CORPUS, VOUS AVEZ DIT CORPUS ! DE LA NOTION DE CORPUS À LA CRÉATION D'UN « CORPUS INFORMATISÉ »

Céline Vaguer
UMR 7114 – MoDyCo – Université Paris X-Nanterre

1. INTRODUCTION

On ne peut mener un travail linguistique sans référence à des « données » : ainsi toute grammaire ou tout dictionnaire arbore des « exemples » ; on ne parle pas pour autant, dans ces cas, de « corpus » : il semble que la notion soit entendue (en particulier dans les débuts de la grammaire générative) comme « un ensemble de données produit indépendamment du linguiste et de la recherche linguistique », par opposition aux données que le linguiste est susceptible de produire lui-même : Chomsky s'oppose à l'idée que l'étude d'un corpus puisse mener à la construction d'une grammaire appropriée, comme à l'idée que le corpus des énoncés que l'enfant entend autour de lui soit la base de sa « compétence » (de la grammaire qu'il se construit mentalement). Ainsi le raisonnement linguistique de Chomsky s'opère bien sur des données concrètes, mais qu'il n'appelle pas « corpus ». Donc le débat instauré par Chomsky – étant donné le sens qu'il donne à « corpus » comme, disons, un « ensemble de discours produit extérieurement au linguiste et au travail linguistique » – c'est celui de la pertinence du « corpus » par rapport à ce que le linguiste (de par sa compétence de sujet parlant) peut produire lui-même, d'une part, ou par rapport à ce que la grammaire telle qu'il l'a construite peut prédire, d'autre part. L'argument de Chomsky à l'encontre du « corpus » (comme base pertinente de la description et du raisonnement linguistique), c'est le fait que, pour raisonner sur la langue, il faut pouvoir confronter ce qui est possible et ce qui ne l'est pas, or – par définition – le « corpus » (tel qu'il l'entend) ne peut pas fournir d'exemples de ce que la langue ne permet pas ; de plus, en tant que texte produit à un moment donné, par un ou des locuteurs particuliers, selon un thème, une intention, une situation, des interlocuteurs particuliers, un « corpus » ne peut évidemment illustrer tous les cas de figure d'un phénomène linguistique donné (par exemple : tous les auxiliaires et

combinaisons d'auxiliaires) ; et enfin, en tant que produit fini, le « corpus » ne peut pas non plus laisser voir certaines propriétés linguistiques comme la récursivité (le retour potentiellement infini d'une même structure).

Le présupposé est que le linguiste, de par sa propre compétence de sujet parlant, est à même de produire les données pertinentes (grammaticales et agrammaticales), permettant de faire l'hypothèse de règles dont il vérifiera la pertinence en jugeant si l'ensemble des énoncés qu'elles peuvent produire est, ou non, conforme à ce qu'autorise la langue – c'est-à-dire ce que le linguiste lui-même considère comme acceptable ou inacceptable. Ainsi, dans ce cadre, le travail du linguiste suppose nécessairement le recours à l'intuition pour constituer les données, les manipuler, raisonner sur le résultat de ces manipulations, mais en même temps, il y a un doute sur la pertinence de l'exercice de l'introspection – ce pourquoi justement les structuralistes et les distributionnalistes avaient prôné le recours au « corpus ». Mais on sait aussi que ce dernier n'est pas la panacée, ainsi que l'a pointé Chomsky. Toute recherche entreprise doit donc se mettre au clair sur ce point méthodologique :

- qu'est-ce qu'un corpus ?
- quel est ou quel doit être le statut du corpus dans l'investigation linguistique ?

Dans un premier temps, nous ferons un bilan sur cette notion de *corpus* – bilan né du constat que bien souvent, dans les articles de linguistique, rien n'est dit par les linguistes sur le statut des données : « ressources dont les natures différentes ne sont pas nécessairement distinguées par le linguiste, qui les nommera toutes "corpus" » (Gasiglia, 2003), mais aussi du fait que les supports de recherche d'occurrences ont évolué et qu'il est donc primordial de réfléchir sur la nature des données ainsi récoltées. Pour ce faire, nous mettrons en évidence l'existence de différentes conceptions de la notion de *corpus*, de différentes attitudes à l'égard des données, de différentes démarches pour élaborer les corpus, de différents jugements que l'on produit sur les données. Puis, nous justifierons le point de vue que nous avons adopté en tant que chercheur, et nous exposerons la démarche retenue pour constituer notre corpus : la méthodologie et la constitution d'une base de données.

2. LA NOTION DE *CORPUS*

Quelles que soient la théorie et la méthodologie retenues, 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. Saussure (1916 *in* 1972) avait raison de dire que « en matière de langue, on s'est toujours contenté d'opérer sur des unités mal définies ».

2.1 Les différentes conceptions de la notion de corpus

L'existence de différentes conceptions de la notion de *corpus* apparaît lorsque l'on regarde comment les linguistes l'abordent et la définissent. Pour les uns, il faut entendre par là un ensemble d'énoncés retenus, écrits ou oraux (parmi l'*univers*[1] des possibles), qui sera soumis à l'analyse : « base d'observation permettant d'entreprendre la description et l'analyse de la langue en question » (Arrivé *et al.*, 1986). 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 « représentatifé » ; c'est le cas de Riegel *et al.* (1994) qui spécifient de surcroît que les données doivent être « attestées » :

> « On peut rassembler un ensemble de textes ou d'énoncés jugés représentatifs de la langue… Une telle collection ne comprenant que des données attestées (des énoncés effectivement produits) constitue un *corpus* ».

