Des corpus numériques à l'analyse linguistique en langues de spécialité

Langues, Gestes, Paroles

Collection dirigée par Agnès Tutin et Nathalie Vallée

La collection « Langues, Gestes, Paroles » se propose d'accueillir des ouvrages relevant du champ des Sciences du langage, et situés dans un large domaine de recherche incluant des travaux de descriptions et traitements linguistiques ainsi que des travaux ancrés dans les thèmes de la parole, de l'acquisition et de la multimodalité. La collection a pour objectif de faire le point sur les derniers développements des connaissances dans ces domaines.

Éléments de catalogage
Des corpus numériques à l'analyse linguistique en langues de spécialité. Sous la direction de Cécile Frérot et Mojca Pecman. 374 p. : couv. ill. en coul. ; 21,5 cm. Collection « Langues, Gestes, Paroles », ISSN 2105-9497 — ISBN 978-2-37747-261-1

Ce travail a bénéficié du soutien du Centre de Linguistique Inter-langues, de Lexicologie, de Linguistique Anglaise et de Corpus-Atelier de Recherche sur la Parole (CLILLAC-ARP) de l'Université de Paris, de l'ILCEA4 et du programme IDEX Université Grenoble Alpes

Université Grenoble Alpes
CS 40700
38058 GRENOBLE CEDEX 9

Des corpus numériques à l'analyse linguistique en langues de spécialité

Sous la direction de
Cécile Frérot et Mojca Pecman

UGA Éditions
Université Grenoble Alpes
Grenoble
2021

Préface

Natalie Kübler

> Bas Aarts: *What is your view of modern corpus linguistics?*
> Noam Chomsky: *It doesn't exist*
> (Aarts 2000, p. 5[1])

En 2021, cette réponse, devenue célèbre, de Noam Chomsky à Bas Aarts au cours d'un entretien, a bien vieilli. En effet, si Noam Chomsky avait été étudiant dans les années quatre-vingt-dix du siècle dernier, il aurait peut-être pensé les choses autrement. Il n'est plus nécessaire, aujourd'hui, de justifier l'existence de la linguistique de corpus puisque celle-ci a percolé dans toutes les disciplines des Sciences du langage, de la linguistique la plus formelle à la plus appliquée !

Si la linguistique de corpus s'est largement déployée, dès la fin des années quatre-vingt, dans le monde anglophone et dans les pays du Nord de l'Europe, elle a pris place plus tardivement en France.

Parallèlement à l'ouvrage fondateur de John Sinclair, se constituent différentes approches faisant appel aux corpus, telles que la théorie des universaux de traduction chez Mona Baker, le *data-driven learning* chez Tim Johns, la lexicographie bilingue chez Wolfgang Teubert ou encore les études sur la néologie menées par Antoinette Renouf. En France, ce n'est que vers la fin des années quatre-vingt-dix que l'on commence à constituer des corpus électroniques et à les interroger à l'aide d'outils, mais les études portent surtout sur la langue générale. Des pans entiers de la langue restent encore peu ou pas étudiés : les langues et discours spécialisés.

On voit alors apparaître progressivement des travaux sur corpus dans les domaines spécialisés en français, notamment avec les travaux d'Anne Condamines, de Monique Slodzian et de Didier Bourrigault

1. Aarts, Bas, 2000, « Corpus linguistics, Chomsky and Fuzzy Tree Fragments », dans C. Mair and M. Hundt (dir.) *Corpus Linguistics and Linguistic Theory*, Amsterdam, Rodopi, p. 5-13.

Introduction à la place des corpus et des outils numériques pour l'étude des langues de spécialité

Cécile Frérot
Mojca Pecman

Dans le numéro inaugural de la revue *Corpus* paru en 2002, Sylvie Mellet attribuait au corpus la capacité qu'a cet objet de représenter une « médiation consciente entre le chercheur et le fait linguistique » (Mellet, 2002, p. 9), en faisant un « lieu de confrontation entre la théorie et l'empirie » (Mayaffre, 2005, p. 5). C'est autour des années 2000 que l'introduction des corpus en France s'est progressivement imposée dans les cercles linguistiques, comme le soulignait alors Damon Mayaffre : « Plus une discipline, plus un comité scientifique, plus un chercheur qui n'y fasse référence » (ibid.). La linguistique de corpus, telle qu'elle s'envisage depuis plus de vingt ans dans le monde anglo-saxon (e.g. Sinclair, 1991 ; Biber, Conrad & Reppen, 1998 ; Tognini-Bonelli, 2001 ; Aijmer & Altenberg, 2002) ou en France, sous l'impulsion d'une communauté de linguistes (e.g. Habert, Nazarenko & Salem, 1997 ; Kübler, 2003a ; Rastier, 2005 ; Williams, 2005) a ainsi fortement contribué à admettre cet objet qu'est le corpus comme un observable nécessaire en linguistique.

