Identifying usage patterns in corpus data:
A multifactorial account of epistemic and evidential markers in English and French

Introduction: This is a contrastive corpus-based analysis of a set of epistemic and evidential markers in British English and metropolitan French. The markers under investigation are *I think*, *I believe, it seems, it appears* and *je pense, je crois, il semble* and *il paraît*. The goal here is to determine the usage patterns for the markers across both languages through a combination of detailed manual analysis and multivariate statistical modeling. Patterns, in this context, are understood as statistical tendencies.

Method: The approach adopted here is known as the multifactorial feature analysis (Geeraerts *et al.* 1994; Gries & Stefanowitsch 2006; Glynn & Fischer 2010; Glynn & Robinson 2014). The underlying assumption behind this method is that contextualized language use is an index of language structure and, by extension, of conceptual structure. In other words, by way of generalization across many usage-events (i.e., observations), it is possible to reveal frequency-based patterns of language use. These patterns are identified in terms of associations and disassociations of contextual clues, also referred to as usage-features.

Data: The data for the study, amounting to 120 occurrences per marker, were extracted from the blog components of the *TenTen* corpora (SketchEngine Kilgarriff *et al.* 2014) with the use of regular expressions. All the examples will be manually annotated for a range of syntactic and semantico-pragmatic variables (e.g., Marker position: <Initial>, <Medial>, <Final>; Temporal scope: <Present>, <Past>, <Future>, <General Present>; Information verifiability: <Verifiable>, <Non-verifiable>; or Reality: <Personal> vs. <Public>). The resulting metadata will then be submitted to multivariate statistical modeling (correspondence analysis and multinomial logistic regression) in order to identify the frequency-based patterns associated with the investigated markers of epistemicity and evidentiality.

Expected results: We expect that the agentive markers *I think* and *I believe* and their respective equivalents in French will be more frequently associated in use with situations referring to personal and non-verifiable reality. Social reality and verifiable information, in turn, are here expected to be more readily correlated with the non-agentive markers in the two languages.

References

Geeraerts, D., S. Grondelaers & P. Bakema. 1994. *The Structure of Lexical Variation.* Berlin: Mouton de Gruyter.
Glynn, D. & J. Robinson (Eds.). 2014. *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy*. Amsterdam: John Benjamins.
Glynn, D. & K. Fischer (Eds.). 2010. *Quantitative Methods in Cognitive Semantics: Corpus-driven approaches*. Berlin: Mouton de Gruyter.
Gries, St. Th., & Stefanowitsch, A. (eds.) 2006. *Corpora in Cognitive Linguistics: Corpus-based approaches to syntax and lexis*. Berlin: Mouton de Gruyter.
Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý & V. Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* ASIALEX 1: 7–36.