Le corpus retenu, qui aura alors subi un jugement d'acceptabilité de la part du linguiste, puisque « le linguiste trie les énoncés qu'il va soumettre à l'analyse » (Dubois *et al.*, 1999), sera considéré comme un « échantillon de la langue » (*op. cit.*) que tout linguiste souhaite représentatif[2], en ce sens qu'il espère qu'il illustre l'ensemble des possibilités structurelles existantes (par exemple de l'emploi de la préposition *dans*), tout en sachant qu'il ne sera pas exhaustif[3] puisqu'on ne peut prétendre rassembler tous les énoncés possibles…

2.2 Les différentes attitudes à l'égard des données

Ainsi, existe-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, ou encore selon que l'on est lecteur ou chercheur (Vaguer, 2004b & 2005b). On peut, en effet, retenir le point de vue du lecteur, qui prend connaissance d'un certain travail, d'une part, et le point de vue du chercheur qui opère le travail en question ; les deux « corpus » ainsi délimités ne se recoupent que partiellement : si nous nous définissons en tant que lecteur, le corpus de Vandeloise (1986), par exemple, correspond alors à l'ensemble des phrases constituant l'objet de l'analyse présenté dans l'ouvrage, mais ce n'est sans doute qu'un sous-ensemble (celui que l'auteur a retenu comme pertinent pour l'exposé) de la totalité des exemples effectivement examinés par Vandeloise ; c'est ainsi que Milner (1978) peut écrire :

> « Les exemples, comme il est d'usage dans la grammaire transformationnelle, sont censés valoir pour la classe entière des phrases construites de manière analogue. De façon générale, nous laisserons à l'intuition du lecteur le soin de reconstituer la classe pertinente. »

2.3 Les différentes démarches pour élaborer les corpus

Si l'on adopte le point de vue du chercheur, il y a à nouveau à distinguer entre deux démarches possibles (Fillmore, 1992)[4] : ou bien les hypothèses s'élaborent à partir d'exemples « forgés » (l'« introspection » dans le cadre d'une « linguistique de bureau », Corbin 1980), ou bien le travail s'opère sur des exemples « attestés » (le « corpus » dans le cadre d'une « linguistique de terrain », *Ibidem*) ; dans le premier cas, le linguiste construit lui-même les énoncés, dans le second cas, il les relève dans des textes de divers genres qui n'ont pas été produits pour les besoins de la cause (romans, articles de presse, entretiens radiophoniques, etc.).

2.3.1. CORPUS FORGÉ : AVANTAGES ET DÉSAVANTAGES. L'une des façons pour un linguiste de constituer les données, sur lesquelles il va travailler, repose sur ce que l'on appelle « les corpus forgés » : « corpus basés sur la pratique expérimentale et dynamique qui consiste à utiliser la compétence des locuteurs pour obtenir des données selon les besoins de l'étude » (Riegel *et al.*, 1994). Le linguiste peut alors s'adresser à des informateurs[5] pour savoir quels sont leurs jugements d'acceptabilité sur l'ensemble des énoncés, pour leur faire produire des énoncés et ainsi vérifier la représentativité de ses propres réactions.

a) Le principal avantage de l'exemple forgé est qu'il permet les manipulations dont le linguiste a besoin pour procéder à son analyse et observer celles qui ne sont pas possibles[6] (éventualité peu probable dans les énoncés attestés). Soit, par exemple, l'énoncé *Il est dans les dix heure*s : il peut être soumis à diverses commutations permettant de conclure, rapidement et économiquement (par rapport au temps que représenterait la recherche effective des phrases attestées correspondantes), que la préposition peut se voir substituer *vers* mais non *à, de, pour*…, et que le déterminant est incommutable [**Il est dans (ces + mes + des + quelques + plusieurs) dix heures*]. De même, si l'on cherche quels compléments de verbe *dans* peut introduire, plutôt que de procéder à des relevés dans des textes, il est peut-être plus sûr de tester à partir de la liste fournie par un dictionnaire quels verbes sont susceptibles de se construire avec *dans*, et quelles sont les propriétés permettant de les classer… L'intérêt de cette démarche est qu'elle est relativement objective parce qu'indépendante des aléas des corpus attestés (on peut avoir en effet un article de presse ou une page de roman sans un seul complément en *dans – a fortiori* un complément de type précis que l'on cherche à étudier). De plus, les corpus forgés ne nécessitent pas de longues et fastidieuses manipulations d'exemples, tel que c'est le cas avec les corpus attestés où les phrases sont généralement plus longues et complexes. La constitution de corpus forgés s'avère alors plus souple et plus

économique (en temps et en investissement notamment) que le dépouillement de corpus divers.

b) Le principal désavantage de l'exemple forgé est qu'il est tributaire des jugements d'acceptabilité et de grammaticalité du chercheur (nous reviendrons plus loin sur ces notions), et que ces derniers peuvent être faussés (involontairement) par la prégnance de l'hypothèse que l'on a en tête ; ainsi Melis (2003) considère t-il que *dans les* ne peut introduire un sujet (il met l'astérisque à **Dans les deux cents kilos suffiront* et **Restent dans les trente semaines à planifier*), alors que le lecteur forgera facilement (d'ailleurs précédé par Gross, 1977) entre autres *Dans les trente personnes sont venues*, ou acceptera les énoncés incriminés. De plus, étant donné que nul n'est parfait, on n'est jamais sûr de penser à toutes les possibilités qu'offrent tous les items, et on peut fausser les tests (plus ou moins consciemment) en fonction de l'hypothèse qui se fait jour. Enfin, un autre désavantage des corpus forgés est qu'ils ne permettent pas de décrire (qualitativement et quantitativement) la représentativité des données dans l'usage effectif de la langue : « l'introspection est impuissante à décrire leur [les variations dans les pratiques langagières] distribution dans la population : le social lui échappe par définition » (Corbin, 1980).