Dans ce contexte, la linguistique de corpus a orienté les recherches dans toutes les disciplines des Sciences du langage vers de nouveaux horizons. Elle a offert aux linguistes la possibilité de fonder leurs recherches sur des faits de langue authentiques observés en corpus et de pratiquer leur science de manière heuristique. Damon Mayaffre qualifie à ce titre la linguistique de corpus de « puissance heuristique » : elle « doit permettre de nourrir l'interprétation et la théorisation […] et susciter de nouvelles hypothèses de travail […] grâce à de nouveaux outils » (Mayaffre, 2011, p. 325). L'avènement du numérique a par ailleurs profondément modifié le rapport au texte et l'analyse de corpus textuels, la lecture linéaire étant désormais combinée à des lectures « tabulaire et réticulaire » (Viprey, 2005) dans la perspective d'une philologie et/ou herméneutique numérique où les logiciels d'analyse de données textuelles

font « exploser la linéarité du texte pour présenter leurs données en tableaux : tableaux alphabétiques, tableaux de fréquences, tableaux de distance » (Mayaffre, 2007, p. 18). Le corpus aiguise ainsi le regard du linguiste qui doit désormais se focaliser sur des outils et instruments, tout à la fois méthodologiques et théoriques, et sur son objet d'étude premier qu'est la langue (lexique, syntaxe, sémantique, discours, entre autres). Qui plus est, les linguistes ont de nos jours l'opportunité de s'interroger sur la manière dont ce croisement leur ouvre de nouvelles perspectives pour décrire et modéliser la langue.

Ces propos ne sauraient en aucun cas occulter les deux approches, parfois opposées, parfois complémentaires, qui envisagent dans un cas le corpus comme le terrain d'observations d'une théorie construite *a priori*, dans l'autre le terrain d'observations permettant de définir des modèles *a posteriori*. En effet, la linguistique de corpus offre la possibilité de cette double approche bien connue des linguistes qui explicitent généralement leur recours à l'une ou l'autre, voire aux deux, « *corpus-based* », c'est-à-dire fondée sur corpus, et « *corpus-driven* », orientée par corpus (Biber, 2009).

L'histoire de la linguistique de corpus est ainsi marquée par des interrogations épistémologiques portant sur l'objet « corpus » dans les Sciences du langage. Chemin faisant, les linguistes se sont emparés des corpus pour en faire un outil de prédilection. Appréhendés en tant qu'objet théorique, les corpus sont devenus des outils méthodologiques incontournables pour acquérir ensuite un statut en tant qu'outils pédagogiques. Tout au long de cette évolution, les travaux de Natalie Kübler ont joué un rôle essentiel, figurant au premier plan dans l'introduction des corpus dans l'enseignement universitaire en France. Déjà, en 1999, Kübler et Foucou (Foucou & Kübler, 1999 ; Kübler & Foucou, 2000), expérimentaient avec la création du projet Wall[1] l'un des premiers concordanciers en ligne en anglais de spécialité, permettant de générer automatiquement des exercices pour l'enseignement de l'anglais aux informaticiens. Avec ce projet, Natalie Kübler développe les premiers corpus spécialisés interrogeables en ligne adaptés à des objectifs scientifiques et pédagogiques spécifiques, à savoir l'analyse terminologique en vue de la traduction (Kübler, 2003a), dessinant les prémices des

1. La page Internet consacrée au projet Web-Assisted Language Learning (WALL) est accessible en ligne sur https://www.eila.univ-paris-diderot.fr/recherche/wall/index [consulté en décembre 2020].

études en langues de spécialité fondées sur corpus. Depuis ces premières expérimentations avec les corpus (ibid., Kübler 2002a/b/c, 2003a/b, 2008, 2011 ; Kübler & Frérot, 2003) jusqu'à des travaux plus récents sur l'évaluation de leur efficacité pour l'enseignement de la traduction spécialisée (e.g. Kübler, Loock & Pecman, 2018 ; Kübler, Mestivier & Pecman, 2018), elle n'a cessé de creuser le sillon de la linguistique de corpus en langues de spécialité. Dans cette dynamique, ses travaux, comme ceux d'Agnès Tutin et d'Olivier Kraif sur le lexique et la phraséologie transdisciplinaire (e.g. Kraif, 2001, 2008, 2016 ; Tutin, 2005, 2007 ; Tutin & Grossmann, 2014) ainsi que ceux d'autres chercheurs (notamment Condamines, 2002 ; Boulton, 2016 ; Williams, 2003) ont contribué à conférer à ce champ disciplinaire le statut et l'assise scientifique qu'on lui reconnaît désormais.

En langues de spécialité, les recherches appuyées par corpus sont devenues le paradigme quasi incontournable puisque la multiplicité des langues et des discours spécialisés est plus facilement accessible et observable depuis que la linguistique de corpus a mis à la disposition des chercheurs des ressources textuelles sur support numérique, et des outils informatiques pour les interroger et les analyser. Bien que les recherches en langues de spécialité comme en linguistique de corpus témoignent d'une incontestable maturité et vitalité, il n'existe pas encore d'ouvrage qui offre un regard croisé sur ces deux disciplines et qui se donne comme objectif de mettre en lumière les problématiques et les pistes de recherche qui émergent de leurs interactions. Les quelques rares ouvrages abordent parfois certains aspects des discours spécialisés à l'ère du numérique ; c'est le cas de *Languages for Specific Purposes in the Digital Era* (Bárcena, Read & Arús, 2014) consacré notamment aux approches pédagogiques des langues de spécialité. Il nous a ainsi paru nécessaire d'envisager un ouvrage axé sur la problématique de l'analyse des langues de spécialité orientée par ou fondée sur des corpus numériques. C'est là tout l'objet du présent ouvrage, qui s'adresse naturellement aux chercheurs qui s'intéressent aux langues de spécialité et à la linguistique de corpus, que ce soit dans une perspective de traitement automatisé des langues visant leur modélisation, ou bien dans des perspectives à visée didactique, lexicologique ou encore traductionnelle. L'ouvrage vise à offrir à l'ensemble de ces communautés de chercheurs, et d'enseignants-chercheurs, la possibilité d'explorer la richesse et la diversité des analyses qu'offre la rencontre de deux objets d'étude : les corpus et les langues de spécialité. Par ailleurs, les étudiants en troisième cycle trouveront