Corpus et recherches linguistiques

Introduction

Sylvie Mellet

Édition électronique

URL : <https://journals.openedition.org/corpus/7>

DOI : 10.4000/corpus.7

ISSN : 1765-3126

Éditeur

Bases ; corpus et langage - UMR 6039

Édition imprimée

Date de publication : 15 novembre 2002

ISSN : 1638-9808

Référence électronique

Sylvie Mellet, « Corpus et recherches linguistiques », *Corpus* [En ligne], 1 | 2002, mis en ligne le 15 décembre 2003, consulté le 15 avril 2022. URL : <http://journals.openedition.org/corpus/7> ; DOI : <https://doi.org/10.4000/corpus.7>

Ce document a été généré automatiquement le 15 avril 2022.

© Tous droits réservés

Corpus et recherches linguistiques

Introduction

Sylvie Mellet

- 1 Créer une nouvelle revue est une aventure, stimulante certes, mais risquée aussi, et particulièrement gourmande en temps et en énergie. Il y faut donc de bonnes raisons. Au nombre de celles qui nous ont poussés à créer se trouve le constat que la linguistique de corpus, actuellement très en vogue, ne disposait d'aucun périodique qui lui fût spécifiquement consacré. Les ouvrages collectifs, les manuels de haut niveau, les numéros spéciaux fleurissent, mais de revue régulière, point. L'UMR « Bases, Corpus et Langage », équipe membre de l'Institut de Linguistique française (dont les programmes fédératifs sont pour l'essentiel orientés vers la constitution, la diffusion et l'exploitation de corpus écrits et oraux pour l'étude du français, de ses variantes et de ses dialectes) s'est donc fixé comme objectif de fonder et animer une revue de qualité consacrée à la linguistique de corpus ; celle-ci sera envisagée sous tous ses aspects : théoriques, épistémologiques, méthodologiques ; la revue sera ouverte également à la présentation des diverses applications dans les champs disciplinaires et sur les domaines géolinguistiques les plus variés. Le comité de rédaction veillera cependant à ce que chaque numéro, quelle que soit son orientation spécifique, apporte sa contribution à une meilleure appréhension du rôle des corpus dans les pratiques linguistiques contemporaines et à une analyse réflexive mettant en question le mode de constitution et/ou d'exploitation du corpus dans les études présentées. L'entreprise se veut ambitieuse : nous remercions tous ceux qui ont bien voulu s'y associer en participant à l'un ou à l'autre des comités qui en garantiront la qualité scientifique.
- 2 La notion de corpus paraît, de prime abord, assez simple et bien ancrée dans certaines traditions des sciences humaines et sociales, philologique ou juridique par exemple. Il s'agit d'un recueil, formé d'un ensemble de données sélectionnées et rassemblées pour intéresser une même discipline. Néanmoins, dans le champ linguistique, la notion s'est complexifiée au cours des dernières décennies en fonction de la diversité des pratiques et des objectifs assignés à la constitution et à l'exploitation des corpus. Or le moment semble venu d'explicitier ces pratiques, de les questionner et de tenter de mettre au jour leur impact épistémologique sur l'évolution de la discipline, tant il est vrai –

comme l'écrit Damon Mayaffre – que la qualité première d'une démarche scientifique est d'être une démarche qui s'explique.

