

Le rôle des corpus dans la recherche linguistique

Si les corpus linguistiques tels qu'ils sont aujourd'hui constitués et mis à la disposition des chercheurs grâce aux nouvelles technologies, en ont complètement renouvelé la problématique, le travail de description d'une langue à partir de données attestées et systématiquement recensées n'a rien de vraiment nouveau. Pour s'en tenir au français et aux seules productions modernes du xx^e siècle, nul ne peut ignorer l'apport en ce domaine de chercheurs tels que le danois Sandfeld, les français Damourette et Pichon ou le belge Grevisse, et dont la tradition se maintient fermement de nos jours dans les départements d'études romanes de nombreuses universités scandinaves. Et ces auteurs avaient bien conscience de l'intérêt de cette démarche de collecte d'exemples vrais, tel Sandfeld qui écrit dans l'avant-propos de son ouvrage sur les pronoms en 1928 :

« Le but que je me suis proposé en écrivant ce livre, a été de donner du français de nos jours une peinture aussi détaillée et aussi vivante que possible. A quelques exceptions près [...], tous les exemples ont été puisés dans des écrivains contemporains ou dont l'activité principale tombe après 1870. »

Et, se référant aux travaux d'un de ses collègues disparus, il ajoute :

« Comme lui, je me suis efforcé, dans la mesure de mes moyens, de peindre “la vie pétillante de la langue vivante avec ses lois et ses libertés, ses différentes couches et sa diversité bariolée”, et j'ai cherché comme lui à faire parler surtout les exemples. »

F. Desonay redit de même en 1955, dans la préface qu'il rédige pour la 6^e édition du *Bon Usage* de Grevisse :

« Le principal mérite de M. Grevisse est de se tenir, greffier vigilant et diligemment informé, aux écoutes des meilleurs écrivains contemporains [...]. A cet égard, les listes d'exemples que multiplie M. Grevisse après l'exposé toujours clair et souvent “élégant” de chaque “vérité” grammaticale sont constamment mises à jour, tant se révèle exigeant le scrupule d'épouser la ligne fluctuante de [la] langue écrite [...]. »

Enfin Damourette et Pichon ouvrent leur monumental *Essai* en affirmant dès le premier paragraphe :

« Il ne s'agit pas ici d'une synthèse hâtivement faite d'après les travaux d'autrui. Bien que nous n'ayons pas négligé de nous entourer, à l'occasion, des lumières des nombreux et savants linguistes qui ont étudié la langue française, nous pouvons néanmoins dire que ce n'est pas leurs travaux qui

constituent le fond de notre ouvrage. Attelés depuis l'année 1911 à la confection du présent Essai, nous nous sommes attachés à rassembler une quantité importante de matériaux : on rencontrera, dans le livre premier et dans le livre II, un certain nombre de théories très générales, mais toutes sont étayées sur des faits concrets, car elles ne sont que la synthèse d'études de détail qui figureront, dans la suite de l'ouvrage, avec un grand nombre d'exemples justificatifs. »

Si leur *Essai* se distingue en effet par l'abondance des exemples — plus de 34 000 au total, et dont le nombre précis est toujours soigneusement indiqué sur la couverture de chaque tome — il se distingue aussi des autres travaux cités en ce qu'ils ne sont pas uniquement écrits, mais également oraux, et que ces exemples oraux sont présentés dans une transcription soigneusement pensée et décrite par les auteurs, et scrupuleusement respectée par eux tout au long de leurs sept tomes.

Ceci étant, et aussi respectables qu'ils soient, ces travaux ne peuvent être considérés comme des *corpus*, au sens que l'on donne aujourd'hui à ce terme, et sur lequel on reviendra plus bas, mais apparaissent plutôt comme des recueils de données, dont les limites sont liées à la façon dont ils ont été ou sont d'abord constitués, puis exploités :

- la collecte d'exemples reste forcément artisanale, au fil des lectures effectuées par les auteurs dans des ouvrages, presque toujours écrits, et retenus de façon assez arbitraire ;
- la tâche de classement des exemples ainsi collectés demeure prépondérante, avec le souci de ne laisser aucun exemple de côté, et souvent au détriment de la description linguistique, qui reste assez sommaire, sinon même traditionnelle.

