**Marcel Cori**
(MoDyCo, Université Paris10, CNRS)
**Sophie David**
(MoDyCo, CNRS, Université Paris10)

# Les corpus fondent-ils une nouvelle linguistique ?

## INTRODUCTION

Il existe actuellement un courant de pensée, aux contours non précisément définis, selon lequel l'usage des corpus (constitués par le rassemblement (de millions) de données langagières effectivement attestées), serait à l'origine d'un renouvellement profond en linguistique, remettant en cause tout un pan de la recherche en ce domaine. Une partie de ce courant s'inscrit sous l'étiquette de *linguistique(s) de corpus* – ou de *Corpus Linguistics* – mais certains auteurs, qui posent d'une manière similaire le problème de la collecte et de l'analyse des données, et par suite le problème de la définition de l'objet de la linguistique, n'emploient pas le terme. C'est pourquoi nous préférerons faire référence à ce courant par le terme d'*approche sur corpus*.

Dans cet article, nous présentons les principaux arguments développés par ce courant, en essayant de les articuler afin de construire une position cohérente, même si cette position n'est pas celle d'un auteur en particulier. La discussion de ces positions permettra de revenir sur certaines questions fondamentales qui se posent à la recherche en linguistique[1].

## 1. LES POSITIONS

Nous étudions tout d'abord ce que les auteurs qui se réclament de la linguistique de corpus ou de la *Corpus Linguistics* entendent par ces termes. Nous développons ensuite les arguments avancés par les tenants des approches sur corpus, en particulier ceux qui nous semblent converger pour former une position théorique à peu près cohérente. Ces arguments peuvent être regroupés en trois séries : la première a trait à la constitution des données (§ 1.2.) ; la deuxième consiste en la remise en cause de l'approche chomskyenne, mais aussi plus largement de ce qui est rassemblé sous le terme de « linguistique

1. Nous remercions Françoise Kerleroux pour sa relecture d'une première version de cet article.

introspective » (§ 1.3.) ; la troisième se prévaut d'une prise en compte de nouveaux champs, ou d'une meilleure prise en compte de champs déjà considérés (§ 1.4.).

## 1.1. L'emploi du terme « linguistique de corpus »

En Grande-Bretagne, le terme *Corpus Linguistics* désigne le programme de recherche de Sinclair (1996), ou bien celui de Leech (1992)[2]. Leech (1992 : 106) emploie également le terme de *Computational Corpus Linguistics* (CCL) après avoir noté les liens particulièrement étroits entre la *Corpus Linguistics* et l'usage des ordinateurs. Pour lui, la *Computer Corpus Linguistics* n'est plus seulement une méthode, mais « a new research enterprise, and in fact **a new philosophical approach** to the subject ». Chez d'autres auteurs, notamment américains, le terme de *Corpus Linguistics* n'apparaît pas, mais l'accent est mis très fortement sur le renouvellement des méthodes en linguistique.

En France, l'usage du terme « linguistique de corpus » s'est étendu, ceux qui l'emploient situant ou non leur démarche par rapport à la tradition britannique. On peut citer notamment Habert *et al.* 1997, la revue AFLA (1999), les articles rassemblés par Bilger (2000), ou par Williams (2005). Cependant, rares sont ceux qui prennent la peine de définir précisément le terme de *linguistique de corpus*, la plupart des travaux faisant comme si son objet et ses fondements allaient de soi. Quand la linguistique de corpus est définie, elle constitue pour certains une discipline : « La linguistique de corpus est une **discipline** qui s'est développée à partir des années dix-neuf cent quatre-vingts mais qui a vraiment pris son essor avec l'arrivée sur le marché d'ordinateurs personnels. Largement issue de l'école contextualiste, la **discipline** s'intéresse à la langue en contexte sous la forme de grands ensembles de textes, les corpus. C'est une **discipline** qui est très liée à l'utilisation de l'informatique, mais qui reste une **discipline** des sciences humaines et non de l'informatique. Les maîtres mots sont linguistique et corpus » (Williams 2005 : 13)[3].

On notera à l'inverse la grande prudence de Habert *et al*. (1997), qui, malgré le titre de leur ouvrage, ne disent nulle part que la « linguistique de corpus » constitue une discipline ; et l'emploi du pluriel est un signe de cette prudence. Pluriel qui rend bien compte de la diversité des champs et des approches proposés.

En fin de compte, ce qui apparaît, c'est une volonté de regrouper des travaux s'inscrivant dans des positionnements conceptuels différents :

– Se trouvent ainsi rassemblés des travaux fondés sur le plan méthodologique ou bien sur le plan théorique. Ce que note bien Tognini-Bonelli (2001) par le sous-titre de son introduction : *Corpus Linguistics : between theory and methodology*.

2. Voir l'article de Jacqueline Léon dans ce numéro.

3. Dans la partie 1., tous les éléments en gras dans les citations ont été soulignés par nous.

– Se trouvent ainsi rassemblés des travaux dans lesquels la part faite au Traitement automatique des langues varie, ce que recouvre partiellement l'opposition entre « corpus-based » et « corpus-driven », construite par les écoles anglaises.

On ne s'étonnera pas alors que la linguistique de corpus « ne constitue aucunement un domaine de recherche unifié » (Rastier 2005 : 31).

Cependant, c'est la valorisation de cette diversité qui est retenue ultimement, au nom de la constitution de la discipline : « […] il faut accepter que pour construire une discipline en France, il faille inclure une variété d'approches de la langue qui va prendre en compte les tendances et l'histoire de la linguistique en langue française […] La diversité est la clé de voûte de notre discipline » (Williams 2005 : 14-15).

Cette diversité a un coût : celui de l'impossibilité de constituer une définition unifiée de la notion de « corpus ». L'introduction de Williams (2005 : 13-14), qui présente les différents corpus utilisés par les auteurs des textes qu'il a rassemblés, est de ce point de vue exemplaire : il définit les corpus comme « des ensembles de textes choisis et ordonnés selon des critères précis » (référence est faite aux travaux de Atkins *et al*. 1992, Biber 1993, Sinclair 1996, Rastier 2005), pour refuser ensuite « une définition trop stricte [qui] exclurait les corpus littéraires et les corpus non constitués. La notion de corpus doit pouvoir englober des corpus monosource, basés sur des journaux par exemple […] ». Les différents qualificatifs associés au terme *corpus* montrent aussi l'impossibilité d'une définition unifiée : échantillonné/de référence/nu/annoté/comparable/parallèle/aligné/représentatif/de suivi, etc. (pour une présentation hiérarchisée, voir Blanche-Benveniste 2000 ; Habert 2000). Quant à considérer que ce qu'on trouve sur la Toile est un corpus ou pourrait faire partie d'un corpus, là encore les avis divergent. Si, malgré tout, on cherchait une définition unifiée, celle-ci ne pourrait viser qu'un plan purement technique, celui du support électronique. C'est la position finalement adoptée par Williams (2005 : 14) pour caractériser les corpus utilisés dans les études qu'il a rassemblées : « le point commun [est] qu'il s'agit de textes stockés électroniquement sous un format texte ».

