ANalyse et exploitation des données de corpus linguistiques : présentation

Gabriel Bergounioux, Bernard Colombat, Jacqueline Léon

▶ **To cite this version:**

Gabriel Bergounioux, Bernard Colombat, Jacqueline Léon. ANalyse et exploitation des données de corpus linguistiques : présentation. Dossiers d'HEL, 2017, Analyse et exploitation des données de corpus linguistiques, 11, pp.3-6. hal-01511187

HAL Id: hal-01511187
https://hal.science/hal-01511187

Submitted on 24 Apr 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DOSSIERS D'HEL – 11

ANALYSE ET EXPLOITATION DES DONNÉES DE CORPUS LINGUSTIQUES : PRÉSENTATION

Gabriel BERGOUNIOUX, Bernard COLOMBAT, Jacqueline LÉON

Cette publication des *Dossiers d'HEL*, supplément électronique à la revue *HEL*, comporte huit articles issus des communications du colloque SHESL-HTL 2015 « Corpus et constitution des savoirs linguistiques ». Ce colloque a eu lieu les 30 et 31 janvier 2015 à Paris et a été co-organisé par la SHESL, le laboratoire d'Histoire des Théories Linguistiques (UMR 7597) et le Laboratoire Ligérien de Linguistique (UMR 7270), sous la direction de Gabriel Bergounioux, Bernard Colombat et Jacqueline Léon (cf. l'appel à colloque et le programme [http://www.shesl.org/spip.php?rubrique76]). Une autre partie des communications a fait l'objet d'une publication dans le numéro XXXVIII/2 d'*Histoire Épistémologie Langage*, « Corpus et constitution des savoirs linguistiques et pérennisation des données » (décembre 2016). Nous en donnons la liste à la fin de cette présentation.

La référence aux corpus est devenue l'une des orientations méthodologiques majeures de la linguistique contemporaine en lien avec le développement de la numérisation et le recours aux outils de traitement automatique. Pour en donner un exemple dans l'actualité scientifique, on a constaté en quelques années la création de la TGIR Humanités Numériques (Huma-Num) déclinée en plusieurs consortiums, d'un Equipex (Ortolang) et d'un appel de l'ANR (Corpus en SHS). Avec le projet Huma-Num et la mise en place de DARIAH et de CLARIN, c'est au niveau européen que la question se trouve transposée.

Le travail sur des données destinées à l'établissement, la collation, la vérification et l'analyse des faits linguistiques est une pratique ancienne. Elle correspond d'abord à une tradition philologique et exégétique, ininterrompue de l'Antiquité à nos jours, qui reste liée à la fondation des bibliothèques et des dépôts d'archives comme à la rédaction des compilations (par ex. les Alexandrins, les Bénédictins). Cette relation des lettrés au classement et à l'exploitation des documents se retrouverait dans la plupart des civilisations, en particulier en Orient.

Avec l'expansionnisme européen, l'accumulation – qui existe dans d'autres traditions – s'est étendue à un travail de description des langues que transforment l'usage des techniques d'enregistrement (à la fin du XIX^{e} siècle) et l'application, sur les récits recueillis, de méthodes de transcription et de segmentation pour lesquelles le *Handbook of American Indian Languages* demeure emblématique.

L'automatisation des corpus commence dans les années 1960 et pose les questions d'échantillonnage (*vs* textes intégraux), de recherche systématique de structures. À partir de la fin des années 1980, de grandes masses de données sont devenues disponibles grâce au développement technologique des ordinateurs et à un

perfectionnement des logiciels. Parallèlement, la numérisation des ouvrages légitime les entreprises d'accumulation des sources écrites et des documents sur la représentation des langues, comme le montre l'exemple du *CTLF* et du *Corpus des grammaires françaises*, affectant, après les langues, le métalangage.

*

Les deux premiers articles de ce dossier sont consacrés aux travaux de constitution de corpus. Gerda HAßLER évoque l'établissement du corpus qui a servi de soubassement au grand *Lexique des principaux concepts linguistiques des XVII^e^ et XVIII^e^ siècles* paru en deux volumes chez Walter de Gruyter en 2009, sous sa direction et celle de Cordula Neis. Par l'étude de concepts comme ceux d'« analogie », de « génie de la langue », qui manifestent les difficultés de la conceptualisation, elle montre ce que peut apporter une démarche proprement sémasiologique appuyée sur une base de données spécifique à la connaissance de l'élaboration du métalangage linguistique.