Le cas le plus frappant à cet égard reste sans doute celui des ouvrages de Sandfeld, dans lesquels on ne trouve vraiment pas grand'chose ni sur le fonctionnement des pronoms par exemple, ni sur celui de l'infinitif — avec quelques classements contradictoires en prime, un même exemple pouvant être analysé d'une façon puis d'une autre dans des chapitres différents. Damourette et Pichon, qui ont eu sans aucun doute de plus grandes ambitions théoriques, n'échappent pas à ce travers qui, dans un chapitre sur les auxiliaires, proposent dans un seul paragraphe (§1612) 130 exemples, qui occupent treize pages, à l'appui d'une présentation des emplois du verbe *être* comme auxiliaire d'antériorité qui tient exactement en trente lignes. Par ailleurs, relevés pour la plupart dans des textes écrits, sans doute plus diversifiés aujourd'hui qu'au début du siècle, ces exemples ne peuvent donner une image que de l'attesté, et plus

précisément de ce qui apparaît comme la norme du français écrit reconnu comme correct. Mais surtout, et en dépit de leur nombre, les exemples collectés le sont de façon aléatoire, et sans valeur statistique. Si bien que rien n'autorise à penser que leur rassemblement couvre effectivement l'ensemble du système linguistique. En tout cas les descriptions s'en ressentent manifestement, qui proposent ainsi une image du français correct, sans en avoir forcément perçu clairement les véritables principes de fonctionnement.

J'illustrerai ces limitations par deux exemples, qui me semblent assez caractéristiques.

Premier exemple : Dans le chapitre qu'il consacre au relatif *dont*, et en s'appuyant seulement sur les exemples recueillis, non seulement Grevisse ne rend nulle part correctement compte des répartitions d'emploi,¹ pourtant fortement contraintes, entre les relatifs apparemment concurrents *dont* / *duquel* / *de qui*, tous trois liés à la même reprise d'un GN précédé de la préposition *de*, mais il ne dit nulle part de façon explicite que la possibilité d'apparition de *dont* dépend fondamentalement du statut non prépositionnel du Nom tête dont dépend, directement ou non, le groupe de GN relativisé.

- 1 Un vieil homme était assis [sur le seuil [de la porte]]
 ⇒ ... *la porte dont le vieil homme était assis [sur le seuil [-]]
 ⇒ ... la porte [[sur le seuil de laquelle] le vieil homme était assis [-]]
- 2 Il a été élu [au conseil d'administration [du Crédit Lyonnais]]
 ⇒ ... ?le Crédit Lyonnais [dont il a été élu [au conseil d'administration [-]]]
 ⇒ ... le Crédit Lyonnais [[au conseil d'administration duquel] il a été élu [-]]
- 3a J'entends [les cris [des oiseaux du parc]]
 ⇒ ... les oiseaux du parc [dont j'entends les cris [-]]
- 3b J'entends [les cris [des oiseaux [du parc]]]
 ⇒ ... *le parc [dont j'entends les cris des oiseaux [-]]
 ⇒ ... *le parc [les cris des oiseaux duquel] j'entends [-]]

Second exemple : D'après les grammaires traditionnelles, le verbe de certaines complétives peut être soit à l'indicatif soit au subjonctif, mais il ne peut être au subjonctif qu'à la condition que la phrase supérieure (contenant le verbe introducteur) soit négative ou interrogative. Mais cette dernière formule est beaucoup trop vague, car, parmi les constructions dites interrogatives, c'est la seule inversion du pronom sujet qui autorise le subjonctif dans la complétive (et

¹ Sur ce point, on peut trouver néanmoins une description plus fine, même si elles reste incomplète, dans Godard, 1988, 74 et sqq.

sans doute avec des restrictions supplémentaires sur la personne du pronom sujet, et le temps du verbe introducteur).

4 a Crois-tu que Jean est un bon candidat ?

b Crois-tu que Jean soit un bon candidat ?

5 a Est-ce que tu crois que Jean est un bon candidat ?

b *Est-ce que tu crois que Jean soit un bon candidat ?