On peut enfin remarquer que la définition de la discipline se donne souvent par opposition. Par exemple, Rastier (2005) reprend le terme de « linguistique de corpus » en indiquant : « Au milieu des années 1990, la linguistique de corpus, issue des humanités (et notamment de la collectivité réunie autour de la revue *Computers and the Humanities*), l'a emporté sur le courant orthodoxe-néochomskien de la linguistique computationnelle » (Rastier 2005 : 41). Jacques (2005 : 21) définit elle aussi la linguistique de corpus par opposition à une autre linguistique, qu'elle appelle « la linguistique introspective »[4]. Ce mode

4. Tout comme le terme de « linguistique de corpus » se trouve employé à rassembler des approches différentes, voire contradictoires, la « linguistique introspective » rassemble elle aussi des approches très différentes. On peut se demander si on n'est pas en présence d'un strict effet d'étiquetage, qui fait exister « quelque chose », mais « quelque chose » qu'on ne sait ou qu'on ne peut précisément définir.

d'exposition par opposition est aussi utilisé dans la présentation des arguments avancés, ces derniers étant plus négatifs que positifs (§ 1.2., 1.3.).

## 1.2. Les données

Un premier ensemble d'arguments porte sur ce que la linguistique a à reconnaître comme données (on emploie ici le terme de *donnée* dans un sens non technique ; le terme couvre dans cette section le « matériau » linguistique sur lequel les linguistes travaillent).

La description des pratiques supposées des linguistes quant au recueil des données repose sur les affirmations suivantes :

a) Les linguistes travaillent sur un (tout) petit nombre de données, trop peu de données. Ainsi, pour la syntaxe, Abney (1996 : 4) note : « Current theoretical syntax gives a systematic account of **a very narrow class of data** ». Plus généralement, Manning (2003 : 195) remarque : « Traditional fields linguists have been constrained by **quite small** collections of texts and conversations ».

b) Le reproche qui suit vise leur caractère « inventé » et ses conséquences : « Generative approaches **resort to inventing the primary data**, on the basis of intuitions. » (Manning 2003 : 295). Les données travaillées ne sont pas alors connectées à un usage réel : « […] as a result categorical judgments are overused where not appropriate, while a lack of concern for observational adequacy has meant that successive versions have tended to treat a shrinking subset of data **increasingly removed from real usage** » (Manning 2003 : 296). Ces données finissent par constituer des « stilted linguist's sentences » (Manning 2003 : 300), elles sont jugées non représentatives et non vérifiables : « […] generative grammar has produced many explanatory hypotheses of considerable depth, but is increasingly failing because its hypotheses are disconnected from **verifiable linguistic data** » (Manning 2003 : 296)[5]. Elles ne peuvent constituer des « tests » pour les théories. Leech (1992 : 111) insiste ainsi sur la nécessité de maintenir : « the empirical emphasis on testatibility against independently obtained data ».

c) Enfin, dernière caractéristique, les données travaillées sont « ajustées » au nom de différents principes de simplicité : « Syntacticians focus on the data that they believe the fewest complications factors, and **« clean up » the data** to remove what they believe to be remaining complications that obscure simple, general principles of language » (Abney 1996 : 11). Conséquence, constituée en reproche ou bien en trait distinctif : les linguistes travaillent sur des données artificielles : « […] the simple principles we arrive at are **artifacts of our data selection and data adjustment** » (Abney 1996 : 11) ; les données sont « des **productions artificielles** produites par l'introspection des linguistes », à la diffé-

5. Manning reconnaît cependant que les données qui intéressent les syntacticiens demandent de recourir à de très grands corpus, nécessairement annotés, et requièrent encore à l'heure actuelle de la part des syntacticiens des « compétences informatiques considérables » (2003 : 294-295). De même, les données qui intéressent les sociolinguistes sont tout aussi difficiles à construire, sauf à rassembler des données confondant temps, espace, classe sociale, méthode de recueil, etc.

rence des données de la linguistique de corpus « qui s'intéresse aux textes, aux textes réels » (Williams 2005 : 13).

Si on peut considérer que ce premier ensemble d'arguments définit par la négative l'intérêt des approches sur corpus, cet ensemble la définit aussi de manière positive dès lors qu'on en formule la version opposée :

– À la critique portant sur le petit nombre, les approches sur corpus opposent la masse (considérable) de données[6], sur laquelle on peut travailler.

– À la critique portant sur leur caractère inventé et sur les simplifications opérées, les approches sur corpus revendiquent leur caractère réel, leurs propriétés d'objectivité, de représentativité, de vérifiabilité.

Il s'en déduit que travailler à l'aide de corpus permettrait de trouver de nouveaux faits (Kennedy 1998, Manning 2003) et, par là même, de récuser des connaissances faussement établies (Manning 2003).

### 1.3. Remise en cause des concepts fondateurs de la linguistique générative

Pour nombre d'auteurs, Chomsky est la cible principale. Comme le fait remarquer Léon (2005), il occupe la place d'un « antiprécurseur », contre lequel s'unifie tout un courant. Sont visés plus particulièrement les travaux de Chomsky de 1957, 1965 et 1986, et notamment le concept de grammaticalité dans son articulation avec l'acceptabilité, la compétence et la performance. Nous passons en revue successivement les critiques :

***1) La grammaticalité et l'acceptabilité.*** Plus précisément ce sont les jugements de grammaticalité et d'acceptabilité qui sont critiqués. Et ce sous trois angles : leur caractère binaire ; leur caractère subjectif (compris comme variable) ; l'absence de précision quant à leur objet.

a) Les jugements ne doivent plus être considérés comme des jugements binaires. On est en présence d'un continuum. Ce que notent Bod *et al*. (2003 : 1) : « […] it has become increasingly clear that alternations and judgments display **properties of continua** and show markedly **gradient behavior**. » Ou encore Abney (1996 : 10), qui oppose le caractère graduel du jugement (d'acceptabilité) et le caractère catégorique de la grammaticalité : « Judgments are rather **smoothly graded** but goodness according to the grammar is all or nothing ». Dans la même ligne, Habert *et al*. (1997 : 9) notent que « les linguistiques de corpus se trouvent confrontées à un éventail de réalisations langagières qui remet en cause les distinctions tranchées entre acceptable et non-acceptable ».

Les linguistes (et particulièrement ceux qui s'inscrivent dans un cadre génératif), en s'intéressant essentiellement aux données jugées selon l'une ou l'autre des extrémités de l'échelle (acceptable/grammatical ou bien inacceptable/

6. Du moins depuis le développement de la puissance des ordinateurs, qui a permis de construire de (très) grands corpus.

agrammatical), ont laissé de côté bon nombre de phénomènes ou n'ont pas su fournir les explications adéquates (Bod *et al*. 2003 : 1).

