Histoire Épistémologie Langage
38/2 (2016), p. 5-8

DOI : 10.1051/hel/2016380201

Disponible en ligne sur :
www.hel-journal.org

Présentation
Constitution de corpus linguistiques et pérennisation des données

Gabriel Bergounioux, Bernard Colombat, Jacqueline Léon

Ce numéro d'*HEL* est constitué d'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 (UMR7597) 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 et la table ronde font l'objet d'une publication dans les *Dossiers d'HEL*, supplément électronique à la revue *HEL.* Voir la liste des articles ci-dessous.

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.

Plusieurs questions ont été abordées dans les articles composant ce numéro :

- Quels critères permettent de définir de grandes masses de données comme étant des « corpus » ? Quels sont les fondements épistémologiques des corpus de référence et quels sont leurs principes de légitimation ? Quels sont les enjeux pour les langues à tradition écrite de la place croissante donnée aux corpus oraux ?
- Une approche fondée sur des corpus peut-elle être qualifiée de strictement empirique ou bien répond-elle à des exigences théoriques spécifiques ? Dans quelle mesure des théories et/ou des écoles se sont-elles appuyées sur des corpus ?
- Quel est le statut des corpus en tant que producteurs de données dans la construction d'une représentation linguistique ? En quoi l'élaboration de corpus implique-t-elle une instrumentation des langues (par ex. les outils de transcription) et en quoi ces outils permettent-ils de modifier, d'infléchir ou d'affiner cette représentation ?

Ce numéro comporte à peu près pour moitié des articles portant sur les corpus contemporains et pour moitié des articles historiques. À l'exception de l'article de G. Bergounioux, tous traitent de corpus numérisés et informatisés.

Trois articles abordent la question de l'importance des corpus pour certaines théories et écoles linguistiques. Maëlle Amand traite de la constitution d'un corpus d'anglais parlé, le *Tyneside Linguistic Survey* à partir des années 1960. Celui-

ci marque une étape dans le développement des corpus dans la linguistique en Angleterre, entre l'apparition de l'école anglaise de dialectologie en 1873, et l'essor de la *Corpus Linguistics* dans les années 1990. Ce type de corpus met en jeu un grand nombre de variables linguistiques, contrairement à l'approche labovienne alors dominante, permettant une analyse plus complexe qui ne privilégie aucune variable explicative a priori.

Nicolas Ballier s'intéresse à une école de phonologie anglaise, celle de Lionel Guierre, apparue à la fin des années 1960. À partir de la numérisation du dictionnaire de prononciation de Daniel Jones, Guierre a établi des règles de placement accentuel en anglais sur la base de critères linguistiques (et pas seulement du signal sonore) qui constitue une critique de *Sound Pattern of English* de Chomsky et Halle reposant sur de l'oral transcrit, autrement dit de l'écrit.

Gabriel Bergounioux établit une genèse de la conception des corpus, dont la première acception apparaît au début du XIX^e^ siècle, en comparant le statut des corpus pour les trois structuralismes, russe, américain et français. Pour les Américains, la notion de corpus est liée à une linguistique de terrain aux préoccupations anthropologiques, alors qu'elle participe à une visée plus politique en Russie. Pour le structuralisme français, l'objectif est une reconstruction à partir d'un corpus fermé de langues mortes.

Deux articles émanent de « producteurs » de grands corpus informatisés, outillés et accessibles en ligne pour un large public. Les auteurs, Rachel Panckhurst *et al.* pour un corpus de SMS, et Wendy Ayres-Bennett et Bernard Colombat pour un corpus de grammaires françaises du XVI^e^ au XVIII^e^ siècles, s'interrogent sur leurs conditions de constitution, d'analyse et de traitement. La constitution de ces deux types de corpus soulève des problèmes tout à fait distincts. W. Ayres-Bennett et B. Colombat se posent la question de la représentativité (notamment les difficultés posées par la notion de canon), de l'exhaustivité, et des enjeux qui se trouvent déplacés quand on étend le corpus.

Un des principaux problèmes que pose Rachel Panckhurst *et al.* est celui du transcodage qui ne peut être universel mais dépend de l'approche théorique adoptée ainsi que du type d'analyse envisagée. L'étude de l'évolution de la langue (française) constitue un objectif commun pour ces deux projets : l'histoire de la langue sur le temps long, plusieurs siècles, pour l'un ; la variation entre langue écrite et langue orale sur le temps court pour les SMS.

L'article de Marie-Paule Jacques pose la question plus générale du type d'approche linguistique rendue possible par les corpus, quand on dépasse l'opposition classique inductive/logico-déductive. En mettant au premier plan le rapport des formes au sens, l'auteur montre que l'approche onomasiologique, du sens vers les formes, consistant à « comprendre comment ça se dit » est une orientation moins

prisée par les linguistes que par les acteurs de la linguistique appliquée qui ont des besoins plus concrets. En revanche l'approche sémasiologique, qui part des formes vers le sens, impose pour les linguistes plus de contraintes et une sélection plus stricte des données de corpus.

Enfin l'article de P. Cordereix, en explorant les procédures techniques mises en œuvre, pose la question de la pérennité des archives sonores. Il retrace l'histoire de l'automatisation de la documentation, concernant notamment la recherche d'un système de classification universelle, la standardisation de l'information bibliographique et son informatisation, en montrant la nécessité d'une interopérabilité entre les acteurs et les institutions permettant des échanges à l'extérieur des domaines de spécialité. L'exemple de la BnF permet de saisir les contraintes imposées par la conservation d'un patrimoine phonographique.

Les articles publiés dans les *Dossiers d'HEL* n° 10 « Analyse et exploitation des données de corpus linguistiques » sont les suivants :

– Mireille Bilger et Paul Cappeau, « L'apport des corpus aux grammaires et/ou aux dictionnaires : l'exemple de *contre, même* et *entre »*
– Anne-Marie Chabrolle-Cerretini, Cyril de Pins, Narcís iglésias franch, Christophe Rey, « Repenser l'histoire de la linguistique romane par la constitution de nouveaux corpus : l'expérience du projet *Dictionnaire HIstorique des COncepts Descriptifs de l'Entité Romane* (D.HI.CO.D.E.R.) »
– Rossana De Angelis, « Textes et *documents* dans l'analyse des corpus. Nouveaux objets pour la linguistique ? »
– Gerda Hassler, « Les corpus métalinguistiques et l'histoire conceptuelle des théories linguistiques – une contradiction ? »
– Vanice Medeiros, « Les glossaires brésiliens dans la littérature : les savoirs sur la langue »
– Christiane Morinet, « De l'empirique au théorique ou de la difficulté à objectiver les '*phénomènes peu visibles'* » : le cas de l'implication cognitive de l'énonciateur dans l'acquisition »
– Catherine Pinon, « Corpus et langue arabe : un changement de paradigme »
– Table ronde : « Des corpus linguistiques avant les corpus électroniques ? ». Intervenants : Émilie Aussant / Marc Baratin / Franck Cinato / Anne Grondeux / Cendrine Pagani-Naudet. Modérateur : Bernard Colombat. Compte rendu : Pascale Rabault-Feuerhahn.