindre, comme on l'a vu), on ne peut pas savoir si ceux que l'on décrit correspondent à une petite partie, à la moitié, à la majeure partie d'entre eux. Donc cette question de la représentativité n'est pas pertinente pour le linguiste « de langue », car tout au plus peut-il se préoccuper de rendre compte des emplois connus (*i.e.* ceux qui ont émergé en tant que données / observables) ; en revanche, la notion peut être pertinente pour celui qui analyse les discours (pour caractériser des genres de textes, par exemple, ou opérer des comparaisons d'ordre sociolinguistique, etc.)[4].

4. Notre objectif premier est d'étudier la langue et de caractériser certains types de constituants et non d'observer comment ces derniers sont utilisés dans les discours, ce qui relativise l'importance de leur représentativité. Nous aurions pu nous limiter à l'œuvre de Rabelais ou à une année du journal *Le Monde* ; ce type de corpus ainsi défini aurait permis de proposer une étude quantitative des emplois de la préposition *dans* chez tel auteur ou dans tel support mais ne nous aurait pas permis de trouver attesté l'ensemble des verbes susceptibles de sous-catégoriser cette préposition. C'est pourquoi nous avons eu recours à des énoncés attestés issus de sources différentes. Notre propos étant de vérifier que tous les verbes décrits comme se construisant avec *dans* chez A. Dugas & H. Manseau (1996) recevaient bien une illustration attestée, l'hypothèse est que ce n'est pas chez un seul auteur (fût-il celui de la *Recherche du Temps perdu*) qu'on les trouvera tous employés. Diversifier les sources est apparu comme la garantie d'un plus grand éventail de verbes. Ce choix n'est pas intangible, variant selon l'objectif : dans l'optique par exemple de comparer la fréquence relative des GP en *dans* à sens spatial et à sens temporel dans la langue littéraire du XVIe s. et du XXe s., il peut être justifié de partir de deux œuvres quantitativement importantes : Rabelais et Proust. De même si l'on cherche à évaluer la fréquence des GP en *dans* à sens causal par rapport aux GP en *dans* à sens spatial et temporel, avec l'idée de vérifier que le sens prototypique de *dans* est d'ordre locatif.

Il n'en reste pas moins que notre corpus est représentatif d'un certain type d'emploi de la préposition *dans*, qui n'a pas été enregistré jusqu'ici (ce que l'on a étiqueté « coïncidence ») ou de cas qui n'ont été que peu abordés (ce que les lexicographes rangent sous « le temps », « l'état », « l'approximation ») du fait que les études jusqu'ici se sont essentiellement consacrées à la complémentation spatiale[5] : comparé à la sur-représentation des emplois locatifs dans les corpus antérieurs, le nôtre au contraire les sous-représente au profit des autres cas de figure possibles. Ce corpus est donc représentatif de ce qu'on cherche à mettre en évidence : la complémentation verbale en *dans*.

Le linguiste, en déclarant que son corpus est représentatif de ce qu'il souhaite mettre en évidence, ne prononce-t-il pas la **clôture** de son corpus ? Là encore, nous ne disposons pas d'outil objectif permettant de dire à quel moment il faut clore son corpus et si c'est nécessaire de le faire :

> La clôture du corpus a relevé de la responsabilité du chercheur, et la représentativité du corpus est exclusivement du ressort du chercheur. Le corpus apparaît dès lors clairement comme un objet construit… et travaillé (Dalbera, 2002 : 93-94).

Nous avons considéré notre corpus clos pour les besoins de l'analyse. Évidemment, le corpus est complété chaque fois que l'on trouve des attestations pour des verbes non mentionnés. C'est en quelque sorte une clôture fictive ! mais qui permet de passer à autre chose. Car si la constitution du corpus est pour nous incontournable, elle n'est qu'une des étapes de la recherche linguistique : après la récolte des énoncés, qui monopolise beaucoup de temps, il faut passer à leur analyse et il est difficile de mener les deux de front. Le test des propriétés et la justification des

Cependant, à supposer que les GP en *dans* à sens spatial soient plus fréquents que les autres, l'observation reste de l'ordre du discours (témoignant de ce que les locuteurs actualisent le plus souvent) ; du point de vue du fonctionnement du système linguistique lui-même, il n'y a pas de raison de dire que le sens locatif est premier puisque l'analyse des constructions ne permet pas de le mettre en évidence. Mais l'interprétation d'une fréquence supérieure à une autre n'est pas non plus très claire du point de vue du discours même : que peut-on conclure de la mise en discours que de mettre en relief que les emplois locatifs sont plus nombreux que les emplois temporels dans l'œuvre de Rabelais ? Que Rabelais décrit les lieux plutôt que des périodes temporelles ou des relations logiques ? Que la prééminence quantitative des GP en *dans* à sens spatial est typique du genre de texte adopté (ce n'est pas de la poésie, ce n'est pas du théâtre, ce n'est pas un texte scientifique, etc.) ? Que la langue écrite littéraire du XVIe s. privilégie pour *dans* l'emploi locatif ? Etc.