Par ailleurs, il importe de noter que les jugements (de grammaticalité) sont corrélables à la fréquence des formes jugées. C'est ce qu'indique par exemple Manning (2003 : 298-302), à propos de cadres de sous-catégorisation proposés par Pollard et Sag (1994)[7] : « First, the line of grammaticality has apparently been drawn at a fairly arbitrary of low frequency. Roughly things that occur at least 1 time in 100 are regarded as grammatical, while things that occur less commonly are regarded as ungrammatical ». Comme, par ailleurs, ce jugement est prédictible à partir d'un calcul faisant usage de probabilités : « […] speakers'well-formedness judgments of words and sentences are extremely well-predicted by the combined probabilities of their subparts » (Bod *et al*. 2003 : 1), il s'en déduit que le recours à des approches statistiques est justifié, et ce contrairement à ce qu'affirmait Chomsky (1957). Celles-ci permettent de rendre compte de l'existence de degrés dans le jugement d'acceptabilité, degrés dont on ne peut rendre compte quand on ne dispose que de grammaires algébriques : « Gradations of acceptability are not accommodated in algebraic grammars : a structure is grammatical or not » Abney (1996 : 14)[8].

b) Le jugement repose sur l'introspection des linguistes. Or l'introspection n'est pas un moyen fiable : différentes expériences ont montré qu'il existait des variations inter et même intra individuelles. Sont ainsi rappelés par Manning (2003) les travaux de Labov (1975 [2001][9]). Abney (1996, note 5) note aussi que les désaccords ne sont pas exceptionnels, que le jugement peut varier en fonction des hypothèses théoriques avancées, que le jugement peut varier dans le temps.

c) Enfin, on ne sait (jamais) précisément ce sur quoi porte le jugement, étant donné qu'interviennent et des facteurs contextuels et notre capacité à imaginer des contextes favorables. Manning (2003 : 310) en conclut : « I tend to feel that there are good reasons for getting at synchronic human grammars as gradient phenomena from large data sets of actual language use rather than from human judgments. In part this is due to the unclarity of what syntactic acceptability judgments are actually measuring […] ».

---

7. Plus précisément, il s'agit de la sous-catégorisation de verbes tels que *consider*, *regard*, *turn out*, *end up*, pour laquelle Manning a systématiquement trouvé, dans le *New-York Times*, des contre-exemples.

8. Toutefois, on reconnaît que les grammaires algébriques ont parfois admis que certaines structures agrammaticales puissent être pires que d'autres. Par exemple « ECP violations versus subjacency violations. […] such degrees of grammaticality as have been considered have not been accorded a formal place in the theory » (Abney 1996 : 14).

9. Jacques (2005 : 22) rapporte ainsi les conclusions de Labov (1975 [2001]) (en omettant toutefois de faire précéder sa citation de « À l'heure actuelle ») : « aucun résultat ne permet d'entretenir l'espoir que les jugements introspectifs des linguistes soient fiables, reproductibles ou généralisables dans leur application au langage de la communauté ».

***2) Les concepts de compétence et de performance.*** Les critiques portent sur leur délimitation et par suite sur l'objet même de la linguistique, qui ne peut plus être identifié à l'étude de la seule compétence.

a) Est d'abord souligné le fait que les délimitations des concepts de compétence et de performance n'ont cessé d'évoluer. Ainsi Manning (2003 : 307), à la suite de Abney (1996), souligne que des faits, initialement considérés par la linguistique générative comme relevant de la performance, ont été considérés comme des faits de compétence dès qu'on a su les traiter (le cas échéant dans le cadre d'une approche probabiliste) : « In recent decades, the scope of grammar has been expanded in various ways : people now routinely put into grammar semantic and discourse facts that would have been excluded in earlier decades ».

b) Ensuite, ce sont la distinction et l'articulation de ces deux concepts qui sont visées : « Dividing the human language capacity into grammar and processor is only a manner of speaking, a way of dividing things up for theoretical convenience […] it is not a distinction with any empirical content » (Abney 1996 : 10). Les arguments invoqués sont fondés sur le caractère graduel du jugement (cf. ci-dessus) ; sur le fait que la séparation entre grammaire (compétence) et processing (performance) ne correspond à aucune « meaningful physiological division » ; enfin, sur la nature des jugements, qui sont aussi des jugements perceptifs, et ce quelle que soit la nature de ce que l'on juge (une phrase occurrence ou une structure). Comme les données qui intéressent le linguiste sont « the result of the interaction of grammatical constraints with computational constraints », la distinction entre compétence et performance ne tient plus.

Cette critique de Abney rejoint des critiques plus anciennes qui visaient aussi bien l'opposition compétence/performance que l'opposition saussurienne langue/parole.

– Leech (1992 : 108) indique que : i) les faits de performance sont observables, alors que l'étude de la compétence repose sur des considérations abstraites et spéculatives ; ii) l'étude de la performance est plus « utile » que celle de la compétence (dans des domaines tels que l'éducation, la traduction, le TAL) ; iii) mais surtout, l'opposition compétence/performance n'est pas aussi nette qu'on a bien voulu le dire : « the putative gulf between competence and performance has been overemphasised, and the affinity between (say) the grammar of a language as a mental construct and the grammar of a language as manifested in performance by the native speaker must be close, since the latter is product of a former […] I-language and E-language are likely to be closely related : in a sense, like different sides of the same coin ».

– Halliday (1992 : 66) indique ainsi : « We are so accustomed to thinking about language and text in terms of dichotomies such as the Saussurean langue and parole, or Hjelmslev's system and process, that we tend to objectify the distinction : there is language as a system, an abstract potential, and there are spoken and written texts, which are instances of language in use. But the "system" and the "instance" are not two distinct phenomena. There is only one phenomenon here, the phenomenon of language : what we have are

> two different observers, looking at this phenomenon from different depths in time. If I may use once again the analogy drawn from the weather : the instance-observer is the weatherman, whose texts are day-to-day weather patterns displaying variations in temperature, humidity, air pressure, wind direction and so on, all of which can be observed, recorded and measured. The system-observer is the climatologist, who models the total potential of a given climatic zone in terms of overall probabilities. »[10].

c) La compétence chomskyenne repose sur la notion de « locuteur idéal appartenant à une communauté homogène » (Chomsky 1957). Si Abney (1996), par exemple, reconnaît que, pour Chomsky lui-même, ce type d'idéalisation était un expédient, un moyen commode de circonscrire ce qu'on ignorait, il n'empêche que la linguistique « has lost sight of its original goal, and turned Chomsky's expedient into an end in itself ».

d) C'est alors l'objet même de la linguistique qui est redéfini. Il englobe et la compétence et la performance (tout au moins une large part) chez Abney : il s'agit de décrire la langue/le langage « in the sense of that which is produced in language production, comprehended in language comprehension, acquired in language acquisition, and, in aggregate, that which varies in language variation and changes in language change » (1996 : 12). Dans ce cadre, « the grammar should predict which structures humans actually perceive, and what the relative weights are in access where humans are uncertain about which structures the speaker intended » (Abney 1996 : 13). Pour Rastier (2005 : 35), « les règles de la langue sont sans doute des normes invétérées et les performances de la parole ne restent évidemment pas exemptes de normativité : elles instancient et manifestent les règles de la langue et diverses normes sociolectales. Bref, la linguistique peut prendre *de droit* pour objet de description l'espace des normes : au lieu de les édicter, comme elle le faisait naguère en frappant d'inacceptabilité des énoncés alors même qu'ils sont attestés, elle doit les décrire et pour cela exploiter des corpus ». De manière générale, tout ce qui a trait à l'usage des formes (identifié par leur fréquence, leur distribution, etc.), soit des données considérées comme des données de performance, doit être considéré comme central (Leech 1992, Kennedy 1998).

