A corpus-based contrastive analysis of English and French lexical bundles in five genres

Natalia Grabar (STL UMR8163 CNRS, Université de Lille 3) & *Marie-Aude Lefer* (Marie Haps School for Translators and Interpreters, Brussels)

This presentation reports on a large-scale analysis of English and French lexical bundles (also called recurrent word combinations) in five genres. Phraseological units, especially lexical bundles, have been largely under-researched in corpus-based contrastive studies so far (cf. Ebeling & Ebeling 2013). However, bundles are definitely worth investigating, as they can help uncover metadiscursive and rhetorical cross-linguistic contrasts that could not be revealed otherwise.

Looking at parliamentary debates and newspaper editorials, Granger (2014) found that lexical bundles are used more pervasively in French argumentative discourse than in English. This can be attributed to a systemic difference between the two languages: French has often been said to rely heavily on rhetorical markers, to be explicitly emphatic and, more generally, to be more verbose than English (see e.g. Vinay & Darbelnet 1995). Drawing on insights from recent cross-register contrastive studies (Hansen-Schirra et al. 2012, Neumann 2013, Lefer & Vogeleer 2014), we wish to examine the genre-sensitivity of lexical bundles in English and French across five genres. One of our starting-point hypotheses is that, in view of the pervasiveness of bundles in French, genres should share more bundles in French than in English.

The contrastive analysis relies on comparable data extracted from four corpora, representing five genres: Europarl (transcripts of parliamentary debates; Koehn 2005, Cartoni & Meyer 2012), KIAP (research articles in medicine, economics and linguistics; Fløttum et al. 2006), Multed (editorials) and PLECI (fiction and news). Together they total 25+ million tokens. The case study presented here is based on trigrams (3-word lexical bundles) only. All trigrams were automatically extracted, of which the most frequent 1.5% trigram types in each corpus were kept for further analysis (this corresponds to 63,150 and 78,756 trigrams in French and English, respectively). These were then classified as being genre-specific (i.e. attested in one genre only) or shared by genres (i.e. attested in 2, 3, 4 or all genres), representing a continuum from high genre-sensitivity (or specificity) to high genre-insensitivity.

Contrary to our initial expectations, preliminary quantitative results suggest that English and French are strikingly similar as regards the distribution of bundles across genres: ca. 75% of the top-1.5% trigrams are genre-specific in each language (English examples include *however I believe*$_{\text{Europarl}}$, *his voice was*$_{\text{PLECIfiction}}$, *empirical analysis of*$_{\text{KIAP}}$), while a mere 0.5%-1% is genre-insensitive, i.e. shared by all five genres (e.g. *some sort of*, *in the process*, *than that of*, *on the contrary*, *no reason to*). In our presentation we will examine the cross-genre similarities and differences in the two languages, so as to uncover some of the typical phraseological features of the genres and language systems under investigation. We will also try to characterize the genre-specific and genre-insensitive trigrams, relying, among other things, on the distinction between referential bundles, discourse organizers and stance markers (Biber *et al*. 2004). The presentation will end with some of the implications of our study for cross-linguistic phraseological research.

References

Biber, D., Conrad, S. & Cortes, V. (2004). *If you look at* … Lexical Bundles in University Lectures and Textbooks. *Applied Linguistics* 25, 371-405.

Cartoni, B. & Meyer, T. (2012). Extracting directional and comparable corpora from a multilingual corpus for translation studies. In: *8th International Conference on Language Resources and Evaluation* (LREC).

Ebeling, J. & Oksefjell Ebeling, S. (2013). *Patterns in Contrast*. Amsterdam & Philadelphia: John Benjamins.

Fløttum, K., Dahl, T. & Kinn, T. (2006). *Academic Voices – across languages and disciplines*. Amsterdam & Philadelphia: John Benjamins.

Granger, S. (2014). A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast* 14(1), 58-72.

Hansen-Schirra, S., Neumann, S. & Steiner, E. (2012). *Cross-linguistic Corpora for the Study of Translations - Insights from the Language Pair English-German*. Berlin: de Gruyter Mouton.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In: *MT Summit X*, 79-86.

Lefer, M.-A. & Vogeleer, S. (eds). (2014). *Genre- and register-related discourse features in contrast*. Special issue of *Languages in Contrast*, 14(1).

Neumann, S. (2013). *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German*. Berlin: de Gruyter Mouton.

Vinay, J.-P. & Darbelnet, J. (1995). [1958]. *Comparative Stylistics of French and English. A Methodology for Translation*. Amsterdam & Philadelphia: John Benjamins.