5. Cf. C. Vandeloise (1986) et la critique que lui adresse sur ce point C. Vaguer (2004c).

interprétations correspondent à un tout autre travail[6], que gêne et ralentit la gestion de grandes quantités de données à analyser, qui pourraient finalement s'avérer inutiles. Mais comme nous l'avons déjà signalé, rien ne garantit qu'on ne trouvera pas demain quelque chose que ne contient pas le corpus accumulé jusqu'à aujourd'hui, si vaste soit-il : la lecture quotidienne d'un journal n'apporte que de loin en loin des attestations des emplois de *dans* au sens que nous appelons « coïncidence », et des années de publications successives de grammaires et de dictionnaires ont passé sans que personne ne repère ce type d'emploi... Mais il a suffi d'un exemple pour déclencher l'intuition et rendre attentif aux autres actualisations possibles de ce type de complément.

Un dernier problème qui se pose au linguiste est celui de son **objectivité** à l'égard des données. Dès que l'on manipule les énoncés, on fait intervenir une intuition (la sienne et par conséquent nécessairement une certaine subjectivité) – même dans l'application de critères, le résultat affecté au test dépend du sentiment linguistique du linguiste. De plus, dans l'ensemble des emplois rencontrés pour un type de verbe, on ne va en garder qu'un certain nombre, sur la base là aussi de jugements personnels : on élimine ce qui *paraît* redondant et l'on garde les énoncés qui *semblent* illustrer ce que l'on cherche à mettre en évidence mais on ne signale pas ce sur quoi on n'a rien de particulier à observer... L'objectivité revendiquée par les tenants du corpus attesté n'est donc qu'apparente, cachant un jugement d'acceptabilité refoulé :

> Nous nous fabriquons tous une grammaire subjective qui entraîne des différences profondes dans les jugements d'acceptabilité (Culioli, 2000 : 17).

Ainsi les énoncés de notre corpus, qu'ils soient attestés ou construits, comportent finalement tous un jugement personnel d'acceptabilité[7],

6. Quel que soit le type de corpus retenu, utilisé, le linguiste devra décider à un moment ou à un autre si son corpus est saturé : « un corpus est *saturé* quand le linguiste juge qu'il est inutile de l'étendre davantage » (Chiss *et al.*, 1993 : 61). C'est seulement quand le linguiste aura décidé de la saturation de son corpus qu'il pourra commencer son analyse afin de mettre au jour les régularités observées dans cet ensemble fini de données.

7. Pour R. Coppieters (1997 : 23), « en syntaxe et en sémantique il a d'ailleurs toujours été indispensable de tenir compte des données de la réflexion intuitive : jugements de grammaticalité, d'acceptabilité, d'opposition, de congruence avec un contexte donné, etc. ». Cf. également sur ce point D. Willems (2000 : 151) : « L'observation fournit des données qualitatives et quantitatives précieuses, l'introspection permet des manipulations syntaxiques et

jugement que portera à son tour le lecteur. Non seulement le linguiste travaille sans cesse avec ces jugements d'acceptabilité (Milner, 1978 : 21) mais de surcroît le travail linguistique lui-même n'est possible qu'à cette condition, en ceci que la simple « observation » (qui ne ferait intervenir aucun *a priori* – au moins conscient) ne peut concerner qu'un objet que l'on a au préalable constitué comme « observable » : l'attention ne peut s'appliquer qu'à quelque chose qui apparaît problématique, qui suscite interrogation – et donc réflexion pour trouver une réponse, une explication.