- 3 Quelques questions récurrentes sont déjà posées dans la littérature consacrée au sujet : cette revue permettra aux chercheurs d'horizons divers d'y apporter leurs réponses personnelles. La première de ces questions – chronologiquement au moins, et peut-être hiérarchiquement – porte sur la définition du corpus et sa clôture. Un corpus ne peut être **clos et exhaustif** que dans le cadre d'une monographie, auquel cas il sera étudié en tant que tel, sans prétendre à être représentatif d'autre chose que de lui-même ni à ouvrir sur aucune forme de généralisation ou modélisation. Un tel corpus est aussi, généralement, très homogène. On le rencontre notamment dans les études stylistiques ou en analyse du discours. Néanmoins, le respect intégral des contraintes de clôture et d'exhaustivité semble difficile à tenir ; si l'on prend l'exemple d'un corpus textuel littéraire, le développement de la recherche elle-même, ses nécessités internes, poussent souvent à intégrer petit à petit au corpus initial des gloses de l'auteur étudié, des commentaires critiques, voire d'autres œuvres littéraires à des fins de comparaison. Le corpus initial subit donc une extension à un corpus connexe ou contextuel dont l'ampleur varie avec l'empan du regard. Comme le souligne F. Rastier, « la pensée du contexte est une pensée des points de vue » et un texte, loin d'être une « monade », n'est qu'une « globalité transitoire » au sein d'un contexte¹. Plusieurs articles de ce recueil reviendront sur ce rôle du point de vue qui oriente la sélection du corpus et sur le problème de l'intégration du contexte – ou, plus largement, de l'intertexte – au corpus de travail.
- 4 A l'opposé des corpus homogènes et exhaustifs se trouvent les **corpus échantillonnés** ; là, le problème se déplace : l'enjeu n'est plus celui de l'exhaustivité, mais celui de la **représentativité**. Il s'agit alors de constituer des échantillons représentatifs d'une réalité plus large – en statistique on dirait : d'une population. Les objectifs d'une telle entreprise peuvent être assez variés : appréhender et donner à voir cette réalité trop vaste pour être embrassée dans sa totalité (par exemple, décrire le français oral du XX^{ème} siècle, comparer l'anglais britannique et l'anglais américain, etc.) ; se donner les bases empiriques nécessaires pour répondre à un questionnement théorique ou étayer une hypothèse structurale (par exemple décrire, comprendre et unifier les emplois du conditionnel dans le système verbal français contemporain) ; constituer enfin les bases de connaissances indispensables au développement des nouveaux outils réclamés par l'essor de l'ingénierie linguistique. Dans tous les cas, il s'agit de constituer ce que S. Aurox appelle un « observatoire »². Plusieurs méthodes s'offrent aux équipes de chercheurs pour constituer cet objet intermédiaire entre les faits empiriques et le modèle théorique : ils peuvent faire confiance aux grands nombres selon l'adage bien connu « *more data, better data* » ; les développements récents de l'informatique ont facilité ce choix méthodologique en permettant la création et le stockage de « très grands corpus » ; mais aux débuts du renouveau de la linguistique de corpus, la puissance informatique était moindre et les pionniers ont souvent adopté une méthode inspirée plutôt de la pratique des quotas, familière aux instituts de sondage, mêlant dans des proportions jugées représentatives des usages des extraits textuels empruntés aux différents genres et sous-genres des discours répertoriés (oral familier, oral académique, presse, roman, théâtre, poésie, discours politique, ouvrages scientifiques, etc.). Ces corpus se veulent généralement **des corpus de référence**, exploitables pour des recherches variées par plusieurs générations de linguistes. Enfin, dans le cas d'un

corpus échantillonné au contraire aux fins d'une recherche spécifique, la sélection des données pourra être délibérément orientée en fonction de cet objectif particulier : pour reprendre l'exemple d'une étude sur le conditionnel français, on conçoit aisément que le chercheur engagé dans cette étude surreprésente dans son corpus les extraits de presse (du *Monde* notamment), les romans d'Echenoz et certains dialogues oraux. Un biais est introduit, voulu et en principe contrôlé par une hypothèse de travail. L'« observatoire » se transforme alors en lieu d'« expérimentation » ; celle-ci « consiste en ce que le sujet connaissant construit un montage qui lui permet : i) de déterminer à volonté la production du phénomène ; ii) d'isoler les effets des différents paramètres constituant le phénomène. A l'aide de ce montage, on doit pouvoir corroborer ou infirmer une hypothèse précise »³. L'article de J.-Ph. Dalbera donne ici un exemple concret des enjeux d'une telle démarche et des difficultés auxquelles elle se heurte inévitablement.