Or, en travaillant sur cette question (Huot 1986), je suis tombée sur un article norvégien (Börjeson 1966) consacré à « la fréquence du subjonctif dans les subordonnées introduites par *que* étudiée dans des textes français contemporains » et dans lequel les nombreux exemples rassemblés contenaient donc uniquement des tours interrogatifs avec inversion du pronom sujet. Néanmoins l'auteur n'avait manifestement pas vu ce lien entre le subjonctif de la complétive et l'inversion du pronom sujet du verbe principal, ni *a fortiori* l'importance dans le fonctionnement du français moderne de cette inversion du clitique sujet, dont l'interprétation essentielle n'est sans doute pas strictement interrogative. Ces deux anecdotes signifient sans doute aussi que le relevé d'exemples ne peut dispenser d'hypothèses sur le fonctionnement de la langue, et que celles-ci ne peuvent sortir des seules données, mais sont étroitement dépendantes des cadres théoriques disponibles. Ces réserves pourraient valoir également pour les corpus dont on parle aujourd'hui, mais si l'objection n'est pas complètement infondée, elle a cependant une portée limitée, en rapport avec la nature même des corpus actuels. Car ce qui distingue fondamentalement ces corpus de ce que j'ai appelé des recueils de données, c'est leur taille qui garantit une certaine exhaustivité; et surtout leur informatisation qui en permet une exploitation systématique (cf. à ce sujet l'excellent livre de Habert *et al.* qui vient de sortir). A titre d'exemples, et pour s'en tenir aux corpus portant sur l'anglais (écrit et oral), qui sont aujourd'hui les plus développés,

- le corpus connu sous le nom de *Bank of English*, d'où a été tiré le Collins *COBUILD English language Dictionary*, comporte aujourd'hui près de 320 millions de mots, et s'accroît au rythme de 50 millions de mots par an ;
- le British National Corpus contient aujourd'hui 4000 textes, avec un total d'environ 100 millions de mots.

Les corpus portant sur le français sont aujourd'hui peut-être moins développés, mais certains, commencés très tôt pour la fabrication du *Trésor de la Langue française*, sont immenses, même si la partie aujourd'hui consultable sur ordinateur sous le titre de *Frantext*, ne contient que des textes du XIX^e et du XX^e siècle. En ce qui concerne le français parlé, l'équipe d'Aix-en-Provence a constitué depuis une

vingtaine d'années un corpus qui atteint aujourd'hui près de deux millions de mots, et apparaît de fait comme le corpus le plus important en la matière. Mais il existe bien d'autres corpus sur le français, en cours de constitution et d'informatisation, et dont nous avons essayé de donner un aperçu dans le n° 2 de la *Revue Française de Linguistique Appliquée*, consacré justement aux corpus. La taille même de ces corpus, et surtout leur indexation systématique, qui en permet une exploitation, y compris statistique, à la fois rapide et plus fiable, ont complètement renouvelé notre approche des faits linguistiques, dans les domaines les plus divers : syntaxe, lexique, construction de dictionnaires, traduction, acquisition de la langue et des langues. C'est sans doute dans le domaine du lexique et du fonctionnement syntaxique des unités lexicales que les choses ont été le plus spectaculaires, ouvrant d'ailleurs de nouvelles perspectives didactiques qui restent pour une bonne part encore à explorer. L'exploitation de grands corpus permet en effet de décrire beaucoup plus finement les caractéristiques syntaxiques et sémantiques des unités lexicales, mais aussi, de façon plus originale, leur combinatoire, au travers de ce que l'on appelle désormais leurs collocations, c'est-à-dire leurs possibilités d'association avec d'autres éléments du lexique. Et dans ce domaine, ce sont sans doute les travaux liés au *Dictionnaire Explicatif et Combinatoire* développé par I. Mel'chuk qui sont à la fois les plus avancés et les plus prometteurs,² ouvrant de nouvelles perspectives sur le fonctionnement des unités lexicales dont certaines par exemple (Tutin 1997) peuvent être couramment accompagnées de collocatifs à valeur d'évaluation ou d'intensité, mais qui sont très spécifiés :

envie, n *folle* : avoir une envie folle de...
pluie, n *torrentielle* : une pluie torrentielle
conseil, n *précieux* : de précieux conseils
peur, n *bleue*, *grosse* : une peur bleue, une grosse peur
mal, n *fou*, *de chien* : un mal fou, un mal de chien