3.2.4. Exploitation du corpus en syntaxe

Du point de vue de son exploitation au sein de la recherche, le corpus rassemblé dans notre base de données a été sans cesse étoffé d'informations diverses (identification de la source, analyses syntaxique, distributionnelle et sémantique...). Par exemple, nous avons une table qui contient l'application de l'ensemble des tests syntaxiques (suppression, déplacements...) permettant l'identification des groupes prépositionnels, donc nous avons forgé un corpus de phrases pour les besoins de l'analyse puisque l'on ne trouvera jamais dans le corpus attesté une même phrase soumise à ces différentes transformations. Il ressort de ces manipulations des énoncés que nous jugerons inacceptables et/ou agrammaticaux, ils le seront pour nous (pour d'autres locuteurs, il peut en être autrement). De plus, notre corpus est « annoté » puisque pour chaque énoncé une analyse syntaxique, distributionnelle et sémantique a été effectuée. Du point de vue syntaxique par exemple, nous avons procédé à l'étiquetage de la construction par l'identification du syntagme et de sa fonction notamment en distinguant entre $GP_{complément}$ et $GP_{modifieur}$ et, au-delà de la notion de complément, nous avons aussi pris en compte les structures prédicatives et les expressions figées. Cet étiquetage offre de multiples possibilités en matière de recherche : on peut ainsi envisager de tester quelques analyseurs syntaxiques disponibles. L'article de D. Bourigault et C. Fabre (2000) nous a vraiment confortée dans cette idée. La nouveauté de leur analyseur SYNTEX réside dans le désir d'extraire des syntagmes verbaux alors que d'ordinaire les analyseurs se focalisent sur le repérage de syntagmes nominaux (cf. l'analyseur LEXTER). L'idéal aurait été de pouvoir tester

lexicales minimales, indispensables à la reconnaissance des éléments pertinents de la structure ».

SYNTEX sur notre corpus de $GP_{compléments}$ en *dans* et cé, d'autant qu'il n'existe pas :

> pour le français, de base lexicale complète et facilement accessible qui recenserait les propriétés de complémentation des verbes, noms et adjectifs (*op. cit.* : 137-138).

Par notre thèse et la création de notre base de données centrée sur la complémentation verbale de la préposition *dans*, nous pouvons peut-être commencer à combler ce manque (le corpus ne devient plus un simple passage obligé pour la recherche mais est au cœur de celle-ci). D'où l'importance de la création de liens entre les recherches théoriques en linguistique et les travaux de conception d'outils de traitement automatique des langues :

> nous croyons qu'une confrontation entre approches des deux types, menée dans le cadre du programme de recherche de l'aide à l'interprétation de corpus, doivent conduire à une fécondation réciproque [...] le souci de réaliser des analyses plus fines, pour améliorer la qualité des résultats fournis par l'analyseur, nous conduit naturellement à rechercher un appui du côté des théories syntaxiques (*op. cit* : 145).

4. Références bibliographiques

Arrivé, M. *et al.* (1986). *La grammaire d'aujourd'hui – Guide alphabétique de linguistique française*. Paris : Flammarion.

Bouix-Leeman, D. (1990) Le problème du sens dans la constitution du corpus. In C. Normand (Ed.), *La quadrature du sens. Questions de linguistes* (pp. 111-129). Paris : P.U.F.

Bourigault, D. & Fabre, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. In A. Condamines (Ed.), *Cahiers de grammaire, 25*, 131-151.

Chiss, J.-L. *et al.* (1993). *Linguistique française - Notions fondamentales, phonétique, lexique*. Paris : Hachette Supérieur.

Coppieters, R. (1997). Quelques réflexions sur la question des données : corpus et intuitions. *Recherches sur le français parlé, 14*, 21-41.

Culioli, A. ([1990] 2000). *Pour une linguistique de l'énonciation. Opérations et représentations*, T. 1. Gap/Paris : Ophrys.

Dalbera, J.-P. (2002). Le corpus entre données, analyse et théorie. *Corpus, 1*, 89-104.

Dubois, J. *et al.* ([1994] 1999). *Dictionnaire de linguistique et des Sciences du langage*. Paris : Larousse.

Dubois, J. (1969). Grammaire distributionnelle, *Langue française, 1*, 41-48.

Ducrot, O. (1979). L'imparfait en français. *Linguistische Berichte, 60*, 1-23.

Dugas, A. & Manseau, H. (1996). *Les verbes logiques*. Montréal : Les Éditions Logiques.

Fillmore, C. J. (1992). "Corpus linguistics" or "Computer-aided armchair linguistics". In J. Svartvik (Ed.), *Directions in Corpus Linguistics* (pp. 35-60). Berlin/New York : Mouton de Gruyter.

Gross, M. (1975). *Méthodes en syntaxe*. Paris : Hermann.

Habert, B. (Ed.) (1995). *Traitement Automatique des Langues (T.A.L.), 36* (1/2). ATALA, CNRS.

