Analyzing lexical variation in regional varieties of Chinese: A concept-based approach

Weiwei Zhang, Kris Heylen, Dirk Geeraerts

University of Leuven

weiwei.zhang@kuleuven.be; kris.heylen@kuleuven.be; dirk.geeraerts@kuleuven.be

Keywords: lexical variation, concept, Cognitive Sociolinguistics, Chinese, basic level

Research question – In the framework of Cognitive Sociolinguistics, this study empirically tests to what extent there is lectally structured onomasiological variation in the Chinese lexicon and specifically, to what extent such lexical variation is influenced by features of the concepts.

Background – With decontextualized evidence from dialect dictionaries and psychometric data, previous studies in Cognitive Sociolinguistics have shown the impact of concept characteristics on lexical diversity in language varieties (e.g. Speelman & Geeraerts 2008; Franco et al. 2019). The present study intends to provide additional quantitative corpus-based evidence for the hypothesis that lexical variation is constrained by lectal features and also by features of the concepts to be named.

Material – The data for the present study is taken from the 'Tagged Chinese Gigaword Corpus' (Huang 2009), which is a newswire corpus containing texts from three regional varieties of Chinese, i.e. Mainland, Taiwan and Singapore Chinese.

Method – To measure **lexical lectometric distance**, we rely on the onomasiological measure for lexical uniformity designed by Geeraerts et al. (1999), which has been successfully applied in previous lexical lectometry studies (e.g. Ruette et al. 2016). The uniformity measurements allow to quantify the amount of lexical variation between varieties both per concept and on an aggregate level. Next, we zoom in on concepts with low naming uniformity across varieties and investigate four concept characteristics that may explain higher degrees of lexical variation: (1) **conceptual entrenchment**. We use concept frequency in the corpus as an approximation and interpret the correlation against the background of the potentially opposing forces idiosyncratic naming preferences and expressive diversification. (2) **lexical fields**: Based on Chinese WordNet, we set a high-level cutoff in the WordNet taxonomy and use the hyponym label as an approximation for lexical field; (3) **vagueness**. We use token-level Word Space Models to model semantic vagueness of lexical items (cf. Heylen et al. 2015); (4) **affect**, which is operationalized as a sentiment score per concept that aggregates over the scores assigned to each occurrence that concept by a sentiment analyzer. Finally, we build a linear **regression model** to assess the effect of the four explanatory variables (concept characteristics) on the response (lexical lectometric distance).

Relevance – The relevance of the study is double. First, the findings of the study contribute to the current lectometric research by testing the uniformity measurement's viability on a language that is typologically not related to the languages scholars have looked at so far, such as English (Ruette et al. 2016), Dutch (Geeraerts et al. 1999) and Portuguese (Soares da Silva 2010). Second, theoretically, the paper adds to the growing recognition of concept characteristics as a determinant of (degree of) variation (Speelman & Geeraerts 2008; Franco et al. 2019). This idea has been part of Cognitive Linguistics since the first formulation of the basic-level hypothesis but is now being studied on a broader and more systematic semantic basis.

References

Franco, K. et al. (2019). Concept characteristics and variation in lexical diversity in two Dutch dialect areas. *Cognitive Linguistics,* 30(1): 205–242.

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

Heylen, K., Wielfaert, T., Speelman, D., & Geeraerts, D. (2015). Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua*, *157*, 153-172.

Huang, C. (2009). *Tagged Chinese Gigaword Version 2.0*. Philadelphia: Linguistic Data Consortium.

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.

Speelman, D., & Geeraerts, D. (2008). The role of concept characteristics in lexical dialectometry. *International Journal of Humanities and Arts Computing*, 2(1-2), 221-242.