### 1.4. Nouveaux champs, thèmes mieux pris en compte

On a vu précédemment que la possibilité de « découvrir » de nouveaux faits était l'un des points avancés, de manière positive, en faveur des approches sur corpus. Plus souvent, les arguments avancés, toujours de manière positive, portent sur la possibilité de prendre en compte des champs considérés comme ayant été ignorés par les courants dominants en linguistique, principalement

10. On remarquera que Rastier (2005 : 33-34) s'arrête, de même, sur l'opposition langue/parole de Saussure : « […] aucune théorie linguistique n'est parvenue à résoudre le problème de leur articulation […] Or, seule la linguistique de corpus peut offrir les moyens théorique et technique d'étudier l'espace des normes et de transformer en dualité l'antinomie entre compétence et performance ».

l'acquisition, la variation, le changement (Leech 1992, Abney 1996, Kennedy 1998, Manning 2003 ; voir aussi Jacques 2005). Sont aussi notés les domaines pour lesquels il est difficile, voire impossible d'émettre un jugement d'acceptabilité : « […] les corpus peuvent rassembler des énoncés sur lesquels l'analyste n'est pas forcément à même de porter des jugements d'acceptabilité. C'est le cas par exemple pour des corpus de langues mortes […], pour des corpus de langues de spécialité […] » (Habert *et al*. 1997 : 9). Ont aussi émergé de nombreux travaux portant sur la notion de genre, étudiée à la fois en tant que telle mais aussi comme dimension nécessaire à prendre en compte quand on veut construire des corpus (représentatifs) d'une langue donnée (Biber 1988, 1993 ; article Corpus dans Asher & Simpson 1994 ; Rastier 2005). Ce que résument Habert *et al*. (1997 : 9) de la manière suivante : « Il est [alors] possible d'étudier dans le détail, en dépassant les caractérisations trop globales, et donc caricaturales, l'opposition entre oral et écrit, l'organisation globale des textes, mais aussi les contrastes socio-linguistiques. »

Par ailleurs, les tenants des approches sur corpus soulignent que ce type d'études a fait progresser de manière notable les travaux et les applications en TAL (Svartik 1992, Kennedy 1998). Les méthodes sont des méthodes probabilistes fondées sur des calculs statistiques effectués à partir de corpus[11]. Mais, au-delà du TAL, les modèles probabilistes ont pour fonction de modéliser les langues. « Certaines constructions […] sont extrêmement fréquentes, d'autres rares ou exceptionnelles. […] On peut aussi chercher à articuler les règles et le poids comparé des différentes régularités observées. Dans cette conception, les règles ne sont pas toutes sur le même plan : certaines sont centrales, d'autres périphériques. Les règles changent alors de statut. C'est une vision probabiliste de la grammaire. » (Habert *et al.* 1997 : 9). Ces modèles peuvent par ailleurs être appliqués à divers champs : « I think the most compelling, though least well-developed, arguments for statistical methods in linguistics come from the areas of language acquisition, language variation, and language change » (Abney 1996 : 2). Enfin, ils permettent de renouveler les descriptions phonologiques, morphologiques, syntaxiques, etc. (pour un aperçu, voir l'introduction de Bod *et al*. 2003).

## 1.5. Bilan

Les arguments avancés par les tenants des approches sur corpus se laissent séparer en deux ensembles : des arguments construits de manière négative et des arguments construits de manière positive. Les premiers cherchent à démonter le programme de recherche issu des travaux de Chomsky (caractéristiques des données manipulées et concepts fondateurs). Les seconds cherchent d'abord à promouvoir des domaines tels que l'acquisition, la variation, le changement, ou bien des approches probabilistes. Les deux types d'arguments se rejoignent dans l'ambition, parfois implicite, de définir un nouveau programme de recherche et par suite l'objet même de la linguistique.

11. Voir l'article de Marcel Cori dans ce numéro.

## 2. DISCUSSION DES POSITIONS

La discussion des arguments avancés par les « linguistes de corpus » va s'opérer dans deux directions. D'une part, afin de répondre aux critiques concernant les données, nous revenons sur la notion de *fait* en linguistique. Nous insistons notamment sur le caractère de science expérimentale qui peut être accordé à la linguistique, quand elle effectue des manipulations sur les *exemples* qu'elle construit (et qui s'opposent d'une certaine manière aux données brutes extraites des corpus). D'autre part, afin de répondre aux arguments positifs en faveur de l'usage des corpus, nous montrons que ceux-ci, quelle que soit l'aide qu'ils peuvent apporter au travail des linguistes, ne changent pas les difficultés rencontrées dans l'élaboration scientifique.

### 2.1. Impossible et possible de langue

Comme toutes les sciences, la linguistique ne considère pas qu'un fait est une évidence brute, issue directement de l'observation. Un fait linguistique est nécessairement construit par le linguiste qui, devant produire des hypothèses, a à décider de ce qui est significatif dans ce qu'il observe. On ne peut ainsi s'en tenir au seul terme de « donnée », entendu jusqu'à maintenant comme « matériau linguistique brut », qui ne présente pas le caractère construit du fait.

Milner a, d'une certaine façon, répondu par avance et de manière très détaillée aux approches basées sur les corpus : « On peut émettre un jugement différentiel concernant les données de langue. Le principe de ce jugement différentiel est que tout ne peut pas se dire. Autrement dit, le jugement concerne ce qui en matière de langue est possible ou impossible ; il suppose donc qu'il y a un impossible de langue. Toutefois, cet impossible de langue n'est pas un impossible matériel. Autrement dit, une donnée de langue peut être possible matériellement, c'est-à-dire attestée, et impossible en langue, ou inversement » (Milner 1989 : 55)[12]. Ce qu'il résume un peu plus loin de la manière suivante : « le possible de langue et le possible matériel sont disjoints » (Milner 1989 : 83). Les deux points importants qui se dégagent de cette position sont d'une part la disjonction entre le possible de langue et l'attesté, et d'autre part l'émission d'un jugement différentiel, comme éléments fondateurs de l'activité grammaticale.

1) Le premier point est sans doute le lieu où la confrontation avec les approches sur corpus est la plus aiguë. L'insistance à caractériser les données linguistiques comme devant être des données « réelles » laisse à penser que les approches sur corpus refusent la disjonction entre le possible et l'attesté. Ce refus pouvant se décliner selon deux versions distinctes :

(i) Soit, comme le dit Milner (1989 : 109 et suiv.) : « on considère qu'on ne peut savoir si une donnée de langue est possible que si elle est attestée ». Mais cela

12. Le terme « impossible » n'est peut-être pas le meilleur, étant susceptible de prêter à confusion : comment peut-on voir ou entendre ce qui est impossible ?

revient à promouvoir un projet « résolument anti-grammatical »[13]. De plus, c'est bien parce qu'il y a du possible non attesté qu'il y a nécessité de se doter de grammaires qui permettent de rendre compte de cet état de fait, de prédire, non pas comment la langue évoluera, mais quelles expressions seront considérées comme faisant partie de la langue par les locuteurs. Et c'est un des grands apports de Chomsky d'avoir défini, en se fondant sur la logique mathématique, des grammaires formelles finies qui permettent d'engendrer des langages (potentiellement) infinis. Comme le dit alors Milner (1989 : 120-121), si l'on reconnaît comme propriété fondamentale du langage celle de la créativité, alors « il est paradoxal que la science linguistique se donne comme principe de constitution l'hypothèse exactement contraire ». Si, en outre, on admet que la tâche du linguiste est de construire des grammaires qui décrivent une langue, on ne peut lui interdire d'écarter des données « réelles » qui n'entrent pas dans son système, tout au moins dans le cadre d'une démarche expérimentale (cf. ci-dessous). Sa grammaire est alors une hypothèse sur le monde, dont on pourra se demander si elle est féconde.