De grands corpus seuls permettent également de confirmer ces limitations d'emploi que respectent intuitivement des locuteurs natifs, mais qui étaient jusqu'à maintenant assez mal décrites et cause de pas mal d'erreurs de la part des locuteurs non natifs. J'ai découvert moi-même avec un certain étonnement, en

² Mais commencent également à être disponibles dans le commerce d'autres travaux novateurs, tels ceux d'un jeune collègue belge T. Fontenelle sur l'utilisation et l'exploitation d'un dictionnaire bilingue, en l'occurrence le Collins-Robert *English-French Dictionary*.

explorant systématiquement *Frantext* à propos des adjectifs en *-aire* et *-ier / -ière*, que l'on appelle couramment relationnels, plusieurs points peu signalés :

- que certains de ces adjectifs, peu courants, n'apparaissent qu'en relation à des N très spécifiques,
 - mobilier* { objets, valeurs, capitaux, biens, revenus }
 - centenaire* { arbre(s) }
 - censitaire* { suffrage, régime }
 - concordataire* { régime }
 - disciplinaire* { matière, questions, régime, mesures, quartier... }
 - forfaitaire* { indemnité, somme, subvention, prix }
 - grégaire* { instinct }
 - hebdomadaire* { journée, horaire, durée, service }
 - indiciaire* { classement, échelonnement }
 - paritaire* { commission, accords, négociations, réunions }
 - réglementaire* { textes, mesures }
 - sédimentaire* { dépôt(s), roches, terrain(s), couche(s) }
- que leur emploi attributif n'est pas possible seulement avec des N sujets [+ Humain] comme il est couramment dit, mais qu'il est attesté aussi avec quelques N sujets [- Humain] très limités.
 - Le problème est financier.
 - Cette assimilation nous paraît arbitraire.
 - Les objets sont généralement plus documentaires qu'esthétiques...
 - La subvention de l'état est désormais forfaitaire.
 - Le circuit de distribution commerciale est lacunaire.
 - ... des disques qui allaient devenir légendaires.
 - La gestion est paritaire.
 - L'attribution de ressources devient ainsi prioritaire.
 - La répartition individuelle est rarement égalitaire.
 - Elle [cette population] est stationnaire lorsque ...
 - Le "bibliobus" est révolutionnaire.

Mais il apparaît que, dès lors qu'on peut étudier la distribution d'un verbe, d'une préposition, d'un nom ou d'un adjectif en consultant de tels corpus, c'est la technique même de l'analyse distributionnelle qui est changée. Comme chaque mot semble avoir sa propre grammaire, il n'est plus possible de faire une grammaire pour des classes complètement homogènes correspondant aux parties du discours, des noms, des adjectifs ou des verbes. Si l'idée d'étudier le lexique de la langue selon les environnements que révèlent les concordanciers est aujourd'hui assez largement acceptée, il n'en est pas encore de même pour la grammaire. Et pourtant les corpus font apparaître de la même manière des phénomènes encore peu connus. Dans un article récent (1996), C. Blanche-Benveniste cite le cas du passif dont l'emploi semble lié à certaines classes de verbes: certains sont fréquemment utilisés au passif, d'autres rarement et d'autres

jamais. Il ressort de l'exploration des corpus de langue parlée que les verbes à effet statif se rencontrent facilement au présent inaccompli

le bar est fréquenté par les Gitans
la porte est entourée de guirlandes

alors que les verbes non-statifs sont rares dans ces emplois. Néanmoins ces mêmes verbes se manifestent au passif avec des compléments d'agent, mais accompagnés d'un auxiliaire qui leur donne un aspect accompli

