Quantifying lexical semantic change across centuries: what can we still learn from Ancient Greek and Latin?

Barbara McGillivray[1,2]

[1] *University of Cambridge,* [2] *The Alan Turing Institute*

bm517@cam.ac.uk

The phenomenon of semantic change, and its relation with polysemy, is of great relevance to a range of humanistic disciplines. Scholars working on historical texts benefit from analyses of words' meaning in their historical context, as these allow them to search and find documents relevant to the topics of interest. A long and rich scholarly tradition has provided us with valuable insights into the mechanisms of semantic change for historical languages (cf. e.g. Leiwo 2012). On the other hand, semantic change can be successfully studied using probabilistic models, following the general paradigm put forward by Jenset and McGillivray (2017), and qualitative analyses can be leveraged to support and complement computational modelling. Recent research in Natural Language Processing has provided great advances in automatic methods for semantic change detection (cf. e.g. Tahmasebi et al. 2018), but its focus has been on modern languages and relatively recent corpus data.

Ancient Greek and Latin enjoy a fortunate status among the historical languages. The availability of large high-quality text corpora with basic annotation such as Diorisis for Ancient Greek (Vatri and McGillivray 2018) and LatinISE for Latin (McGillivray and Kilgarriff 2013), make them an ideal testing ground for analysing linguistic phenomena diachronically and at scale. On the other hand, the long diachronic span of the texts, coupled with uneven distributions and complex interactions, make these particularly challenging datasets, which cannot be adequately analysed within the scope of one single discipline.

In this talk I will report on two interdisciplinary projects exploring the challenges of quantitatively modelling semantic change and polysemy in Ancient Greek and Latin (McGillivray et al. 2019; Perrone et al. 2019). I will focus on some important methodological challenges, such as the presence of gaps and the lack of balance in the corpora, and how they can be addressed computationally. Finally, I will present some quantitative analyses based on a diachronic lexical semantic annotation of Ancient Greek and Latin corpora, highlighting semantic change and semantic variation effects.

References: Jenset, G. B. & B. McGillivray (2017). Quantitative Historical Linguistics. A corpus framework. Oxford University Press, Oxford. Leiwo, M. (2012). Introduction: Variation with Multiple Faces. In Leiwo, M., Halla-aho, H., and Vierros, M. (eds.). Variation and change in Greek and Latin, 1–11. Suomen Ateenan-instituutin säätiö, Helsinki. McGillivray, B. & A. Kilgarriff (2013). Tools for historical corpus research, and a corpus of Latin. In Paul Bennett, Martin Durrell, Silke Scheible, Richard J. Whitt (eds.), New Methods in Historical Corpus Linguistics, Tübingen: Narr. McGillivray, B., Hengchen, S., Lähteenoja, Palma, M. & A. Vatri (2019). A computational approach to lexical polysemy in Ancient Greek, Digital Scholarship in the Humanities: https://doi.org/10.1093/llc/fqz036. Perrone, V., Palma, M., Hengchen, S., Vatri, A., Smith, J. Q. & B. McGillivray (2019). GASC: Genre-Aware Semantic Change for Ancient Greek. In Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 56–66 Florence, Italy, August 2, 2019. Tahmasebi, N., Borin, L. & A. Jatowt (2018). Survey of computational approaches to diachronic conceptual change. arXiv preprint arXiv:1811.06278. Vatri, A. & B. McGillivray (2018). The Diorisis Ancient Greek Corpus. Research Data Journal for the Humanities and Social Sciences.