(ii) Soit on considère que toute donnée attestée est possible (quelquefois en assortissant cette affirmation d'un seuil en terme numérique, ou en terme de variété des sources). Ce qui entre en contradiction avec l'idée selon laquelle il existe de l'impossible attesté. La question qui se pose alors est : peut-on déduire de la simple figuration d'un mot ou d'une suite de mots dans un corpus son appartenance à une langue ?

2) Le deuxième point soulève une réelle difficulté de la recherche en linguistique. Le jugement différentiel[14] vise avant tout une différence, le fait que les données puissent être jugées de manière inégalitaire. Ses caractéristiques sont les suivantes (Milner 1989 : 56) :

– c'est un jugement bivalent, même si l'on reconnaît qu'il puisse y avoir de la gradation dans les jugements[15] ;

– c'est un jugement qui a la forme logique de la norme (car ce qui est « incorrect » peut aussi être attesté).

La question de la gradation est une question très sensible, au cœur des critiques dirigées contre le paradigme génératif[16]. Elle relève en fait de plusieurs aspects qui, sans être indépendants, sont à distinguer :

---

13. Milner poursuit ainsi : « […] il est possible que la science du langage doive renoncer à ce qui fondait l'autonomie de la grammaire. On conçoit pourtant qu'on puisse hésiter et qu'il faudrait de grandes raisons pour accepter une décision aussi drastique ».

14. Le terme de « jugement différentiel » présente l'avantage de ne pas distinguer entre jugement de grammaticalité (défini comme se rapportant à la compétence) et jugement d'acceptabilité (défini comme se rapportant à la performance), ni même entre jugement de grammaticalité ou jugement d'appropriation (Marandin 2004).

15. Le paradigme génératif a admis depuis longtemps qu'il puisse y avoir de la gradation dans les jugements : dès 1955, les notions d'acceptabilité et de grammaticalité sont, pour Chomsky, une question de degré. Ce qui est très souvent oublié par les adversaires des générativistes.

16. Une question qui est sans doute pour beaucoup dans le succès relatif des approches sur corpus.

(i) Les productions linguistiques seraient plus ou moins acceptables, et non pas soit acceptables, soit non acceptables.

(ii) Il y a de la variation dans la langue, que celle-ci se manifeste entre des communautés linguistiques différentes ou bien à l'intérieur d'une même communauté. De même, on observe de la variation chez un même locuteur[17]. Et il n'est pas toujours simple de démêler ce qui ressortit à chacun des niveaux.

Les adeptes des corpus traitent ces deux premiers points d'un seul mouvement ou presque, à l'aide de statistiques sur les données attestées, qui permettraient de donner une réalité numérique à cette gradation.

(iii) Le jugement émis par les linguistes peut varier, sous l'influence des facteurs précédents mais pour d'autres raisons également. Ainsi, quand le linguiste est lui-même locuteur de la langue qu'il étudie, il pourrait être tenté de se passer d'autres locuteurs et de se fier uniquement à son jugement. Le linguiste serait à ce moment-là observateur et objet de l'observation – ce qui est précisément critiqué sous le terme d'*introspection*[18].

Le paradigme génératif ne nie pas qu'il puisse y avoir de la variation dans la langue, et par suite dans les jugements. À l'intérieur d'une même communauté, même socialement homogène, on peut avoir des systèmes de répartition différents sur la grammaticalité/agrammaticalité. La solution consiste alors à ne traiter qu'un système de répartition à la fois, ou bien à ne traiter que des systèmes de répartition compatibles (Milner 1989 : 124)[19].

Il est clair que les réponses qui ont été apportées ne résolvent pas définitivement le problème[20]. L'utilisation systématique de procédures explicites, moins discutables[21], reste cependant une voie intéressante. Elles auraient au moins le mérite de circonscrire les éventuelles zones d'achoppement. Mais force est aussi de constater que le problème ne relève pas uniquement de cet ordre :

---

17. Les raisons en sont diverses : effet de l'exposition, compétences différentes, « accès à plusieurs univers de langue » (Milner 1989 : 124), etc.

18. La notion de norme intervient aussi crucialement dans le débat, et la crainte est évidemment de confondre une norme académique et une norme effective, autrement dit de confondre grammaire prescriptive et grammaire descriptive. Le problème est général en sciences humaines : les êtres humains modifient leur comportement selon la description qui en est faite. La pratique langagière est sans doute la plus touchée, si on pense au rôle que jouent l'école ou d'autres institutions.

19. Pour une illustration de la prise en compte, dans l'analyse linguistique, de la variation dans les jugements, voir Godard & Jayez 1996, qui montrent les différences et les corrélations entre des jugements sur des données concernant le typage des unités lexicales et la reprise pronominale

20. Voir par exemple, les longs avertissements méthodologiques de Chomsky (1965).

21. Comme a été inventé le test « en double aveugle » pour juger de l'efficacité d'une action thérapeutique (les statistiques pourraient intervenir, mais pas des statistiques sur l'existence de certaines constructions dans des corpus, des statistiques portant sur des jugements émis par différents locuteurs, linguistes ou pas). On s'était rendu compte que la subjectivité des personnes chargées de conduire les tests était susceptible d'en fausser les résultats, ce qui n'est pas loin des difficultés rencontrées par les linguistes. Mais nul n'a songé à rejeter l'hypothèse selon laquelle il y avait des substances qui pouvaient améliorer le sort de patients atteints de certaines maladies.

parce qu'à différents moments de la recherche, on doit faire comme si le problème était résolu alors qu'il peut être en jeu dans l'analyse ; parce que, comme le dit Milner (1989 : 124), « [le] calcul de compatibilité ne ressortit pas à l'observation brute, mais à un raisonnement […] linguistique ».

Il est plus difficile de se mettre d'accord sur l'acceptabilité de telle « donnée » langagière que de mesurer la vitesse d'un solide sur un plan incliné. Toutefois, de la difficulté à émettre **dans certains cas** un jugement différentiel qui ferait consensus (car il n'y a bien évidemment pas uniquement des cas compliqués), à conclure à l'invalidité d'un objet ou d'un programme de recherche, il y a là un saut logique et épistémologique qu'on ne saurait faire.