Harris, Z. (1968). *Mathematical structures of Language*. New York : John Wiley & Sons, Interscience Publishers (trad. Française (1971). *Structures mathématiques du langage*. Paris : Dunod).

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London/New York : Longman.

Le Goffic, P. (1995). La double incomplétude de l'imparfait. *Modèles linguistiques, 31* (XVI : 1), 133-148.

Leeman, D. (2002). *La phrase complexe. Les subordinations*. Bruxelles : De Boeck/Duculot.

Mayaffre, D. (2002). Les corpus *réflexifs* : entre architextualité et hypertextualité. *Corpus, 1*, 51-69.

Mellet, S. (Ed.) (2002). Corpus et recherches linguistiques. Introduction. *Corpus, 1*, 5-12.

Milner, J.-C. (1978). *De la syntaxe à l'interprétation. Quantités, insultes, exclamations*. Paris : Le Seuil.

Riegel, M. *et al.* (1994). *Grammaire méthodique du français*. Paris : P.U.F.

Touratier, C. (1996). *Le système verbal français*. Paris : Armand Colin.

Vaguer, C. (2003). *Corpus, vous avez dit corpus ! De la notion de corpus à la création d'un « corpus informatisé »*. Communication aux *3èmes Journées de la linguistique de corpus*, Lorient (11-13 septembre). A par. aux Presses Universitaires de Rennes.

Vaguer, C. (2004a). Constitution d'une base de données : les emplois de *dans* marquant la « coïncidence ». *Revue Française de Linguistique Appliquée*, IX-1, 83-97.

Vaguer, C. (2004b). *Les constructions verbales "V* dans *GN". Approches syntaxique, lexicale et sémantique*, Thèse de doctorat, Université Paris X–Nanterre.

Vaguer, C. (2004c). La préposition *dans* et les verbes dits "de mouvement". Du "spatial", au sens propre et au sens figuré ? *Recherches linguistiques*, *27*, (à paraître), sous la direction de P. Dendale.

Vaguer, C. (2005). Une base de données comme moyen de communication scientifique ? Actas-I, *IXème Simposio International de comunicación social*, organisé par le *Centro de lingüistica Aplicada* y *El Ministerio de Ciencia Tecnologia, y Medio ambiente*. Santiago de Cuba, 134-138.

Vandeloise, C (1986). *L'espace en français*. Paris : Le Seuil.

Willems, D. (2000). Chapitre III : Diversité des domaines d'application. Introduction. In M. Bilger (Ed.), *Corpus. Méthodologie et applications linguistiques* (pp. 149-155). Paris : Honoré Champion et les Presses Universitaires de Perpignan.

Wilmet, M. (1997). *Grammaire critique du français*. Paris/Louvain-la-Neuve : Duculot.

Etude des collocations épistémiques en corpus spécialisé : de la théorie à l'empirie…

RÉSUMÉ

A travers l'étude des collocations lexicales épistémiques en corpus électronique spécialisé, l'imbrication entre théorie linguistique et empirie de corpus est constante, mais de nature variable selon les étapes du travail. En adoptant un point de vue méta sur la recherche en cours, nous avons identifié trois types d'imbrications possibles, qui chaque fois ont contraint nos orientations théorico-pratiques et inversement, suivant en substance, un schéma circulaire progressif et cumulatif.

1. Introduction

Toute analyse linguistique, qu'elle soit phonologique, syntaxique, sémantique ou encore pragmatique (…), est sous-tendue par un modèle théorique, garant de l'hypothèse posée. Aussi, les analyses linguistiques sur corpus, basées sur des données attestées (linguistique de corpus), sont le lieu privilégié des confrontations entre théorie et empirie, posant inéluctablement la question du degré de leur imbrication et des contraintes théorico-pratiques qui en découlent.

A travers l'étude des collocations lexicales épistémiques en corpus électronique spécialisé nous avons dû réfléchir aux rapports mutuels qui régissent les relations entre théorie linguistique et empirie de corpus. En adoptant un point de vue *méta* sur la recherche en cours, nous avons identifié trois types d'imbrication possibles, correspondant à trois étapes suivant la progression chronologique du travail, qui chaque fois ont contraint nos orientations théorico-pratiques et inversement :

- l'étape où la théorie linguistique contraint l'empirie de corpus ;
- l'étape où les données du corpus orientent l'analyse linguistique ;
- l'étape où les résultats nourrissent la théorie linguistique.