Construction et exploitation d'un grand corpus historique annoté du portugais: Le Corpus Tycho Brahe

Charlotte Galves, Université de Campinas, Brésil
Section 16

Selon Labov (1994), la linguistique historique est "l'art de faire un bon usage de mauvaises données". Pour la grammaire générative, cependant, il a été longtemps assumé qu'un bon usage de données historiques était impossible puisque celles-ci ne sauraient donner au linguiste la part essentielle des évidences empiriques nécessaires à son argumentation: les évidences négatives. Marianne Adams, dans sa thèse de 1987, affirme cependant que la théorie générative, à cette époque, a atteint un degré de maturité suffisant pour que l'on soit à même de poser les bonnes questions aux données uniquement positives, et que l'on puisse, à partir des énoncés, inférer la compétence, ou grammaire, de leurs auteurs. Dans cette communication, je me propose de montrer comment, 25 ans après le travail pionnier d'Adams, l'avènement des grands corpus électroniques annotés permet de faire un usage de plus en plus fécond, tant d'un point de vue empirique que théorique, des données fragmentaires que le temps a conservé de façon aléatoire.

Je présenterai le *Corpus annoté du portugais historique Tycho Brahe*, librement accessible sur la toile mondiale, construit sur le modèle des *Penn Corpora of Historical English* dirigés par Anthony Kroch et ses collaborateurs (Pour le français, voir aussi Martineau 2008). Le schéma d'annotation syntaxique de ces corpus s'inspire de la théorie de la grammaire générative. Les représentations suivent une version simplifiée de la théorie X-barre et contiennent des éléments abstraits comme catégories vides et indices référentiels. Un élément clef du dispositif est l'outil de recherche *Corpus Search* qui permet de rechercher sur les arbres ainsi étiquetés, outre des relations de précédence, des relations de dominance et de c-commande. La philosophie de ce genre de Corpus, qui consiste à mettre à disposition des chercheurs les textes dans leur intégralité, permet également de travailler de façon très productive sur la relation syntaxe/texte, notamment la relation entre les phénomènes d'ordre et

l'organisation de l'information, domaine très actif de recherche actuellement (cf., entre autres, Hinterholzl & Petrova, 2009). Il en est de même pour la relation genre/syntaxe ou encore style/syntaxe. Enfin, ces grands corpus syntaxiquement annotés permettent d'étudier de façon de plus en plus précise les dynamiques de changement, non seulement grâce à la quantité croissante de données, mais parce qu'ils permettent d'intégrer à la quantification de ces données une approche théorique plus riche.

La présentation du Corpus sera accompagnée d'un échantillon de résultats obtenus à partir des données actuellement disponibles, portant sur l'évolution syntaxique du portugais européen et du portugais brésilien du 16è au 20è siècle.

Adams, Marianne (1987) Old French, null subjects and Verb-Second phenomena, PhD Dissertation, UCLA.

Corpus Search http://corpussearch.sourceforge.net/

Corpus Tycho Brahe http://www.tycho.iel.unicamp.br/~tycho/corpus

Hinterholzl, Roland and Petrova, Svetlana (2009) Information Structure and Language Change, Berlin : Walter de Gruyter.

Labov, William (1994) Principles of Historical Change, Volume 1: Internal factors. Oxford : Blackwell.

Martineau, France. (2008). Corpus MCVF : Modéliser le changement: les voies du français. U. Ottawa, http://www.voies.uottawa.ca/corpus_pg_fr.html

Penn Corpora of Historical English http://www.ling.upenn.edu/hist-corpora