2.3.2. CORPUS ATTESTÉ: AVANTAGES ET DÉSAVANTAGES. Les corpus attestés se définissent par le fait que les données ont été produites indépendamment du travail linguistique, qu'elles relèvent de sources diverses (romans, article de presse, etc.) et qu'elles peuvent être de natures diverses (écrites ou orales).

a) Les corpus attestés présentent certains avantages (par rapport aux phrases forgées) : l'auteur (du roman, de l'article de presse, etc.) fait un usage spontané de tel terme ou de telle structure ; il n'y a donc pas de risque que la phrase qu'il produit soit faussée par une hypothèse (d'ordre linguistique) à démontrer : les données n'ont pas été produites pour les besoins de la recherche linguistique, ni suscitées par elle. Elles n'ont ainsi pas subi l'influence du linguiste (comme cela peut se produire lorsqu'il forge ses exemples).

b) L'utilisation de corpus attestés présente toutefois des désavantages : un corpus – si vaste soit-il – ne comporte pas nécessairement toutes les données pertinentes (par exemple toutes les manipulations permettant, dans la suite Verbe + Infinitif, de distinguer entre semi-auxiliaire (*Il va partir*) et verbe distributionnel (*Il désire partir*)). En revanche, on peut y trouver des cas de figure auxquels on n'aurait pas pensé spontanément. Il faut dire aussi qu'un cas de figure représenté dans un corpus attesté peut tout simplement ne pas être remarqué par le chercheur : il y a une longue tradition grammaticale et lexicographique qui s'appuie sur des exemples attestés mais qui, entre

autres, n'a jamais repéré certains emplois de *dans* ; ainsi tous les dictionnaires signalent-ils le sens spatial, le sens temporel, le fait que *dans* puisse introduire un état (*être dans l'embarras*) ou l'approximation (*Il a dans les trente ans*), mais aucun ne mentionne l'interprétation appositive (Leeman 2000 ; Vaguer 2000) que peut prendre *dans ce tableau, JE vois dans ce tableau une preuve de sa folie*, compris comme « Ce tableau est une preuve de sa folie ». Le recours à des corpus attestés ne garantit donc pas à lui seul la complétude ou la représentativité de la description. De plus, l'objectivité qu'ils procurent n'est pas entière. Si le corpus c'est, par exemple, la liste des compléments en *dans* que l'on peut extraire de *Frantext*, c'est un recensement neutre. Ce qui n'est pas neutre, c'est ce que l'on fait de ce recensement : on va opérer une sélection selon ce que l'on cherche à étudier, par exemple les compléments temporels ; on s'éloigne de l'objectivité dans la mesure où c'est le linguiste qui décide de ce qui est (ou non) temporel, et donc fait intervenir une certaine intuition (par conséquent nécessairement une certaine subjectivité) – même s'il applique des critères, le résultat qu'il affecte au test dépend de son sentiment linguistique. Dans ces compléments temporels, on ne va en garder qu'un certain nombre, sur la base là aussi de jugements personnels : on élimine ce qui paraît redondant, du même type ; on garde ce qui semble le plus propre à illustrer ce que l'on veut dire, mais on ne signale pas ce sur quoi on n'a rien de particulier à observer, etc.

2.3.3. CONCLUSION. Dans les deux cas, donc, il y a le risque que le chercheur manque des données pertinentes, du fait que, aussi bien lorsqu'il forge des phrases que lorsqu'il recherche des énoncés attestés, il est plus ou moins inconsciemment guidé par une certaine chose à découvrir, ce que masquent les formulations passives dans les définitions habituellement fournies du corpus. Ainsi, pour Arrivé *et al.* (*op. cit.*) c'est un « ensemble d'énoncés d'une langue donnée (écrits ou oraux enregistrés) qui ont été recueillis pour constituer une base d'observation permettant d'entreprendre la description et l'analyse de la langue en question » – où rien n'est dit sur les critères qui président au « recueil »[7]. Et si des critères sont précisés par Sinclair (1996) : « une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage », ils le sont sur la base de principes avouables (et même garants de scientificité), effaçant tous les risques liés à la subjectivité du linguiste ! Pour Sampson (1994), « la linguistique de corpus prend le langage comme elle le trouve ». Or, on ne prend pas le langage tel qu'il est, même dans une linguistique de corpus, à partir du moment où l'on interprète nécessairement les énoncés (attestés) – ce dont témoignent précisément les différences d'acceptabilité[8]. Ainsi, quand on relève des énoncés, on les trouve attestés par rapport à un contexte donné. Or

qu'en est-il quand les données sont appréhendées hors contexte, par d'autres lecteurs ? Il se peut que ces derniers, confrontés à des phrases isolées, puissent être dans l'incapacité de trancher sur leur acceptabilité ou puissent leurs attribuer un jugement d'acceptabilité différent. La notion d'acceptabilité est fluctuante dans la langue ; on n'est donc jamais sûr que ce qu'une autre personne qualifie d'acceptable le soit pour les mêmes raisons, selon les mêmes facteurs que soi. C'est en cela que l'établissement d'un corpus est toujours marqué de subjectivité car, qu'on le forge ou qu'on le relève, c'est toujours suivant ses propres intérêts de recherche, en ayant déjà une idée dans certains cas du type de structure que l'on cherche. L'objectivité revendiquée par les tenants du corpus attesté n'est qu'apparente, cachant un jugement d'acceptabilité refoulé.

2.4 Les différents jugements que l'on produit sur les données : l'acceptabilité et la grammaticalité

Le jugement que porte le linguiste sur les énoncés est le seul outil dont il dispose pour séparer, dans les données qu'il observe ou qu'il construit, celles qui peuvent fonctionner comme illustration de ce que la langue permet, de celles qui montrent ce que la langue interdit ; c'est à partir de cette base fondamentale que l'on peut saisir les différences entre les deux ordres d'énoncés qui vont justifier la formulation d'une règle : le fait que l'on puisse affirmer qu'en français le déterminant se trouve devant le nom repose sur l'observation que, si est possible (entre autres) *Le soleil brille*, ne le seraient pas *Soleil brille* ni *Soleil le brille* ni *Soleil brille le*. Cependant, pour étiqueter ces possibilités et impossibilités, deux termes existent : les uns parlent de « grammaticalité », les autres d'« acceptabilité », d'autres encore des deux.

2.4.1. Ainsi pour Milner (1978), « le jugement d'acceptabilité est le seul donné sur lequel le linguiste peut raisonner pour construire, en terme de grammaticalité, sa théorie ». Mais ce jugement n'est pas objectif, en ceci qu'il n'est pas porté pour constater le résultat d'une expérimentation mettant en jeu des outils indépendants de lui (comme dans le cas du chimiste qui constate que l'addition de tel acide dans telle solution la fait virer au bleu, ou a pour effet un bouillonnement, etc.) – d'où les tentatives de saisir ce qui est en jeu dans le jugement afin de l'objectiver – d'en faire la base d'un critère reproductible – du fait que les locuteurs n'ont pas forcément les mêmes réactions face à un même énoncé.