## 2.2. Données *vs* exemples

Dans le cadre esquissé ci-dessus, ce n'est pas une donnée linguistique « brute » qui constitue la base observable, mais une donnée linguistique pourvue d'un jugement différentiel, que l'on appelle « exemple » (Milner 1989 : 51). Le fait linguistique se définit alors de la manière suivante : « un fait est une différence de grammaticalité ou de signification entre deux expressions […]. La différence est « mise en scène » sous forme d'un contraste entre deux (ou plusieurs) exemples » (Marandin 2001 : 1)[22]. Il résulte de la mise en relation d'exemples « possibles » et « impossibles » une caractéristique semble-t-il inédite parmi les pratiques scientifiques : « la présence d'exemples dont on assure qu'ils sont incorrects. C'est quelque chose qui n'a pas de sens dans le domaine des sciences de la nature : un non-corps, un non-phénomène, etc., sont tout simplement des néants d'existence » (Auroux 1998 : 188-189). Auroux parle en ce cas d'« anti-exemples »[23].

Rappelons quelques-unes des propriétés de l'exemple[24] :

– L'exemple est *construit* : (i) « en tant qu'il est *analysé* » ; (ii) en tant qu'on sélectionne parmi les dimensions mises en jeu (thème de la « concrétion » des données de langue) celle(s) qui importe (nt) dans l'analyse (Milner 1989 : 113-114).

– L'exemple est *abstrait*, parce qu'on ne s'intéresse pas à ses « conditions singulières d'énonciation » (Milner 1989 : 109), excepté bien sûr quand cela est requis dans certains domaines d'analyse. De manière générale, ce n'est pas « un document » (Milner 1989 : 118). Il est aussi abstrait « parce qu'il est

22. Voir également l'introduction de Boons *et al.* (1976 : 39-40), qui proposent, dans leur étude des constructions infinitives, qu'un exemple soit constitué d'au moins de 4 séquences.

23. On peut relever toutefois l'ambiguïté de l'astérisque. En effet, comme le remarque Milner (1989 : 51, note 30), ce que note l'astérisque est ambigu entre « cet énoncé ne se dit pas en fait » et « cet énoncé ne devrait pas se dire en droit ». Desclés et Guentcheva (1991 : 23-24) soulignent eux aussi que « l'emploi de l'astérisque n'est pas assez théorisé. Il recouvre en pratique les situations suivantes : faible plausibilité […], mauvaise formation par rapport à une norme grammaticale ou non-grammaticalité […], inénonçabilité […], énonçabilité souvent présentée à contexte vide ».

24. On remarquera que Milner et Auroux, bien qu'ayant des positions épistémologiques très éloignées, convergent sur un certain nombre de points qui ont trait aux données.

choisi pour illustrer au mieux un phénomène, et, du même coup, une règle. Par conséquent, il doit être simple et, si possible, ne doit illustrer qu'un phénomène à la fois. Dans ces conditions, il y a interchangeabilité entre l'exemple et la règle. » (Auroux 1998 : 191)

– L'exemple est « une occurrence qui représente son type », comme le dit Auroux (1998 : 185) : « À dire vrai, l'exemple n'est pas dans une grammaire comme une plante dans un champ, mais comme une plante dans un herbier. Il n'est pas un *usage* de la langue, mais une mention *sui-référentielle* d'un de ses éléments : c'est une occurrence qui représente son type. De la même façon, la plante scotchée sur la page de l'herbier ne présente pas à celui qui la regarde simplement sa propre forme, mais le prototype d'une classe de plantes. À la différence de la plante, l'occurrence est aussi le nom de son type. » Cette propriété est aussi mise en avant chez Milner (1989 : 52) : l'exemple « n'importe pas par sa particularité mais par sa généralité ». Il n'est pas simplement une expression analysée pourvue d'un jugement, mais une classe d'expressions « semblables »[25], pourvue d'un jugement.

– L'exemple présente certaines des propriétés d'une expérimentation. C'est une « configuration de données propre à tester une proposition de la théorie […] une instance minimale de réfutation » (Milner 1989 : 115)[26]. Auroux rappelle que l'expérimentation, en la matière, n'est pas chose nouvelle, puisque des protocoles tels que « couper des phrases, permuter ses éléments, etc. » sont connus depuis longtemps. Ce sont des manipulations sans instrument : « L'une des particularités des sciences du langage, en effet, c'est que le langage est sans médiation à disposition du locuteur : je puis produire, à volonté, des phrases, les tronquer, y introduire tel élément que je choisis, etc. Il se pourrait que ce soit le seul exemple d'une manipulation sans instrument, du moins le seul qui se soit maintenu dans un état développé d'une discipline scientifique. » (Auroux 1998 : 170)[27].

Les exemples peuvent être construits en partant de données attestées. Si celles-ci sont ajustées, voire simplifiées, c'est d'abord et avant tout qu'elles sont trop complexes au regard du problème traité (cf. le thème de la concrétion, Milner 1989 : 113). « Toute donnée de langue […] combine toujours plusieurs dimensions ». Par exemple, l'unité lexicale combine toujours une forme phonologique, une catégorie, un sens ; la constitution d'une phrase met toujours en jeu plusieurs « règles ». En constituant les données en *exemple*, on les ajuste pour

25. Il est clair que la construction du caractère « semblable » ne présente aucune évidence. La généralisation empirique qu'on effectue met en jeu des fragments de grammaire, dont certains doivent parfois être admis (cf. le thème de l'absence d'observatoire) ; elle isole des dimensions (cf. thème de la concrétion) alors même que l'identification de ces dimensions constitue un des résultats.

26. Toutefois, l'expérimentation linguistique n'a pas toutes les propriétés de celles qui sont mises en œuvre dans les sciences expérimentales, notamment : l'exemple n'a pas la structure de l'événement ; on ne dispose d'aucun outillage ; on ne dispose d'aucun observatoire (Milner, 1989 : 126 et suiv.).

27. Auroux (1998 : 171) cite notamment « les permutations des phonologues et leur test des paires minimales [qui] sont incontestablement des manipulations, quoique sans instrument. »

n'exhiber que les éléments ou les dimensions essentiels, ou ce que l'on croit être les éléments ou les dimensions essentiels (en cela, la démarche linguistique est analogue à l'expérimentation mise en place dans les sciences dites dures). Et il est vrai que ces opérations d'ajustement présentent des risques. Il faut que l'exemple reste proférable (s'il est grammatical), il ne faudrait pas avoir supprimé les éléments cruciaux sur lesquels reposerait une analyse valide du problème traité, etc. Rien ne garantit à coup sûr la validité de ces ajustements.

Il en est de même pour les données qui peuvent être qualifiées d'« inventées ». Encore qu'il n'y ait pas de frontière nette entre l'ajustement et l'invention. Allons cependant plus loin. L'invention est une procédure légitime, scientifiquement féconde quand l'objectif du linguiste est d'écrire une grammaire. Il est ainsi possible et nécessaire, afin de tester une grammaire, d'examiner quels types d'expressions elle est susceptible d'engendrer, ce qui d'ailleurs peut s'effectuer de manière automatique. Il reste alors à porter un jugement d'acceptabilité sur les expressions engendrées, ce qui renvoie aux difficultés énoncées ci-dessus.

