From the Cookbook of Corpus-Based Lexical Lectometry: A Taste of Chinese

Kris Heylen, Weiwei Zhang
University of Leuven
kris.heylen@kuleuven.be; weiwei.zhang@kuleuven.be

Lectometric approaches measure distances between language varieties (dialects, sociolects, registers etc.) by aggregating over observed differences in the realizations of a set of linguistic variables. In *lexical* lectometry, a variable consists of the alternative lexical expressions for one concept. In *corpus-based* lectometry, the observed realizations are culled from stratified corpora. Measuring semantically defined variables in corpora, and aggregating over them, poses specific methodological challenges that have been tackled in a number of studies (Heylen & Ruette 2013; Ruette et al. 2014; Ruette, Ehret & Szmrecsanyi 2016) with different statistical techniques, including Distributional Semantic Models. Yet so far, no general framework for corpus-based lexical lectometry has been formulated that systematically describes the issues and options in each step of the procedure so that it can be straightforwardly applied to new data and new languages, other than English (Ruette, Ehret & Szmrecsanyi 2016), Dutch (Geeraerts, Grondelaers & Speelman 1999) and Portuguese (Soares da Silva 2010).

This paper can be characterized as a twofold extension of the previous studies. First, it aims to establish a general framework for lexical lectometry research that considers most if not all options for different steps. Second, we want to go beyond the Indo-European languages by extending the framework on a typologically unrelated language, i.e. Chinese varieties.

For the general framework, we propose that a proper lexical lectometry research normally should involve the following steps: (1) compilation of a lectally stratified corpus; (2) sampling concepts as measuring points for lectometry; (3) identification of lexical expressions per concept; (4) disambiguation of lexical expressions in corpus data; (5) calculation of aggregated lexico-lectometric distances; (6) evaluation of measurement reliability and validity. For each step, we further provide possible options and caveats. For instance, step 2 and 3 can rely on existing concept-based lexical databases, like a synonym dictionary, or use corpus-driven keyword extraction and semantic vector space models. Step 4 can either make use of token-level distributional semantics models or rely on simpler n-gram language models.

To assess the portability of the general framework, both in practical and linguistic-typological terms, we perform a lexical lectometric analysis for varieties of Chinese based on data from large-scale corpora of Mainland Chinese, Taiwan Chinese and Singapore Chinese.

References

Geeraerts, D., Grondelaers, S. & D. Speelman. (1999). *Convergentie en Divergentie in de Nederlandse Woordenschat*. Amsterdam: Meertens Instituut.

Heylen, K., & Ruette, T. (2013). Degrees of semantic control in measuring aggregated lexical distances. In Borin, L. & Saxena, A. (eds.), *Approaches to Measuring Linguistic Differences*: 361-382. Berlin: Mouton de Gruyter.

Ruette, T., Geeraerts, D., Peirsman, Y., & Speelman, D. (2014). Semantic weighting mechanisms in scalable lexical sociolectometry. In Szmrecsanyi, B. & Wälchli, B. (eds.), *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*: 205-230. Berlin: Mouton de Gruyter.

Ruette, T., Ehret, K., & Szmrecsanyi, B. (2016). A lectometric analysis of aggregated lexical variation in written Standard English with Semantic Vector Space models. *International Journal of Corpus Linguistics*, 21(1), 48-79.

Soares da Silva, A. (2010). Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese. In Geeraerts, D., et al. (eds.), *Advances in Cognitive Sociolinguistics*: 41-84. Berlin: Mouton de Gruyter.