- 5 Comme on le voit, nous sommes déjà loin de la conception naïve qui prévalait encore il y a une dizaine d'années, selon laquelle la constitution d'un corpus de données attestées devait permettre d'éviter toute manipulation artificielle de la réalité : le travail sur corpus faisait partie des *behavioral, natural methods*⁴. Et encore : *The data of a corpus, more thoroughly than we have grown to expect in linguistics, are independant of the tenets of the theory they are required to test*⁵. Bien au contraire, le constat s'est partout imposé que le corpus est un objet construit, que ce soit à travers l'effacement symbolique de tout ce qu'il ne contient pas – ce que R. Nicolai a appelé le « paradoxe de l'archéologue »⁶ –, ou que ce soit à travers la structuration, l'organisation des données retenues, voire leur enrichissement au moyen des procédures du balisage et de l'étiquetage. Le fait n'est pas nouveau : toute classification, tout tableau synthétique récapitulatif de la philologie traditionnelle offraient aussi cette part d'interprétation implicite des données. Le développement des bases de données informatisées, le formalisme obligé de leur structure, le questionnement sur l'échangeabilité des données, sur la standardisation des étiquettes et des formats associés ont sans doute ouvert la voie à une prise de conscience plus aigüe du phénomène. Le corpus est devenu une médiation consciente entre le chercheur et le fait linguistique ; sa constitution prend donc appui sur des procédures explicitées qui font partie intégrante de l'herméneutique mise en œuvre.
- 6 Tels sont donc quelques-uns des problèmes bien connus auxquels toute linguistique de corpus se trouve confrontée. Notre ambition est d'ouvrir un lieu de réflexion et de discussion où ils puissent être repris de manière plus concrète à travers l'exposé critique de diverses pratiques de recherche linguistique et la confrontation d'expériences qui n'auront pas toutes le même rapport aux corpus. Pour les disciplines qui font leurs premiers pas en linguistique de corpus, il s'agira donc plutôt de déterminer quels sont les types de questionnements pour lesquels le recours aux corpus semble utile et d'en fixer les règles ; pour les autres, il s'agira de mettre à distance des pratiques déjà anciennes, d'en évaluer la pertinence, d'en mesurer les acquis et les limites.
- 7 Face à ces objectifs, il était normal de consacrer le premier numéro de la revue *CORPUS* à quelques mises au point générales sur la notion même de corpus et sur son utilisation dans diverses sous-disciplines de la linguistique. D'autres numéros suivront, qui se centreront plus exclusivement sur un domaine d'application particulier. Dans ce numéro 1, donc, les trois premiers articles sont consacrés aux problèmes

méthodologiques et théoriques afférents aux corpus textuels, les deux suivants s'intéressent à des corpus lexicologiques.

- 8 En ouverture, J. Guilhaumou nous propose un historique des divers statuts accordés, depuis les années 70, aux corpus en analyse du discours, particulièrement chez les historiens et spécialistes du discours politique. Cet examen critique au fil des ans met en évidence plusieurs périodes successives dans ce champ disciplinaire, marquées par des ruptures épistémologiques importantes. A suivre cette évolution historique, on voit, pour ainsi dire, s'incarner dans les pratiques de l'analyse du discours des rapports dialectiques entre corpus construit et corpus « naturel », entre corpus clos et corpus ouvert, entre corpus et hors-corpus : à chaque décennie sa configuration propre. S'ajoutent à ces questions générales et récurrentes, des problèmes plus spécifiques à la discipline tels que – notamment – le statut qu'il convient d'accorder aux conditions de production (conçues comme paramètre définitoire des entités discursives étudiées, comme variable socio-linguistique contextuelle, etc. ?) ou encore la formalisation du lien avec l'interdiscours constitué de tous les cotextes associés ou associables au discours étudié. Ce dernier point semble avoir joué un rôle déterminant dans la dernière rupture, celle qui a marqué le milieu des années 90 et qui a conduit à la constitution de très grands corpus intégrant en leur sein les ressources nécessaires à leur interprétation, le « contexte dans le texte lui-même », autrement dit à la création des « corpus réflexifs ».
- 9 Cette notion, proposée par D. Mayaffre dans sa thèse, est ici précisée et discutée par son promoteur. En bon représentant de la jeune génération, D. Mayaffre part de deux constats dont l'évidence caractérise la linguistique de corpus actuelle : d'une part les capacités mémorielles des ordinateurs et des outils de stockage, ainsi que les facilités de numérisation des données, permettent au chercheur de ne pas se censurer quant à la taille de ses corpus ; d'autre part, le corpus est désormais pleinement conçu comme un objet heuristique. Or qui dit démarche heuristique dit aussi démarche interprétative. La question est alors de savoir quel statut et quelle place occuperont – par rapport au corpus d'étude initial – les ressources documentaires diverses nécessaires à cette interprétation. En d'autres termes, un corpus textuel peut-il intégrer son intertexte et former « un tout-textuel sémantiquement auto-suffisant » ? Où l'on retrouve les questions de clôture et d'exhaustivité ...
- 10 A. Jaubert revient sur « l'interdépendance qui solidarise l'approche scientifique et l'objet qu'elle s'est donné », à travers l'examen des pratiques de deux disciplines proches, sinon apparentées : l'analyse du discours et la linguistique textuelle. C'est en réalité le point de vue du chercheur sur l'objet étudié qui en définit les contours précis et rattache l'étude à tel ou tel champ disciplinaire. Or ce point de vue non seulement peut être variable, mais peut également assurer une double focalisation sur l'objet : en cela, souligne A. Jaubert, réside « la spécificité de la démarche stylistique qui articule niveau local et niveau global » et qui fait ainsi le pont entre les différentes sciences du discours. L'examen du vocabulaire spécifique de Marivaux d'une part, de quelques ruptures énonciatives entre discours narratif et discours argumentatif d'autre part en apportent l'illustration.
- 11 J.-Ph. Dalbera ouvre le questionnement sur la constitution des corpus lexicologiques à partir de sa double expérience de lexicologue et de dialectologue ; à travers deux exemples précis, il met en évidence la difficile dialectique du donné et du construit, corrigeant quelque peu par là-même les idées reçues sur l'enquête de terrain et le