La construction des exemples n'est cependant pas exempte de difficultés. Comme le souligne Auroux (1998 : 192) : « On court […] le risque d'utiliser des exemples qui ne sont attestés que dans les grammaires, tout comme il y a des formes qui ne sont attestées que dans les dictionnaires. La constitution d'un corps d'exemples est, comme le montre l'histoire de la grammaire, un processus lent et compliqué. Les exemples sont repris d'une grammaire à l'autre ; parfois simplement traduits lorsque l'on rédige la grammaire d'une autre langue. » Il est donc nécessaire de se donner des garde-fous méthodologiques, notamment en rendant explicites les démarches de recueil, de manipulation et d'analyse. Ce que propose Auroux (1998 : 193) : « On ne fera donc pas l'économie de protocoles dans la construction des exemples (ce qu'on peut appeler une théorie des observables) et d'une déontologie dans le choix de l'univers de la langue. » Quoi qu'il en soit, aucune pratique scientifique ne garantit par elle-même qu'elle ne produira pas des erreurs.

## 2.3. Le corpus n'efface pas les difficultés

Les auteurs qui considèrent que la linguistique fondée sur les corpus est beaucoup plus objective que la linguistique qu'ils qualifient d'introspective oublient qu'il y a de l'arbitraire lors de la constitution des corpus. Tout ce qui est attesté ne peut entrer dans un corpus, si grand soit-il. Dans certains cas, on fait entrer dans le corpus les données qui sont disponibles, parce qu'elles ont été stockées sur des supports électroniques, le plus souvent pour de tout autres raisons que celles qui motivent la recherche linguistique. Ainsi on trouvera beaucoup plus d'articles de journaux que de correspondances entre individus par exemple. En ce sens, on peut dire que les résultats de la recherche dépendent des conditions matérielles de la constitution des corpus[28].

28. Quand on utilise des corpus annotés, on est dépendant également d'autres conditions matérielles, de la façon dont les corpus ont été annotés (choix des étiquettes, erreurs dans l'annotation, etc.).

Par ailleurs, le linguiste qui constitue un corpus n'y fait pas entrer toutes les données attestées qui sont à sa disposition. Il doit faire des choix, en se demandant notamment si ses données sont représentatives, ou si tout bonnement elles se rattachent ou non à la langue dont le corpus est censé donner une image. Parmi les questions qui se posent : faut-il écarter des productions orales ou des textes écrits venant de locuteurs non natifs ? Que faire des productions des enfants ? Quelle variété de langue faut-il prendre en compte ? Etc. Il en résulte que le linguiste qui constitue un corpus doit lui aussi émettre un jugement ultimement bivalent, qui, dans les faits, repose en partie sur son intuition. En cela, la situation n'est pas si éloignée de celle à laquelle sont confrontés les linguistes qui émettent un jugement différentiel.

Il importe ensuite de noter que les données « réelles », « objectives », si l'on reprend les termes des tenants des approches sur corpus, ne sont jamais les données sur lesquelles travaille un linguiste utilisant des corpus. On le sait bien : les corpus doivent être préparés. On corrige, on sélectionne, on élimine, en fonction des traitements ultérieurs (voir pour des exemples les articles de Fradin *et al*., Delais, Geyken, ce numéro). Cette première phase peut même s'avérer complexe, car on peut avoir à résoudre des problèmes, pour lesquels on n'a pas de réponse assurée, pour lesquels le travail sur corpus doit justement apporter une réponse, etc. Parmi les précautions méthodologiques indispensables, un inventaire soigneux de toutes les décisions prises en amont doit être tenu.

À partir du moment où on se livre à des manipulations, même minimales, sur les données, comment pourrait-on prétendre à une objectivité que l'on dénie à d'autres ? On comprend qu'il faille éliminer des corpus les fautes de frappe (encore que tous les cas ne présentent pas le même degré de difficulté). Mais comment préciser la différence entre l'« erreur de langage » ou l'émission d'un énoncé qui ne ferait pas partie du « possible de langue », sauf à avoir résolu le problème par avance ? Comme le remarque justement Auroux (1998 : 241), l'existence d'erreurs de langage disqualifie toute approche qui se contente de généraliser ce que l'on trouve dans les corpus : « L'un des principaux obstacles à la construction d'une théorie linguistique à partir de l'*usage* (c'est-à-dire d'un *corpus*) a été mis en lumière par l'analyse des *erreurs de langage*. » En effet, « la reconnaissance de l'erreur de langage comme celle de la faute de langue suppose au préalable la connaissance de la règle et de la correction. » (1998 : 241-242). Auroux évoque notamment le lapsus et observe qu'un locuteur admettra volontiers lui-même que ce qu'il a pu dire – ou écrire – ne correspond pas à son intention. Or qui peut décider dans une approche où c'est le corpus qui dirige d'éliminer certaines données (attestées) ? Ce n'est en tout cas pas la régularité qui peut en décider, car il existe des régularités y compris dans les fautes : « l'erreur de langage est tout aussi « régulière » que le parler correct » (Auroux 1998 : 242). L'erreur est donc répétable, dans certains cas répétée… et exploitée. Pour ne donner qu'un exemple, c'est bien sur cette observation que se fondent les correcteurs d'orthographe qui utilisent des statistiques sur les fautes les plus couramment commises. Auparavant, en amont

des développements statistiques, quelqu'un aura aussi émis un jugement sur le caractère hors norme de certaines productions (attestées).

Il s'ensuit que la présence, même répétée, d'une forme dans un corpus ne prouve pas que l'on doive considérer qu'il s'agit bien d'un possible de langue[29].

Inversement, le corpus ne permet pas de réfuter une hypothèse du type « telle construction est possible ». Car la non-présence d'une forme dans un corpus n'est en rien une preuve d'impossibilité. Les « anti-exemples » ne peuvent être extraits des corpus. Comme le dit Fillmore « we still have to face the reality that they are no corpora of starred examples : a corpus cannot tell us what is not possible » (Fillmore 1992 : 58). À moins d'admettre un équivalent de ce qu'on appelle « l'hypothèse du monde clos » en logique, qui revient à dire que si une proposition n'est pas énoncée, sa négation est vraie.

## CONCLUSION

Il n'est évidemment pas question de nier l'aide que peuvent apporter les corpus à la recherche en linguistique. Selon les domaines ou les questions considérées, le recours à des corpus est plus ou moins aisé, plus ou moins crucial. Mais ce qu'il importe de noter c'est que les corpus, si grands et si bien faits soient-ils, ne changent pas fondamentalement la position du chercheur en linguistique. *Grosso modo*, le linguiste se comporte – ou doit se comporter – de la même manière face à un corpus ou face à des données en quantité beaucoup plus restreinte qu'il aurait recueillies en écoutant des conversations, en lisant un journal ou des copies d'élèves, ou même de données qu'il aurait inventées : choisir les données à considérer, les mettre en forme, les ramener à la dimension qui doit être étudiée, émettre un jugement.

Néanmoins, certaines critiques des approches sur corpus pointent effectivement de réelles difficultés de la recherche en linguistique (rassemblement des données et jugements différentiels notamment). Elles ont de ce fait le mérite d'inciter les linguistes à rappeler ce qui fonde leur travail, à se donner plus de rigueur dans leurs méthodologies.