la route a été ouverte il y a longtemps par un groupe de militaires

On peut déduire de ces observations, présentées ici trop rapidement, que le passif est un phénomène lié aux qualités lexicales des verbes, limité dans ses réalisations aspectuelles et fortement dépendant des contextes dans lesquels il est utilisé. Le recours aux concordances établies d'après un corpus permet aussi de vérifier des hypothèses, telle celle, avancée justement par Blanche- Benveniste, sur la valeur du pronom *on*, qui serait dépendante des pronoms avec lesquels il entre en concurrence dans un syntagme: suivi de *me* ou de *nous*, il a automatiquement le sens d'un complexe de personnes dont le MOI est exclu, sinon il est interprété comme une 3^e personne indéterminée *ils* et n'a jamais le sens de *nous*. Mais s'il n'est pas suivi de *me* ni de *nous*, il peut avoir la valeur d'un complexe de personnes incluant le MOI. D'où les valeurs différentes, et non ambiguës des deux *on* dans un énoncé comme:

quand on était petits on nous grondait souvent.

Enfin les grands corpus permettent d'affiner certains aspects de l'évolution linguistique. Pour reprendre l'exemple du relatif *dont* signalé plus haut, et sous condition que les corpus écrits de français disponibles proposent des types d'écrits plus diversifiés que FRANTEXT actuellement, il devient possible de vérifier si la généralisation de *dont* au détriment des tours prépositionnels en lequel sont vraiment entrés dans la langue écrite courante ou non. Il est clair en tout cas que la vitalité de *dont* reste plus grande à l'oral qu'on ne croit, même s'il apparaît lié à quelques verbes privilégiés, comme l'ont montré les comptages effectués par l'équipe d'Aix-en-Provence.

90 % des emplois relevés se font en effet avec les verbes suivants :

parler 48,4%

<i>avoir besoin</i>	12,9%
<i>faire partie</i>	8,1%
<i>prendre conscience</i>	4,8%
<i>sortir</i>	3,2%
<i>dépendre</i>	3,2%
<i>être question</i>	3,2%
<i>être convaincu</i>	3,2%

tandis que 45% des emplois de *dont* dans le journal *Le Monde* (dont les collections sont désormais consultables par CD-ROM) concernent les verbes suivants :

<i>parler</i>	8,9%
<i>faire preuve de</i>	8,9%
<i>avoir besoin</i>	6,3%
<i>dire</i>	5,1%
<i>disposer</i>	5,1%
<i>rêver</i>	3,8%
<i>souffrir</i>	3,8%
<i>sortir</i>	2,5%
<i>dépendre</i>	1,3%

L'existence de ces grands corpus et les nouvelles possibilités d'exploration et de décompte qu'ils offrent conduisent à revoir certains outils pédagogiques couramment utilisés en matière d'apprentissage des langues (première ou seconde), telles ces listes de mots, souvent justifiées par la notion de fréquence. Le meilleur exemple en reste sans doute le fameux *Français fondamental* de G. Gougenheim, qui fit certainement œuvre de pionnier en travaillant à partir d'un corpus. Mais on voit mieux aujourd'hui que c'est l'étroitesse de ce corpus (en rapport avec les moyens de l'époque) qui explique et éclaire les insuffisances de ces listes, qui n'ont jamais été très opératoires. Les corpus, et notamment les corpus bilingues qui se multiplient, enrichissent aussi incontestablement le travail de comparaison entre langues. Dans un travail récent, effectué à partir d'un corpus relevé manuellement sans doute, et qui demanderait donc à être confirmé par une exploration plus systématique dans des corpus plus larges, une spécialiste des questions de traduction entre le français et l'anglais, J. Guillemin-Flescher, a réussi à faire apparaître des spécificités fines propres à chacune des langues. Ainsi, lorsqu'il s'agit de représenter une activité, une relation à deux arguments ou un état, on observe que l'anglais a tendance à privilégier l'objet de la relation, et la quantification là où le français privilégie plutôt le sujet de la relation et la qualification :

1. There was a rustle behind them, proceeding from a hedge
Ils entendirent derrière eux un bruissement dans une haie

2. There was an indistinguishable murmur of male voices
On entendait un murmure confus de voix d'hommes
3. There is sand on his lip
Il a du sable sur la lèvre
4. She was a tall girl, ungainly...
Elle était grande, et un peu gauche...
5. He was a superstitious man
Le vieil homme était superstitieux.

