L'utilisation de « big data » en linguistique de corpus: enjeux méthodologiques, mutlidimensionnalité et exemple

Résumé

La linguistique de corpus, telle que pratiquée aujourd'hui, est en perpétuelle adaptation à une croissance quasi exponentielle du canon littéraire. En effet, l'accès à des données linguistiques est aujourd'hui favorisé par les technologies de numérisation et d'annotation semi-automatiques d'anciens textes et d'œuvres modernes, mais aussi grâce à l'accès instantané à l'information via les ordinateurs, tant et si bien que la représentativité des phénomènes linguistiques ne sera plus un problème dans le futur, que du contraire. À titre d'exemple, Mark Davies (2010) a compilé le COHA (Corpus of Historical American English), un corpus de plus de 400 000 000 de mots et qui représente l'anglais américain depuis 1810 jusque 2009. Pour le français, l'Université de Lorraine, en association avec l'ATILF et le CNRS, a développé la base textuelle FRANTEXT (s.d.), qui compte près de 286 000 000 de mots et couvre la période du 10^{e} siècle à aujourd'hui. Assez récemment, le projet *nederlab* a vu le jour pour le néerlandais, avec pour but de permettre à chaque chercheur en sciences humaines d'effectuer des recherches tant linguistiques que culturelles ou littéraires.

Avec l'avènement de ces larges corpus, nous nous adaptons progressivement à l'ère du « big data », faisant entrer la langue dans des archives digitales massives. L'issue de la représentativité étant résolue, un autre problème survient: celui du traitement des données linguistiques. En effet, le linguiste doit faire face à une masse d'informations qu'il doit traiter pour dégager des structures de ces données, or les phénomènes linguistiques sont, pour la plupart, multifactoriels (Gries 2013: 20). Au travers d'une étude sur la généralisation (Traugott et Trousdale 2013) dans les sous-catégories des connecteurs complexes en anglais (Béchet & Brems 2016), qui visait à déterminer si des schémas hiérarchiques ont été construits entre ces constructions, il sera démontré que l'analyse qualitative et les méthodes univariées ne permettent pas une représentation fidèle de ce phénomène, a fortiori lorsque la dimension de temps est prise en compte. Sur base d'extractions du corpus COHA, les connecteurs exprimant la substitution (par exemple, *in place of* et *in lieu of*) (voir Schwenter & Traugott 1995), le but (*in the hope of*, *for fear of*) (voir Brems & Davidse 2010), la cause (*for want of*, *for lack of*) et la concession (*in spite of*) seront analysés sur des critères syntaxiques et sémantiques, en incluant la variation de genre textuel et leur évolution sur la période allant de 1810 à 2009.

En vue de pouvoir visualiser le phénomène de généralisation au cours du temps et en tenant compte de tous les paramètres précités, les données linguistiques seront soumises à la technique du positionnement multidimensionnel ou « Multidimensional Scaling » (voir Levshina 2015), un procédé qui permet de visualiser en deux dimensions les distances et les (dis)similarités entre les constructions étudiées. La dimension de temps étant difficilement analysable de manière instantanée, les graphes qui représentent chaque décennie seront visualisés de manière séquentielle au travers d'un graphique linguistique dynamique (Hilpert 2011) généré avec R (Version 3.1.3, R Development Core Team 2015). Cette technique permettra à l'audience d'interpréter de manière intuitive le positionnement des constructions sur le graphe et d'observer à quel degré de généralité on peut classifier les connecteurs logiques.

Sources

ATILF - CNRS et Université de Lorraine (s.d.). *Base textuelle FRANTEXT*. Accessible sur http://www.frantext.fr.

Béchet, C. et Brems, L. (2016, mai). *From complex prepositions to complex subordinators: Constructional motivations and generalizations*. Communication présentée à la Journée Linguistique du CBL (Cercle Belge de Linguistique), Louvain-la-Neuve, Belgique.

Brems, L. et Davidse, K. (2010). Complex subordinators derived from noun complement clauses: grammaticalization and paradigmaticity. *Acta Linguistica Hafniensia*, 42(1), 101-116.

Davies, M. (2010). *Corpus of Historical American English (COHA)*. Dernier accès le 14 avril, 2016, sur http://corpus.byu.edu/coha/.

Gries, S. T. (2013). *Statistics for Linguistics with R: A Practical Introduction*. Berlin: Mouton de Gruyter.

Hilpert, M. (2011). Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics*, 4(16), 435-461.

Levshina, N. (2015). *How to do Linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins Publishing Company.

R Core Team (2015). R: A language and environment for statistical computing [Logiciel]. R Foundation for Statistical Computing, Vienne, Autriche. Accessible sur http://www.R-project.org/.

Schwenter, S. A. et Traugott, E. C. (1995). The Semantic and Pragmatic Development of Substitutive Complex Prepositions in English. In A. H. Jucker (éd.), *Historical Pragmatics: pragmatic developments in the history of English* (pp. 243-273). Amsterdam: John Benjamins Publishing Company.

Traugott, E. C. et Trousdale, G. (2013). *Constructionalization and Constructional Changes*. Oxford: Oxford University Press.