L'analyse linguistique de l'interlangue en anglais : l'intérêt des ré-annotations automatiques de corpus

Intervention au séminaire du Laboratoire CLILLAC-ARP du lundi 13 février 2017

Thomas Gaillat
PhD
Université Paris Diderot - Paris 7
CLILLAC-ARP

L'Interlangue est un objet d'analyse privilégié des problématiques d'apprentissage d'une langue seconde. Son étude permet d'extraire des usages idiosyncratiques propres aux apprenants. En ce sens, les erreurs et fréquences d'usages sont des indicateurs nécessaires mais non suffisants permettant de mettre en lumière la complexité de la langue d'apprenants. Les corpus d'apprenants jouent ici un rôle essentiel en permettant d'avoir accès à cette interlangue sous de nombreux aspects (Granger 2008; Granger, Gilquin & Meunier 2015).

Néanmoins ces outils atteignent leurs limites lorsqu'il s'agit de les faire dialoguer entre eux et ainsi permettre des comparaisons entre locuteurs de L1 différentes y compris native. L'automatisation de l'analyse linguistique de l'interlangue peut contribuer à lever cette limite (Díaz-Negrillo, Ballier & Thompson 2013). Celle ci-peut se décliner en trois volets. Premièrement, l'introduction automatique d'annotations linguistiques fines permet l'enrichissement à moindre coût de textes non structurés composant les corpus. Deuxièmement, la conversion des textes annotés en structures de données formelles facilite l'échange de données entre corpus par le biais de requêtes complexes sur les co-occurrences, adjacentes ou non, de traits linguistiques. Finalement, la conversion des corpus en jeux de données rend les données exploitables par diverses techniques d'analyse automatique telles que la régression (Gries 2013; Baayen 2008), le data mining (Hahsler 2006) ou encore l'apprentissage machine (Mitchell 1997; Flach 2012).

Au final, les données provenant de divers corpus sont mises en regard automatiquement. On peut alors conduire des analyses visant l'extraction de traits liés à des formes prises en contexte et observer les tendances d'usage propres à chaque L1 (Boyd, Gegg-Harrison & Byron 2005; Anders Noklestad 2009; Tono 2014). L'automatisation assiste le linguiste dans le processus d'analyse de la complexité de la langue et ouvre la voie vers une grammaire des probabilités.

Références :

Anders Noklestad. 2009. A Machine Learning Approach to Anaphora Resolution Including Named Entity Recognition, PP Attachment Disambiguation, and Animacy Detection. The University of Oslo.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: a Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Boyd, Adriane, Whitney Gegg-Harrison & Donna Byron. 2005. Identifying non-referential it: a machine learning approach incorporating linguistically motivated patterns. *Proceedings of the*

Second ACL Workshop on Effective Tools and Methodologies for Teaching NLP and CL, 40–47. University of Michigan - Ann Arbor, Michigan, USA: Association for Computational Linguistics.

Díaz-Negrillo, Ana, Nicolas Ballier & Paul Thompson (eds.). 2013. *Automatic treatment and analysis of learner corpus data*. (Studies in Corpus Linguistics 59). Amsterdam, Pays-Bas, Etats-Unis: John Benjamins Publishing Co.

Flach, Peter A. 2012. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge, United Kingdom: Cambridge University Press.

Granger, Sylviane. 2008. Learner corpora in foreign language education. In Nancy H. Hornberger (ed.), *Encyclopedia of Language and Education*, vol. 4, 337–351. Boston: Springer.

Granger, Sylviane, Gaëtanelle Gilquin & Fanny Meunier (eds.). 2015. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Gries, Stefan Th. 2013. *Statistics for linguistics with R: a practical introduction*. 2nd revised. Berlin, New York: De Gruyter Mouton.

Hahsler, Michael. 2006. A Model-Based Frequency Constraint for Mining Associations from Transaction Data. *Data Mining and Knowledge Discovery* 13(2). 137–166.

Mitchell, Tom M. 1997. *Machine Learning*. The McGraw-Hill Companies. Singapore. Tono, Yukio. 2014. Interlanguage Annotations and Association Rule Mining in the Acquisition of

Relative Clause Constructions. Poznan.