Et les corpus oraux qui se constituent actuellement au niveau européen et international, devraient permettre de la même façon de mieux connaître la façon dont se fait l'acquisition du langage et des langues secondes, dans des situations variées d'apprentissage. S'il est clair cependant que les corpus, et en particulier les grands corpus informatisés, ne suppriment pas forcément le recours à l'intuition du locuteur natif, ni l'importance de l'argumentation linguistique, et qu'ils ne peuvent servir de substitut aux théories linguistiques pour la construction d'hypothèses sur le fonctionnement de la langue, il n'en reste pas moins que la description linguistique fondée sur ces corpus renouvelle fondamentalement non seulement les données elles-mêmes, plus complètes et variées, mais également l'outillage de l'analyse et la conception de la grammaire. Ces très grands corpus, qui représentent d'immenses inventaires toujours en expansion, rendent aujourd'hui un peu vaines les discussions anciennes sur la finitude et la non-finitude des données. Les changements d'échelle intervenus ces dernières années grâce aux progrès technologiques ont entraîné des changements dans la nature même de l'analyse. La possibilité de travailler à la fois sur des sous-parties de ces inventaires et sur des grandes masses, et de comparer plus systématiquement des données d'origine diverse, permet d'apprécier autrement le poids et l'importance des phénomènes grammaticaux. Le lexique est ouvert, mais la grammaire aussi. Avec ces nouvelles masses de données disponibles, il n'est plus possible de présenter toute la grammaire comme un ensemble de règles ou d'explications à valeur très générale, car les règles grammaticales apparaissent de plus en plus limitées par les disponibilités lexicales et les genres textuels dans lesquels elles sont utilisées. Et les phénomènes grammaticaux apparaissent ainsi comme bien plus hétérogènes qu'on ne l'avait cru jusqu'ici. Là où l'opposition était forte il y a encore quelques années, les corpus s'imposent aujourd'hui à tous les linguistes car, à peine constitués, ils ont déjà montré quels extraordinaires outils

d'investigation ils étaient, et combien ils contribuaient à renouveler et surtout enrichir les descriptions linguistiques.

Références

- Blanche-Benveniste, C. (1996) « De l'utilité du corpus linguistique ». *Revue Française de Linguistique Appliquée* 1-2: 25-42
- Börjeson, L. (1966) « La fréquence du subjonctif dans les subordonnées introduites par *que* étudiée dans des textes français contemporains » *Studia Neophilologica* 38: 3-64
- Damourette, J. & Pichon, E. (1933-1950) *Des Mots à la pensée. Essai de grammaire de la langue française*. Paris, Ed. d'Artray
- Fontenelle, T. (1997) *Turning a Bilingual Dictionary into a Lexical-Semantic Database* Tübingen Niemeyer, 331 p
- Godard, D. (1988) *La syntaxe des relatives en français* Paris, Editions du CNRS
- Grevisse, M. (1936) *Le Bon Usage* 11^e édition, 1980, Gembloux, Duculot / Paris, Hatier. 12^e édition refondue par A. Goosse, Gembloux, Duculot
- Guillemin-Flescher, J. (1996) « La traduction humaine : contraintes et corpus » *Revue Française de Linguistique Appliquée*, I-2: 43-56.
- Habert B., Nazarenko, A. & Salem A. (1997) *Les linguistiques de corpus* Paris, Armand Colin.
- Huot, H. (1986) « Le subjonctif dans les complétives: subjectivité et modalisation » In Ronat M. & Couquaux D. (eds), *La Grammaire modulaire* Paris, Ed. de Minuit: 81-111
- Sandfeld, K. (1928) *Syntaxe du français contemporain. Les Pronoms* Paris, Champion, 1970
- Sandfeld, K. (1935) *Syntaxe du français contemporain. Les Propositions subordonnées* Genève, Droz, 1965
- Sandfeld, K. (1943) *Syntaxe du français contemporain. L'Infinitif* Genève, Droz, 1965
- Tutin, A. (1997) « Coder les collocations dans un lexique formel pour le TALN ». *Revue Française de Linguistique Appliquée* II-1: 43-57.

Hélène Huot
université Paris 7-Denis Diderot