2.4.2. Pour Picabia & Zribi-Hertz (1981), « sera dite *grammaticale* dans la langue L, une séquence conforme aux principes et contraintes linguistiques qui constituent la *grammaire* de L ». La grammaticalité relève donc de la compétence. La définition de Picabia & Zribi-Hertz semble totalement

circulaire puisque pour constituer une grammaire, on se fonderait sur le jugement de grammaticalité, et que la grammaticalité, c'est le jugement que les phrases appartiennent à la grammaire ! Autrement dit, on retombe dans le problème posé par le recours à l'acceptabilité : pour élaborer une grammaire, on a besoin de savoir ce qu'est une phrase grammaticale, avant même que la grammaire soit élaborée ; sur quelle base alors décide t-on que la phrase est (a)grammaticale ? Sur une intuition – dont les conditions d'exercice ne sont pas claires : la preuve, on ne fait pas de partage net entre acceptabilité et grammaticalité. Cette confusion (relevée par Normand, 1972) explique que l'on ait souvent reproché aux générativistes de se fonder en fait sur un sentiment linguistique reposant sur la norme (traditionnelle : le « bon » usage) de l'écrit : il est évident que l'on ne construira pas les mêmes règles disant ce qui appartient à la grammaire (qui se confond en l'occurrence avec la langue elle-même), selon que l'on part, par exemple dans le cas de l'interrogation, d'un corpus rassemblant comme phrases grammaticales (à l'exclusion des autres, jugées agrammaticales) :

- ou bien : *Où vas-tu ?* et *Où est-ce que tu vas ?*
- ou bien : *Où vas-tu ?*, *Où tu vas ?*, *Tu vas où ?* et *Où est-ce que tu vas ?*, *Où c'est que tu vas ?*, *C'est où que tu vas ?*

(dans le premier cas, mais non dans le second, *Où tu vas ?* et *C'est où que tu vas ?* entre autres seront jugées non conformes à la grammaire).

2.4.3. Mais pour beaucoup de linguistes, définir cette notion de grammaticalité ne peut se faire sans passer par celle d'acceptabilité, elle en serait d'ailleurs une partie (dans l'acceptable, il y a du grammatical) : ainsi pour Riegel *et al.* (1994) « la grammaticalité ne regrouperait que la partie de l'acceptabilité qui est déterminée par les règles de bonne formation intrinsèque des énoncés », ou chez Picabia & Zribi-Hertz (1981) « la grammaticalité est la composante linguistique de l'acceptabilité ». En fait, l'utilisation des termes paraît renvoyer à des niveaux différents : dans le cas de Milner, le jugement intuitif est dit d'« acceptabilité » ; c'est celui qui permet de trier les formes et d'élaborer par hypothèse une grammaire, laquelle produira des formes (dites, donc, « grammaticales ») ; dans le cas de Riegel *et coll.*, la grammaticalité relève de la structure, tandis que l'acceptabilité a trait aux compatibilités distributionnelles : *Le soleil nage* est grammatical mais inacceptable, *Soleil le brille* est agrammatical. Le problème est dans la circularité de la démarche : si l'on juge *Est-ce que le soleil brille-t-il ?* inacceptable, on construira une grammaire de telle sorte qu'elle ne produise pas cette séquence (dite, donc, agrammaticale).

3. LA CRÉATION D'UN « CORPUS INFORMATISÉ » : UNE BASE DE DONNÉES LINGUISTIQUE

3.1 Corpus forgé ou corpus attesté ?

Dans le cadre de notre recherche, centrée sur la préposition *dans*, corpus forgé et corpus attesté sont complémentaires et non concurrents. Les énoncés attestés viendront de sources diverses (essentiellement écrites) et les énoncés forgés émaneront des manipulations produites sur ces données attestées.

3.1.1. Les phrases forgées permettent le test rapide et économique des propriétés (que l'on souhaite aussi représentatif que possible), par exemple concernant l'association d'un verbe à un complément en *dans*, et des propriétés syntaxiques que possèdent l'énoncé ainsi construit (par exemple *dans la fuite* n'est ni supprimable ni déplaçable dans *La solution est dans la fuite*). De plus, elles permettent de pallier les « trous » éventuels (ou inéluctables) des corpus attestés (ainsi, il y a peu de chance *a priori* que l'on puisse constituer, à partir des corpus attestés disponibles, la liste des verbes susceptibles de se construire avec *dans*), et de construire des associations agrammaticales ou inacceptables qui, comparées aux suites recevables, sont susceptibles de donner des idées d'hypothèse pour caractériser le complément dont on s'occupe. Notre objectif premier est d'étudier la langue, c'est en cela que nous nous rapprochons davantage du champ harrissien et que nous nous éloignons du champ chomskyen.

3.1.2. Les extraits attestés permettent de vérifier ou d'amender les propositions de description ou d'explication, avancées à partir de corpus forgés, et d'en pallier les manques (en portant à l'observation des énoncés auxquels le chercheur ne pense pas spontanément), donc de limiter le risque de circularité (lorsque le chercheur muni d'une certaine hypothèse secrète les observables qui vont dans le même sens). En effet, ainsi que le signale Blanche-Benveniste (2000), « comme ils [les corpus] contiennent des données attestées, dont on peut vérifier les sources, ils engagent à faire un travail d'analyse linguistique qui ne repose pas uniquement sur l'intuition mais sur la confrontation avec des données parfois étonnantes, que la simple intuition n'aurait pas pu atteindre ».

