Bi-text: A systemic functional approach and textometic analysis

Maria ZIMINA

Université Paris Diderot, CLILLAC-ARP – mzimina@eila.univ-paris-diderot.fr

Keywords: Bilingual corpus annotation, comparable corpora, lexicogrammar, textometric analysis, translation

Multilingual text corpora with linguistic annotations are becoming increasingly important in translation studies (Tiedemann, 2011). Many software systems may help with systemic functional analyses of these corpora. However, they often have different capabilities and do not support free interchange. In Systemic Functional Linguistics (SFL), an integrated approach is essential to simultaneously explore multiple corpus layers and their interactions statistics.

The fndings of textmetric analysis helped to develop a research framework for multi-layer linguistic corpora with complex annotations (Fleury, 2013). Called *Le Trameur* ("threader" in English), this framework is built upon an XML-based data model. Available to any corpus linguist (http://www.tal.univ-paris3.fr/trameur/), it allows managing all stages of corpus exploration, including corpus maps and statistical analysis of dependency relations, within a single graphical user interface. In this research, the framework implemented in *Le Trameur* is used for mapping translation correspondences in a comparable corpus BBC_Lenta.RU (Klementiev & Roth, 2006). This corpus is composed of BBC News (2001-2005: 1 million words) and their adaptations into Russian published by www.lenta.ru (approximately 500,000 words). The process of adaptation through translation brings important linguistic and cultural shifts that make these English-Russian texts quite challenging for automatic alignment.

Following a lexicogrammar approach (Gledhill, 2011), this research explores characteristic attractions of signifcantly overrepresented linguistic patterns in corresponding text zones (corpus parts) to reveal translation correspondences. The results show that translation mapping is achieved when automatically discovered lexical correspondences are used as anchor points to explore functional equivalence of related linguistic features. In textometric studies, this process based on contrastive analysis of selected text zones is known as

resonance. It relies upon characteristic elements computation (Lebart *et al*., 1998) and can be propagated across multiple annotation layers.

The research fndings suggest that combing systemic functional approach and textometic analysis offers new perspectives for context-based comparable text processing.

References

Banks D. (2005). *Introduction à la linguistique systémique fonctionnelle.* Editions L'Harmattan.

Fleury S. (2013). Le Trameur. Propositions de description et d'implémentation des objets textométriques. *Publication sur le site de l'Université Paris 3* : http://www.tal.univ-paris3.fr/trameur/trameur-propositions-definitions-objets-textometriques.pdf

Fleury S. and Zimina M. (2008). Utilisation de MkAlign pour la traduction philologique. *Actes des 9es Journées internationales d'Analyse statistique des Données Textuelles (JADT'08)*, pp. 483-493.

Gledhill C. (2011). The lexicogrammar approach to analysing phraseology and collocation in ESP texts. *ASP*, 59: 5-23.

Klementiev A. and Roth D. (2006). Weakly Supervised Named Entity Transliteration and Discovery from Multilingual Comparable Corpora. *ACL-44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 817-824 [corpus : http://cogcomp.cs.illinois.edu/page/resource_view/1].

Salem A. (2004). Introduction à la résonance textuelle. *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles (JADT'04)*, pp. 986-992.

Tiedemann J. (2011). *Bitext Alignment*. Morgan and Claypool Publishers (USA).