Autre projet de grande envergure, promu par une équipe de romanistes sous la direction d'Anne-Marie CHABROLLE-CERETINI, le *Dictionnaire HIstorique des COncepts Descriptifs de l'Entité Romane* (D.HI.CO.D.E.R.) a un objectif assez similaire puisqu'il vise également à recenser les grands concepts qui ont permis de décrire l'entité romane depuis le XIX^e^ siècle. Les auteurs établissent le recensement des dictionnaires et des corpus existants, définissent les zones informationnelles de l'outil, avant de fournir trois exemples de résultats, matérialisés dans les articles « panroman », « dialecte » et « limousin ». Ils espèrent ainsi éclairer sous un nouveau jour l'histoire de la description de l'entité romane.

Quatre articles sont consacrés à des analyses de corpus. Catherine PINON se penche sur la question de savoir comment la mise à disposition de nouveaux corpus de l'arabe permettra de renouveler la description d'une langue fondée sur une longue tradition grammaticale. S'appuyant sur des exemples précis (l'expression des modalités du nécessaire et du possible, l'expression de la négation du passé), l'auteure montre comment l'exploitation de « corpus synchroniques contemporains et philologiques des différentes époques » permettra dorénavant de dépasser l'impasse d'une transmission sclérosée fondée sur la simple accrétion.

C'est également dans les corpus, mais ceux de français oral, que Mireille BILGER et Paul CAPPEAU voient le nécessaire renouvellement des grammaires françaises. Après une présentation des corpus utilisés, les auteurs s'attachent à l'usage de *contre*, notamment dans la locution *par contre*, et de *même*, notamment sous les formes *quand même, quand bien même, même pas*. Les données quantifiées fournies par les corpus examinés permettent de réviser des affirmations réductrices données par les grammaires ou les dictionnaires existants.

Christiane MORINET étudie un corpus de copies de lycéens français en vue de la construction d'une représentation linguistique de l'acquisition. Elle étudie les formes d'ancrage énonciatif, effacement des pronoms de première personne et usage du conditionnel, et leur transformation du parler à l'écrit, afin de déterminer l'écart entre les échanges informels et les données écrites.

Vanise MADEIROS analyse les glossaires sur le plan discursif, en tant que métatextes dans la littérature lusophone (en l'occurrence deux écrivains brésiliens et un écrivain angolais). Elle distingue les glossaires produits par l'écrivain qui prend alors la position de lexicographe, et ceux produits par l'éditeur associés à la diffusion des ouvrages au sein de la lusophonie.

Enfin, Rossana DE ANGELIS discute les différentes analyses critiques qui s'intéressent aux transformations du rapport entre texte, corpus, discours, impliquées par le développement du numérique, notamment la dématérialisation et la délinéarisation du texte. Elle montre en particulier comment l'approche herméneutique numérique permet d'appréhender ces nouvelles formes textuelles, et donne l'exemple du site *The World of Dante*, outil multimedia pour l'étude de la *Divina Commedia* de Dante. Elle examine enfin les différentes conceptions et méthodologies des linguistiques de corpus, comme la textométrie, la lexicométrie et la logométrie, permettant de mettre en œuvre une telle herméneutique numérique.

Les Dossiers se terminent par une table ronde réunissant cinq intervenants et un modérateur (Emilie AUSSANT, Marc BARATIN, Franck CINATO, Anne GRONDEUX, Cendrine PAGANI-NAUDET et Bernard COLOMBAT) autour du thème des corpus linguistiques avant les corpus informatisés. Les interventions des auteurs ont donné lieu à une discussion dont le compte rendu a été établi par Pascale RABAULT-FEUERHAHN.

SOMMAIRE DU N° XXXVIII/2 D'*HISTOIRE ÉPISTÉMOLOGIE LANGAGE*

CONSTITUTION DE CORPUS LINGUISTIQUES ET PÉRENNISATION DES DONNÉES

Gabriel BERGOUNIOUX, Bernard COLOMBAT, Jacqueline LÉON : Présentation 5

Maëlle AMAND. La constitution d'un corpus de geordie parlé : choix épistémologiques et réalisations empiriques. Retour sur un demi-siècle de sociophonétique anglaise 9

Nicolas BALLIER. Du dictionnaire lexico-phonétisé aux corpus oraux, quelques problèmes épistémologiques pour l'école de Guierre 23

Gabriel BERGOUNIOUX. La linguistique de corpus et la partition des structuralismes 41

Wendy AYRES-BENNETT & Bernard COLOMBAT. L'extension du *Grand Corpus des grammaires françaises, des remarques et des traités sur la langue.* Questions théoriques et méthodologiques 55

Rachel PANCKHURST, Mathieu ROCHE, Cédric LOPEZ, Bertrand VERINE, Catherine DÉTRIE & Claudine MOÏSE. De la collecte à l'analyse d'un corpus de SMS authentiques : une démarche pluridisciplinaire 73

Marie-Paule JACQUES. Une linguistique outillée, pour quels objets ? 87

Pascal CORDEREIX. Comment indexer les corpus oraux ? 101