3.1.3. L'objectif de notre recherche n'est pas de quantifier les emplois de la préposition *dans*[9], mais bien d'avancer dans son identification syntaxique et sémantique. Notre objectif est donc descriptif et concerne la caractérisation de *dans* en langue, et non pas en discours (c'est-à-dire dans les productions orales ou écrites telles que rassemblées dans les corpus attestés) ; autrement dit, il ne s'agit pas de voir comment les locuteurs utilisent *dans* (ou tel type

de complément en *dans*) : plutôt à l'oral qu'à l'écrit ou inversement, plutôt dans la description que dans la narration ou l'argumentation, ou réciproquement, plutôt dans tel genre que dans tel autre, etc., ni donc de voir quel emploi est le plus représentatif ou le plus fréquent dans les performances. Il s'agit de déterminer à quels différents emplois de *dans* on a affaire dans les discours (seuls observables : les actualisations de la langue sont le passage obligé de tout travail linguistique, comme on l'a vu précédemment), de façon à essayer de construire une identité de la préposition en langue permettant, en retour, de rendre compte des énoncés concrets dans lesquels elle apparaît. Le corpus est donc une base incontournable : ce à partir de quoi on peut avoir un aperçu des différentes possibilités qui guident la recherche d'une définition, ou qui en permettent la vérification lorsqu'on a élaboré une hypothèse, mais qui n'est pas en lui-même l'objet de la recherche (notre objectif n'est pas l'analyse des discours). C'est en cela qu'on se rapproche de la « linguistique de corpus » entendue comme : (a) « le travail que fait le linguiste qui constitue un corpus », c'est-à-dire qui prend un texte (au sens large : écrit/oral transcrit, etc.), l'annote par l'ajout d'informations d'ordre morphologique, syntaxique, sémantique… et le traite informatiquement (étiquetages, arbres, analyseurs syntaxiques) pour le rendre utilisable par d'autres (outil d'exploration) puisque nous constituons un corpus (informatisé), et non en tant que (b) le corpus serait l'objet même de notre étude – puisque ce qui nous intéresse c'est un « fait de langue ». En effet, si l'on se reporte à l'opposition saussurienne langue/parole, reformulée en langue/discours, le corpus tel que défini en (a) est un discours (traité informatiquement), qui nous intéresse en tant qu'il manifeste des emplois (effectifs), en tant qu'il témoigne d'emplois possibles (attestés). Mais notre objectif n'est pas de rendre compte des emplois dans tel ou tel corpus (si étendu soit-il) : on cherche à saisir l'identité de la préposition *dans* en langue, identité formelle et sémantique censée présider aux / déterminer les multiples actualisations en discours. Donc le corpus n'est pas l'objet même de notre recherche (puisqu'on ne cherche pas à décrire un corpus), il n'en est que l'outil (incontournable, certes).

3.2 La démarche adoptée

En ce qui concerne la complémentation verbale, la démarche de constitution du corpus a consisté à se donner dans un premier temps une définition syntaxique (à l'aide de propriétés formelles, donc) du complément de verbe, en tant qu'il s'oppose à l'ajout d'une part, au complément dit « de phrase » d'autre part ; nous avons procédé ici essentiellement à un travail de documentation mettant en jeu des phrases forgées par les auteurs consultés : Bonami (1999), Delaveau (2001), Dubois-Charlier (2001), etc. (donc en un

sens attestées, puisque produites par d'autres que nous-même). Puis, à partir d'une liste de verbes, mentionnés comme étant susceptibles de se construire avec la préposition *dans* (*cf.* l'index de Dugas et Manseau, 1996), nous avons cherché des attestations de ces différentes combinaisons dans des bases de données telles que *Frantext*, *Glossanet*, etc. Les verbes signalés par Dugas et Manseau (1996) n'étant pas tous l'objet d'une attestation, nous avons complété le corpus attesté par des phrases forgées. Enfin, sur le corpus de phrases attestées et forgées ainsi rassemblé, nous avons procédé au test des propriétés retenues pour distinguer entre complément de verbe et ajout, donc nous avons forgé un corpus de phrases (qui correspond aux résultats de l'application des critères).

3.3 Le recours à une base de données

La constitution d'une base de données pour rassembler son corpus[10] n'est pas une pratique naturelle en linguistique (entendue comme non spécialisée en TAL). Ainsi, nous mettons ici en évidence les apports de ce type de traitement et leurs avantages.

3.3.1. AVANTAGES GÉNÉRAUX, POUR LA RECHERCHE, DE LA CONSTRUCTION D'UNE BASE DE DONNÉES

a) L'intérêt pour le chercheur lui-même dans la gestion de son propre travail :

- la construction d'une base de données permet une perspective cumulative, donc de ne pas recommencer à constituer un corpus à chaque nouvelle recherche[11], et évolutive tant par sa structure (modulable) que par son contenu : on peut insérer ainsi des données à volonté (ajouter des informations sans cesse), qui peuvent être modifiées en fonction des usages. On peut ainsi l'améliorer, l'affiner pour finalement obtenir ce que l'on souhaite exactement.

- la souplesse : une base de données, une fois que sa structure est bien définie[12], est plus souple et plus puissante qu'une simple liste sur papier, dans *Word*, dans *Excel*… car elle permet notamment des mises à jour, constantes et en cascade, de données identiques mais enregistrées à différents endroits par exemple.

- le stockage et l'organisation des données : la base de données permet de stocker une quantité quasi illimitée d'informations (ce qui n'est pas négligeable quand on sait le nombre de manipulations que l'on effectue sur un corpus) et elle permet d'organiser des informations de façon significative : ainsi, on peut avoir sous les yeux toutes les données associées à un énoncé (sa source, ses analyses formelle et sémantique, les classes distributionnelles…). Elle contient donc le corpus avec des informations différentes : contextuelle, syntaxique, lexicale, sémantique… Elle permet en

quelque sorte de mieux voir les données (on peut proposer des vues sur les données, par exemple, le nombre d'enregistrements pour tel verbe…). Mais ce point de vue reste celui de la personne qui constitue la base de données. Ainsi ce qui compte pour établir une base de données, c'est de savoir ce que l'on souhaite en faire.

- le traitement automatique : la base de données permet aussi de récupérer des informations selon des critères de sélection (par exemple, on peut extraire la liste de tous les noms présents dans les SN introduits par *dans*, on peut aussi sélectionner tous les énoncés issus d'une même source (c'est ce que permet la table « Source », notamment si l'on veut faire une étude sur un journal particulier, sur un auteur particulier… Le traitement automatique des requêtes évite donc des manipulations fastidieuses à la main et offre un gain de temps qui permet d'approfondir la recherche et de mieux voir d'un coup d'œil les régularités. Il permet également d'opérer des analyses quantitatives (par les requêtes, les tris, les décomptes…), qui permettent alors de s'insérer davantage dans le courant de la linguistique de corpus. On peut, par exemple, s'interroger sur le type de nom qui est le plus fréquemment employé avec *dans* et les verbes de mouvement.

b) L'intérêt scientifique d'une circulation de la recherche et le fait qu'il existe relativement peu de corpus électroniques disponibles.

