Corpus arborés instantanés : des outils pour collecter et annoter des données textuelles

Kim Gerdes, Mcf, Sorbonne Nouvelle - Paris 3, Institut de linguistique et phonétique générale et appliquées (ILPGA)

Intervention au séminaire du Laboratoire CLILLAC-ARP du lundi 16 octobre 2017

La datamasse textuelle n'a pas encore transformé la linguistique autant que d'autres sciences. Outre des raisons épistémologiques, il y a des raisons tout à fait concrètes qui rend la collecte d'un site web entier difficile pour un linguiste : Configurer un crawler (collecteur), encoder correctement le résultat, extraire les parties intéressantes de milliers de pages html, uniformiser les textes et finalement interpréter statistiquement les données brutes est hors de portée du linguiste commun – sans parler de l'annotation syntaxique des données – ce qui l'oblige souvent de se contenter de travailler sur des corpus de très petite taille ou bien sur des corpus déjà préparés et peu configurables à l'instar de Google n-gram.

Je présenterai un survol des fonctions qu'ouvrent deux logiciels que j'ai co-devéloppés : le Gromoteur et l'Arborateur (http://gromoteur.ilpga.fr/ et https://arborator.ilpga.fr/).

Le premier logiciel, le Gromoteur, est un collecteur de données web configurable dans une interface graphique avec quelques spécificités nécessaires pour la recherche en linguistique, par exemple la possibilité de ne collecter que des pages d'une langue particulière, reconnue automatiquement, permettant ainsi la collecte de textes dans des langues peu représentées. Le Gromoteur contient aussi des modules d'extraction de parties de pages web et d'étiquetage morpho-syntaxique pour une série de langues. L'analyse statistique intégrée propose le calcul et l'affichage graphique de la sur- et sous-représentation des types par section de texte ainsi qu'un calcul de collocations dans une fenêtre de mots donnée.

L'arborateur permet « d'arborer » des textes brutes avec des graphes en dépendance. Cela peut se faire complètement manuellement dans une interface graphique en ligne, où chaque annotateur ne voit que ses propres arbres – et le validateur peut ensuite combiner les différents arbres proposés. Ou bien, on peut faire intervenir des parseurs en dépendance soit en amont, si l'on dispose déjà d'un modèle pour la langue, soit dans un processus de bootstrapping, où l'interface en ligne permet d'entraîner un parseur (Mate, Bohnet 2013) dès qu'on a terminé l'annotation dépendentielle d'une nouvelle série de phrases. On peut ainsi considérablement augmenter la vitesse et la qualité des annotations. Le logiciel s'intègre aussi très bien dans l'enseignement de la grammaire – dans un but purement pédagogique ou bien dans un projet de « class-sourcing » de l'annotation (Gerdes 2013).

Finalement, je montre à l'aide d'un treebank parallèle de l'oral Cantonais-Mandarin (développé en collaboration avec la City University de Hong Kong) comment ces corpus arborés ouvrent la voie à une syntaxe comparative quantifiée (Wong et al. 2017).