Colloque SHESL-HTL 2015
« Corpus et constitution des savoirs linguistiques »

Actualité des corpus linguistiques

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 constate qu'en l'espace de deux ans en France, ce domaine a bénéficié d'une Infrastructure de Recherche (IR Corpus) 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.

Questions possibles

L'intérêt actuel pour les corpus correspond-il à une réflexion épistémologique ?

Une approche fondée sur des corpus peut-elle être qualifiée de strictement empirique ou répond-elle à des exigences théoriques spécifiques ?

Quel est le statut des corpus en tant que producteurs de données dans la construction d'une représentation linguistique ? En quoi contribuent-ils à la cumulativité des connaissances ?

En quoi l'élaboration de corpus implique-t-elle une instrumentation des langues (par ex. les outils de transcription) et comment participent-ils à la mise au point de grammaires, à la rédaction de dictionnaires et aux développements du TAL ?

Quels sont les effets des données de corpus sur les théories et les écoles ? Quels manques résulteraient de leur absence ? Ces données sont-elles centrales, incontournables, exclusives, ou, au contraire, complémentaires, annexes, périphériques, etc. ? Quelles preuves peut-on en donner dans les recherches actuelles ?

L'accroissement et la diversification des données apportées par les corpus contribuent-ils à l'amélioration de la théorisation ?

Histoire du travail sur les données

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.

Questions possibles

Quelle signification accorder au changement d'échelle selon que le travail sur les langues se fait au travers d'échantillons limités ou de corpus importants ? En d'autres termes, peut-on parler de « grammaires de corpus », qui proposent une représentation des langues en extension, l'objet produit constituant par lui-même un corpus ? Si l'existence de tels objets est avérée, depuis quand existent-ils et dans quelles traditions ? Inversement, comment se décide la justification de théories qui se dispensent d'un tel recours aux données ?

Comment et pourquoi se sont constitués les corpus d'inscriptions ? Les données épigraphiques par exemple répondent-elles simplement à un besoin de recensement et d'exhaustivité, ou posent-elles des problèmes linguistiques pour ceux qui les ont créées et/ou ceux qui les exploitent ?

Extension aux langues du monde

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. Dans son acception moderne, la linguistique de corpus semble connaître un nouvel essor et une nouvelle définition dans une linguistique de terrain aux préoccupations anthropologiques (aux États-Unis) ou à visée de planification (en Russie).

Questions possibles
Au moment de la grammatisation massive des vernaculaires, comment s'est effectué le passage d'échantillons limités (les collections de Notre Père par exemple) à des données plus importantes ?
Quels ont été les principes opératoires dans la création des nouvelles données ? Comment ces données sont-elles organisées et quel est leur statut ? En quoi leur finalité est linguistique (et non folklorique ou ethnographique) ? En quoi se distinguent-elles d'autres outils comme la cartographie ?

La constitution des grandes bases de données modernes
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.
Questions possibles
Quels sont les critères qui permettent de définir ces grandes masses de données comme étant des « corpus » ? En quoi induisent-elles un changement dans l'écologie de la pratique linguistique, notamment dans la division du travail scientifique ?
Quels sont les fondements épistémologiques des corpus de référence et quels sont leurs principes de légitimation ?
En quoi la place croissante donnée aux corpus oraux dans les langues à tradition écrite peut-elle aboutir à un renversement des perspectives ?

Les « métacorpus »
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.
Questions possibles
Quel est le rôle de ces outils dans la construction de la représentation des langues ?
Permettent-ils de modifier, d'infléchir ou d'affiner cette représentation ?