On peut diffuser l'information contenue dans une base de données : chaque linguiste, face à la spécificité de sa recherche, forge son propre corpus. Par la saisie d'un corpus dans une base de données, nous souhaitons rendre celui-ci accessible à d'autres linguistes pour plusieurs raisons : la première, c'est que nous nous sommes rendue compte, au fil de nos lectures (articles, revues… de linguistique), que nous n'avons pas accès aux corpus sur lesquels ces écrits ont été produits. Or cela nous aurait permis de vérifier les dires de certains linguistes, de compléter leur analyse sur le même corpus de base et de le compléter par de nouveaux énoncés pour confirmer, ou infirmer, ces dires. Ainsi, nous n'avons pas accès aux corpus analysés par Gross par exemple, or, il nous a semblé que certaines de ses analyses et conclusions n'étaient pas tout à fait exactes, mais seulement par rapport aux extraits de corpus qu'il nous donne et par rapport à notre propre corpus. La théorie n'est donc pas reproductible puisqu'on peut ne pas arriver aux mêmes conclusions. Chaque corpus construit par un linguiste meurt donc avec lui. Tant d'heures de recherche d'occurrences qui se perdent… L'accès au corpus des autres linguistes permettrait de gagner du temps et d'approfondir davantage la recherche. Ainsi, ce que nous avons recueilli pour la préposition *dans* peut servir à d'autres linguistes, leur permettre de mettre en évidence d'autres phénomènes que nous n'avons pas analysés (par exemple, quelqu'un qui travaille sur les temps grammaticaux pourra peut-

être y trouver des choses). De ce fait les informations contenues dans une base de données sont consultables et réutilisables par d'autres personnes.

De plus, il existe en France très peu de corpus électroniques disponibles[13], facilement accessibles (sur le français) qui puissent nous aider dans l'établissement de notre corpus d'étude. Il suffit pour s'en rendre compte de faire une recherche sur le Web avec le mot-clef « corpus » ou « base de données linguistique » ou « corpus linguistique » (les résultats sont probants !). Actuellement, le concordancier en ligne *GlossaNet*, le *TLFi*, le *Web*, le *Dictionnaire de l'Académie française*, *ABU* : la Bibliothèque Universelle, le site *Elicop* (Étude Linguistique de la Communication Parlée) sont disponibles et accessibles gratuitement alors que *Frantext*, *Le Monde Diplomatique*, *Le Petit Robert Multimédia* (ou autres corpus sur CD-Rom) restent sous le coût d'une licence (donc payants). Malheureusement, les corpus actuellement accessibles sont peu diversifiés (beaucoup sont centrés sur la littérature) ; ainsi, seul *GlossaNet* permet d'oublier pour un temps la recherche d'occurrences dans la presse munie d'un crayon !

3.3.2. AVANTAGES DE LA BASE DE DONNÉES LINGUISTIQUE ICI CONSTITUÉE : LE CORPUS EST CONSTITUÉ D'ÉNONCÉS MUNIS DE LEUR ANALYSE. Par l'informatisation de notre corpus, nous nous inscrivons dans le courant des linguistiques « de corpus » qui consiste en « l'utilisation de corpus annotés, de grande taille, variés et assortis d'outils d'exploration puissants, permettant d'observer plus finement les phénomènes » (Habert *et al.*, 1997). Par rapport aux faits, nous définirons notre corpus comme un regroupement de phrases isolées les unes des autres (absence de paragraphes, de textes…), mais ayant en commun l'usage de la préposition *dans*. Ces énoncés sont issus de sources différentes (presse, littérature… nous n'avons pas voulu distinguer des niveaux de langue différents et, par exemple, ne travailler que sur du « littéraire », ou que sur du « journalistique ») et récoltés de deux façons : la première reste traditionnelle – la lecture minutieuse armé d'un crayon pour relever ce qui nous semble pertinent. La seconde repose sur l'utilisation du concordancier *GlossaNet* : après avoir saisi nos requêtes de type [<dissoudre> dans] dans notre profil *GlossaNet* (notre recherche étant centrée sur la complémentation verbale en *dans*, nous souhaitions extraire des journaux une liste d'occurrences comportant les verbes se construisant avec cette préposition), le résultat de l'extraction nous était envoyé par courriel, il ne nous restait plus donc qu'à l'analyser et à saisir les occurrences dans notre base de données en suivant toujours la même procédure (les liens existants entre les tables, *cf*. Vaguer 2004) :

Etape 1 : Saisie dans la table « Source » de la provenance des énoncés récoltés (année, [auteur, titre], [journal, type de support : informatique, papier]) ;

Etape 2 : Saisie dans la table « Précisions sur la Source » pour spécifier l'article consulté, la page, le genre…

Etape 3 : La table « Identification distributionnelle du SP » contient l'énoncé retenu, ainsi que les propriétés syntaxiques du complément introduit par *dans* (les manipulations traditionnellement jugées pertinentes pour en permettre l'identification sont ici représentées : suppression, détachement, position préverbale, pronominalisation, test en *le faire*, entre autres) ;

Etape 4 : La table « Identification des constituants V, dét, N » permet de saisir chacun des constituants (en vue d'extraction automatique, par exemple, de l'ensemble des noms) et contient les conclusions de l'identification syntaxique du complément : est-il complément ou modifieur ?

Etape 5 : La table « Propriétés des Noms » permet une première analyse du nom en terme de classes d'objets, classe sémantique ou par ses propriétés morphologiques : est-il dérivé d'un verbe ?