---

29. Le lien entre (a)grammaticalité et fréquence n'est pas simple. À titre d'illustration, on peut citer l'article de Manning (2003), qui fait notamment état de la découverte d'un nouveau « fait » en anglais (p. 294), soit l'emploi de « as least as » plutôt que de « at least as ». Mais il fournit de manière vague le nombre d'occurrences brutes, et surtout, on ne connaît pas (et personne ne pourrait fournir une telle indication) la taille de l'ensemble examiné, parce qu'il s'est notamment appuyé sur le Web. Dans ce cadre, même plusieurs centaines d'occurrences (p. 293) ne permettent pas d'affirmer à coup sûr qu'il ne s'agit pas d'une erreur répétée. Il est évident, néanmoins, qu'une présence massive dans un corpus doit amener le linguiste à s'interroger sur la donnée, sans pour autant laisser le corpus « décider ».
Ainsi, lors d'une recherche que nous avons effectuée sur la Toile à l'aide du moteur de recherche Google le 25 mai 2007, nous avons trouvé 14 400 occurrences de « vous disez », contre 1 320 000 occurrences de « vous dites » et 2 230 occurrences de « vous ditent ». Quelle conclusion faut-il en tirer ? La réponse ne peut venir des seules données brutes.

## Références


ABNEY S. 1996. « Statistical Methods and Linguistics ». Klavans J., Resnik Ph. (éds), *The Balancing Act*. Cambridge : The MIT Press, 1-26.

ASHER R.E., SIMPSON J.M.Y. (éds). 1994. *The Encyclopaedia of Language and Linguistics*. Oxford : Pergamon Press.

ATKINS B.T.S., CLEAR J., OSTLER L. 1992. « Corpus Design Criteria ». *Literary and Linguistic Computing. Journal of the Association for Literary and Linguistic Computing*, 7 (1), 1-16.

AUROUX S. 1998. *La raison, le langage et les normes*. Paris : PUF.

BIBER D. 1988. *Variation across Speech and Writing*. Cambridge : CUP.

BIBER D. 1993. « Representativeness in Corpus Design ». *Literary and Linguistic Computing. Journal of the Association for Literary and Linguistic Computing*, 8 (4), 243-257.

BILGER M. (éd.). 2000. *Corpus. Méthodologie et applications linguistiques, Actes du colloque « Questions de méthodes dans la linguistique sur corpus ». 1998*. Paris : Honoré Champion et Presses Universitaires de Perpignan.

BLANCHE-BENVENISTE Cl. 2000. « Types de corpus ». Bilger M. (éd.), *Corpus. Méthodologie et applications linguistiques, Actes du colloque « Questions de méthodes dans la linguistique sur corpus ». 1998*. Paris : Honoré Champion et Presses Universitaires de Perpignan, 11-15.

BOD R., HAY J., JANNEDY S. (éds). 2003. *Probabilistic Linguistics*. Cambridge-Londres : The MIT Press, 1-10.

BOONS J.-P., GUILLET A., LECLÈRE Ch. 1976. *La Structure des phrases simples en français. Constructions intransitives*. Genève : Droz.

CHOMSKY N. 1955. *The Logical Structure of Linguistic Theory*. Bibliothèque de MIT, miméo.

CHOMSKY N. 1957. *Structures syntaxiques*. Paris : Le Seuil. [trad. 1969].

CHOMSKY N. 1965. *Aspects de la théorie syntaxique*. Paris : Le Seuil. [trad. 1971].

CHOMSKY N. 1986. *Barriers*. Cambridge : The MIT Press.

DESCLÉS J.-P., GUENTCHEVA Z. 1991. « Test et acceptabilité ». *Histoire, Épistémologie, Langage*, 13 (2), 9-25.

FILLMORE Ch. 1992. « 'Corpus Linguistics'or 'Computer-aided Armchair Linguistics' » In Svartvik J. (éd.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium, 4-8 août 1991*. Berlin-New York : Mouton de Gruyter, 35-60.

GODARD D., JAYEZ J. 1996. « Types nominaux et anaphores : le cas des objets et des événements ». *Chronos*, 1, 41-58.

HABERT B. 2000. « Des corpus représentatifs : de quoi, pour quoi, comment ? ». Bilger M. (éd.), *Linguistique sur corpus. Études et réflexions*. Perpignan : Presses Universitaires de Perpignan, 11-58.

HABERT B., NAZARENKO A., SALEM A. 1997. *Les linguistiques de corpus*. Paris : Armand Colin et Masson.

HALLIDAY M.A.K. 1992. « Language as System and Language as Instance : The Corpus as a Theoretical Construct ». Svartvik J. (éd.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium, 4-8 août 1991*. Berlin-New York : Mouton de Gruyter, 31-77.

JACQUES M.-P. 2005. « Pourquoi une linguistique de corpus ». Williams G. (éd.), *La linguistique de corpus*. Rennes : Presses Universitaires de Rennes, 21-29.

KENNEDY G. 1998. *An Introduction to Corpus Linguistics*. Londres-New-York : Longman.

LABOV W. 1975 [2001]. « Qu'est-ce qu'un fait linguistique ? ». *Marges linguistiques*, 1.

LEECH G. 1992. « Corpora and Theory of Linguistic Performance ». Svartvik J. (éd.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium, 4-8 août 1991*. Berlin-New York : Mouton de Gruyter, 105-125.

LÉON, J. 2005. « Claimed and Unclaimed Sources of Corpus Linguistics ». *Henry Sweet Society Bulleti*n, 44, 36-50.

MANNING Ch. 2003. « Probabilistic Syntax ». Bod R., Hay J., Jannedy S. (éds), *Probabilistic Linguistics*. Cambridge-Londres : The MIT Press, 289-341.

MARANDIN J.-M. (éd.). 2001. *Cahier Jean-Claude Milner*. Paris : Verdier.

MARANDIN J.-M. 2004. « Formatage de l'information : focus et contexte ». Corblin F., Gardent Cl. (éds), *Interpréter en contexte*. Paris : Hermès, 31-80.
MILNER J.-Cl. 1989. *Introduction à une science du langage*. Paris : Le Seuil.
POLLARD C., SAG I. 1994. *Head-Driven Phrase Structure Grammar*. Chicago : University of Chicago Press.
RASTIER F. 2005. « Enjeux épistémologiques de la linguistique de corpus ». Williams G. (éd.), *La linguistique de corpus*. Rennes : Presses Universitaires de Rennes, 31-45.
*Revue française de linguistique appliquée : Grands corpus : diversité des objectifs, variété des approches*. 1999. 1.
SINCLAIR J. 1996. http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.
SVARTVIK J. (éd.). 1992. *Directions in Corpus Linguistics. Proceedings of Nobel Symposium, 4-8 août 1991*. Berlin-New York : Mouton de Gruyter.
TOGNINI-BONELLI E. 2001. *Corpus Linguistics at Work*. Amsterdam : John Benjamins.
WILLIAMS G. (éd.). 2005. *La linguistique de corpus*. Rennes : Presses Universitaires de Rennes.