travail du dialectologue. On verra à quel point en effet les hypothèses théoriques sous-jacentes et le projet heuristique sont déterminants dans la collecte des données, sans parler de leur structuration ultérieure au sein d'une base. Mais, inversement, « la délimitation d'un corpus [peut] 'faire' l'objet et, pour partie au moins, configurer la théorie ». Constat qui, une fois encore, suggère combien la linguistique de corpus oblige le chercheur à une critique réflexive sur ses pratiques.

- 12 Enfin, M. Plénat et ses collaborateurs démontrent, à travers un exemple tout à fait spectaculaire, comment l'extension des données grâce à l'exploitation non seulement des grands corpus numérisés, mais aussi des ressources de la Toile, peut faire évoluer une hypothèse explicative, en l'occurrence une analyse de morphologie dérivationnelle. Les résultats convainquent bien sûr de la nécessité d'une quête aussi exhaustive que possible de données attestées. Mais le travail linguistique à partir des occurrences collectées sur la Toile ne va pas sans susciter aussi diverses difficultés et poser de nombreuses questions : concrètement d'abord, ce type d'exploration nécessite le développement d'outils d'extraction et de tri spécifiques ; beaucoup reste à faire dans ce domaine, pour sortir de l'artisanal. Sur le plan théorique ensuite, M. Plénat souligne qu'une grande vigilance et une réflexion critique sont nécessaires pour évaluer la pertinence de ce recours à un corpus difficilement contrôlable, constamment évolutif et totalement hétérogène. C'est sans doute là un des défis qu'aura à relever la linguistique de corpus dans les prochaines années : la Toile sera une source inépuisable et incomparable d'occurrences de faits linguistiques de toutes sortes, mais quelle sera exactement la qualité des résultats obtenus en la parcourant, quel statut sera-t-on en droit d'accorder aux multiples hapax qu'elle ne manquera pas de produire ?

NOTES

1. François Rastier, « Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage », *Langages* 129 (mars 1998), pp. 97-111 ; voir notamment p. 107.
2. Sylvain Auroux, *La raison, le langage et les normes*, Paris : PUF, 1998 (pp. 166-169 notamment).
3. Sylvain Auroux, *ibid.*, p. 166.
4. Wallace Chafe, « The importance of corpus linguistics to understanding the nature of language », in Jan Svartvik (ed.) *Directions in Corpus Linguistics*, Proceedings of Nobel Symposium 82 (Stockholm, 4-8 August 1991), Berlin – New York : Mouton de Gruyter, 1992, pp. 79-97 (p. 88 en particulier).
5. Geoffrey Leech, « Corpora and theories of linguistic performance », *ibid.*, pp. 105-122 (citation p. 111).
6. Robert Nicolai, *La traversée de l'empirique*, Paris : Ophrys (« Bibliothèque de Faits de Langues »), 2000.

INDEX

Mots-clés : échantillonnage, corpus, clôture, exhaustivité, observatoire, représentativité

AUTEUR

SYLVIE MELLET

« Bases, Corpus et Langage », UMR 6039, CNRS <http://www.unice.fr/bcl/tmp/9accueil.htm>