À l'aide de cet échantillon d'emplois de la préposition *dans* (que nous souhaitons représentatif de l'ensemble de ses emplois en discours), nous avons pu mettre en évidence (Vaguer, 2004b) des régularités quant à l'utilisation de cette préposition, par le biais de manipulations réglées, et avancer ainsi dans son identification. Notre corpus comporte donc les énoncés de départ, mais aussi toutes les indications qui leur sont associées, tant du point de vue de la provenance de l'énoncé (source : auteur, genre, année, page…) que du point de vue de l'analyse de l'énoncé lui-même : son analyse syntaxique par le biais de manipulations (quel type de constituant, quelle structure de phrase, quelle fonction des constituants dans la phrase, etc.) et l'analyse de chacun de ses constituants (à quelle classe distributionnelle appartiennent-ils ?), son identité sémantique (locatif, approximatif…). À l'heure actuelle, notre base de données (nommée Zéphyr-V, V comme Verbe) rassemble 1 200 énoncés pourvus de leurs analyses syntaxique, lexicale et sémantique.

4. RÉFÉRENCES

Arrivé M.; Gadet F.; Galmiche M. 1986. *La grammaire d'aujourd'hui : guide alphabétique de linguistique française.* Paris : Flammarion.

Blanche-Benveniste, C. 2000. « Corpus de français parlé » *in* Bilger, M. (éd). *Corpus. Méthodologie et applications linguistiques*. Paris : Honoré Champion et PUP. (p. 15-25).

Bonami O. 1999. *Les constructions du verbe : le cas des groupes prépositionnels argumentaux*. Paris. Thèse de l'Université Paris VII.

Builles J.-M. 1998. *Manuel de linguistique descriptive. Le point de vue fonctionnaliste*. Paris : Nathan.

Chomsky N. 1969. *Structures syntaxiques*. Paris : Le Seuil.

Chomsky N. 1971. *Aspects de la théorie syntaxique*. Paris : Le Seuil.

Corbin P. 1980. « De la production des données en linguistique introspective ». *Théories linguistiques et traditions grammaticales*. Villeneuve-d'Asq : PU de Lille. (p. 121-179).
Delaveau A. 2001. *Syntaxe. La phrase et la subordination*. Armand Colin, Coll. Campus.
Dubois J.; Giacomo M.; Guespin L. 1999. *Dictionnaire de linguistique et des Sciences du langage*. Paris : Larousse (1ère éd. 1994).
Dubois-Charlier F. 2001. « Compléments de Verbe, de Proposition, de Phrase, d'Énoncé ». *Adverbe et Circonstant*. CLAIX. n°17. Aix-en-Provence : PUP. (p. 33-50).
Dugas A., Manseau H. 1996. *Les verbes logiques*. Montréal : Éditions Logiques.
Fillmore C. J. 1992. « "Corpus linguistics" or "Computer-aided armchair linguistics" » *in* Svartvik, J. (éd). *Directions in Corpus Linguistics*. number 65. Berlin : Mouton de Gruyter. (p. 35-59).
Gasiglia N. 2003. « Réflexions autour des coûts et bénéfices pour un linguiste qui recourt à des ressources électroniques et des outils informatiques dédiés à leur dépouillement : le cas d'une étude lexicale relative aux mots du football ». *Pré actes des 3èmes Journées de la linguistique de corpus*. Lorient (11-13/09/03). France.
Gleason H.-A. 1969. *Introduction à la linguistique*. Paris : Larousse.
Gross M. 1977. *Grammaire transformationnelle du français. Syntaxe du nom*. Paris : ASSTRIL.
Habert B.; Nazarenko A.; Salem A. 1997. *Les linguistiques de corpus*. Paris : Armand Colin / Masson.
Habert B. 2002. « Outiller les linguistes/outiller la linguistique : par où, par qui commencer ? ». Intervention à la table ronde *TAL et enseignement*. TALN'02 Nancy. 24/06/02.
http://www.limsi.fr/Individu/habert/Cours/PX/BHabertOutillerLaLinguistiqueTableRondeTALN02.pdf.
Leeman D. 2000. « Compléments circonstanciels ou appositions ? ». *Langue française*. n°125. Paris : Larousse. (p. 19-29).
Melis l. 2003. « Le groupe prépositif comme déterminant du nom » *in* Haderman, P., Van Slijcke, A., Berré, M. (éds). *La syntaxe raisonnée – Mélanges de linguistique générale et française offerts à Annie Boone*. Bruxelles/Paris : De Boeck/Duculot. (p. 235-250).
Mellet S. 2002. « Corpus et recherches linguistiques. Introduction ». *Corpus*. n°1. Nice : Publications de la Faculté des Lettres, Arts et Sciences humaines de Nice. (p. 5-12).
Milner J.-C. 1978. *De la syntaxe à l'interprétation. Quantités, insultes, exclamations*. Paris : Éditions du Seuil.
Normand C. 1972. « De quelques notions fondamentales (sur un enseignement d'initiation à la linguistique) ». *Langue française*. n°14. Paris : Larousse. (p. 32-56).
Picabia L., Zribi-Hertz A. 1981. *Découvrir la grammaire française. Une introduction active à la linguistique française et générale*. Paris : CEDIC.
Riegel M.; Pellat J.-C.; Rioul R. 1994. *Grammaire méthodique du français*. Paris : PUF.

Sampson J. 1994. « Susanne : a domesday book of english grammar » *in* Oostdijk, N., De Haan, P. (éds). *Corpus Based Research into Language*. Amsterdam : Rodopi. (p. 169-187).
Saussure F. de. 1972. *Cours de linguistique générale*. Paris : Payot. (1ère éd. 1916).
Sinclair J. 1996. *Preliminary recommendations on Corpus Typology*. Rapport Technique. EAGLES (Expert Advisory Group on Language Engineering Standards). CEE.
Vaguer C. 2000. *Il s'est trompé dans l'administration du médicament. Un ou des compléments de structure : « Dans + N_{action} » ? Naissance de la notion "complément d'apposition"*. Mémoire de DEA. Université de Paris X–Nanterre.
Vaguer C. 2004a. « Constitution d'une base de données : les emplois de *dans* marquant la "coïncidence" ». *Revue Française de Linguistique Appliquée*. IX-1. (p. 83-97).
Vaguer C. 2004b. *Les constructions verbales "V* dans *GN". Approches syntaxique, lexicale et sémantique*. Thèse de doctorat. Université de Paris X–Nanterre.
Vaguer C. 2005a. « 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. (p. 134-138).
Vaguer C. 2005b. « De l'utilité d'un corpus en syntaxe, mais quel corpus ? ». *in* Vergely P (éd.). *Rôle et place des corpus en linguistique.* Actes du Colloque JETOU'2005. (p. 101-114).
Vandeloise C. 1986. *L'espace en français*. Paris : Le Seuil.

3. NOTES

(1) Tel que Dubois *et al.* (1999, p. 123) le définissent : « L'*univers* est l'ensemble des énoncés tenus dans une circonstance donnée, tant que le chercheur n'a pas décidé si ces énoncés entraient en totalité ou en partie dans la matière de sa recherche ».
(2) La représentativité est pour Gleason (1969, p. 158) un des problèmes essentiels liés à la constitution et à l'utilisation d'un corpus de matériaux, fournis par un ou plusieurs informateurs, et à partir duquel le linguiste doit écrire sa description de la langue. Le problème lié à la représentativité d'un « échantillon » de langue que forme le corpus, c'est que « certains traits grammaticaux ne sont pas fréquents ; ils risquent de ne pas être représentés de façon valable dans un corpus réuni au petit bonheur. D'autres traits, au contraire, sont très courants : même une quantité restreinte de matériaux suffit à les illustrer bien plus qu'il n'est nécessaire pour établir ou confirmer une analyse ». Normand (1972, p. 34) résumait ainsi les propos de Gleason : « des traits importants de la langue peuvent ne pas être représentés et des traits ordinaires l'être trop souvent ».
(3) « Un corpus ne peut être clos et exhaustif que dans le cadre d'une monographie… Il sera étudié en tant que tel, sans pouvoir prétendre à être représentatif d'autre chose que de lui-même ni à ouvrir sur aucune forme de généralisation ou modélisation » (Mellet 2002, p. 6).
(4) *Cf.* la caricature proposée par Fillmore (1992, p. 35): « Armchair linguistics does not have a good name in some linguistics circles. A caricature of the armchair linguist is something like this. He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, "Wow, what a neat fact !", grabs his pencil, and writes something down. Then he paces around for new hours in the excitement of having come still closer to knowing what language is really like. (There isn't anybody exactly like this, but there are some approximations.) Corpus

linguistics does not have a good name in some linguistics circles. A caricature of the corpus linguist is something like this. He has all the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus as the second word of a sentence. (There isn't anybody exactly like this, but there are some approximations) ». L'idéal pour Fillmore serait que les deux types de linguistes soient réunis en un seul homme.

(5) « En français, le terme *informateur* peut prêter à confusion : il fait souvent penser à un indicateur, c'est-à-dire à quelqu'un qui fournit des renseignements à la police ou à un autre service plus ou moins officiel… En anglais, la confusion n'existe pas car il existe deux termes distincts : *informant* (celui qui fournit des renseignements à la police) et *informer* (celui qui fournit des renseignements à un journaliste, à un linguiste, etc.) » (Builles 1998, p. 60).

(6) L'emploi de corpus forgés permet au linguiste d'avoir « la langue accessible à travers une série toujours ouverte de nouveaux énoncés, spontanés ou provoqués… » (Riegel *et al.* 1994, p. 19). Et d'un point de vue quantitatif, le fait d'avoir accès à la langue dans son ensemble, et non uniquement à un échantillon (comme c'est le cas avec les corpus attestés), offre d'autres possibilités : « N'étant plus limités en nombre, les échantillons de performance étayent les hypothèses sur la langue, mais permettent aussi leurs vérifications en les confrontant à de nouvelles données » (*Ibidem*).

(7) Insistons sur le fait que rien n'est dit sur le recueil des données (comment on procède, sur quoi on opère, sur quels types de données). Finalement, la notion de « corpus » semble acquise et admise par l'ensemble des linguistes, qui l'emploient sans juger utile de la définir, comme allant de soi : la consultation de différents ouvrages (dont l'analyse est proposée ici) nous a permis d'observer que cette notion est souvent esquivée, ou non explicitée.

(8) « Moi, je suis de la France. Je ne dis pas : je suis la France. Je suis de la France. Toutes mes pensées, toutes mes façons d'être, toutes mes sensations, toutes mes vibrations, elles sont de la France » (Habert *et al.*, 1997, p. 9). Cet exemple extrait du corpus *Mitterand1* met bien en évidence qu'il n'est pas facile d'établir des distinctions tranchées entre les réalisations langagières jugées acceptables et celles jugées non-acceptables, puisque les constructions employées par F. Mitterand paraissent pour certaines agrammaticales. Or le Président les a employées et son insistance montre qu'il est conscient des structures énoncées (elles ne relèvent pas du lapsus).

(9) Comme cela se fait dans les recherches actuelles en linguistique de corpus : cf. Habert *et al.* (1997) et plus récemment les communications de Gasiglia, Arnaud, Alves, Fujimura, Manguin… aux 3èmes Journées de la Linguistique de Corpus (Lorient, septembre 2003).

(10) Nous entendons par *corpus*, une banque de données ouvertes qui sera alimentée et étoffée régulièrement en fonction des exemples rencontrés et des préoccupations de recherches. Notre corpus sera donc centré sur des énoncés constitués de la préposition *dans* et on le jugera saturé pour des raisons matérielles au moment de finaliser notre thèse.

(11) Il faut, pour ce faire, bien entendu travailler sur le même sujet.

(12) Notons, toutefois, qu'une base de données nécessite un travail long et fastidieux de mise en place : en effet, il faut, dans un premier temps, définir quels sont les éléments que l'on veut y voir figurer et comment on souhaite que cela s'organise (la mise en place de liens entre les tables n'est pas évidente). Mais cette formalisation permet d'avancer dans la compréhension du phénomène étudié puisqu'il faut, à ce moment-là, se demander ce qu'on cherche à mettre en évidence, ce qu'on veut voir apparaître, etc. Si des efforts sont investis dans la constitution d'une base de données, il y a ensuite un « retour sur investissement » (Habert, 2002) non négligeable.

(13) Or, tout linguiste travaillant sur un « corpus » (comme nous l'avons mentionné en première partie de cet article), il y a beaucoup de données riches qui restent inaccessibles.