Evaluating Language Models for Ancient Greek: Design, Challenges, and Future Directions

Barbara McGillivray[1], Malvina Nissim[2], Nilo Pedrazzini[3], Saskia Peels-Matthey[2], Silvia Stopponi[2]
[1]*King's College London (UK)*
[2]*University of Groningen (NL)*
[3]*The Alan Turing Institute (UK)*

Linguistic research on corpora of ancient texts has recently seen a growing interest in the use and development of quantitative and computational methods (McGillivray, 2014). Techniques already widely applied to modern languages have proved useful to the study of ancient languages as well, with the training of distributional models of lexical semantics and historical language models as one of the most promising advances (see e.g. Rodda et al., 2019; Sprugnoli et al., 2020). In computational linguistic research, language modelling is the task of assigning probabilities to word distributions from a corpus, allowing individual word forms to be represented as numerical vectors. Such language models have been employed to tackle questions of great relevance to diachronic approaches to linguistics. Crucially, this includes the detection and tracking of lexical semantic change by comparing word representations in different time periods (see e.g. Kulkarni et al, 2015; Hamilton et al., 2016; Di Carlo et al., 2019). The viability of this approach for ancient Greek has been shown, among others, by Boschetti (2009), Rodda et al. (2017) and Perrone et al. (2021).

These new advances have been made possible by the availability of digitised collections of texts, tokenised and annotated at different linguistic levels. The Lemmatized Ancient Greek Texts (released by Giuseppe Celano) and the Diorisis Ancient Greek Corpus (Vatri & McGillivray, 2018), for example, include lemma and part-of-speech tags, together with morphological analysis, while two manually annotated treebanks of Ancient Greek, the PROIEL (Haug & Jøhndal, 2008) and the Ancient Greek Dependency Treebank (Bamman & Crane 2011) contain syntactic dependency information.

In our contribution we present a framework to evaluate the quality of computational lexical semantic representations of Ancient Greek words, obtained by training language models built with different techniques, including Word2Vec (Mikolov et al., 2013) and count-based models. The availability of manually annotated treebanks allows us to test the extent to which the integration of syntactic information (part of speech and dependency relations) may improve the quality of a model, leading to better, or simply different, semantic representations. Levy and Goldberg (2014) have shown for English that taking into account syntactic dependencies generates embeddings that are markedly different from simple bag-of-word ones, influencing the type of semantics captured by a model. Testing this on ancient languages opens new interesting avenues, since the automatic treatment of languages that cannot rely on native speakers may instead leverage syntax to model different aspects of the syntax-semantics interface. We will present the advantages of applying word embeddings to ancient languages, but we will also present the challenges of such methods and their limitations.

Being aware of the existing gap between the development of new methods in computational linguistics and the computational knowledge and skills actually available to most linguists outside the specific field, one of the aims of this contribution is to expose historical linguists to the computational methods discussed, and to explain and discuss their potentialities for the study of lexical semantics.

References

Bamman, D. and Crane, G. (2011) 'The Ancient Greek and Latin Dependency Treebanks' in *Language Technology for Cultural Heritage. Theory and Applications of Natural Language Processing*. Berlin, Heidelberg: Springer.

Boschetti, F. (2009) *A Corpus-based Approach to Philological Issues*. PhD thesis. Trento: Center for Mind / Brain Sciences, University of Trento.

Di Carlo, V., Bianchi, F. and Palmonari, M. (2019) 'Training Temporal Word Embeddings with a Compass' in *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (1), 6326-6334.

Levy, O., and Goldberg, Y. (2014) 'Dependency-based word embeddings' in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2)*, 302-308.

Hamilton, W. L., Leskovec, J. and Jurafsky, D. (2016), ‘Diachronic word embeddings reveal statistical laws of semantic change’ in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin: Association for Computational Linguistics, 1489–1501.

Haug, D. T. T., and Jøhndal, M. L. (2008) ‘Creating a parallel treebank of the Old Indo-European Bible translations’ in *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH)*, 27-34.

Kulkarni, V., Al-Rfou, R., Perozzi, B. and Skiena, S. (2015) ‘Statistically significant detection of linguistic change’ in *Proceedings of the 24th International World Wide Web Conference*, New York, Association for Computing Machinery, 625–635.

McGillivray, B. (2014), *Methods in Latin Computational Linguistics.* Leiden: Brill.

Perrone, V., Hengchen, S., Palma, M., Vatri, A., Smith, J. Q. and McGillivray, B. (2021) ‘Lexical semantic change for Ancient Greek and Latin’ in *Computational approaches to semantic change (Volume 6)*. Berlin: Language Science Press, 287-310.

Rodda, M. A., Senaldi, M. S. G. and Lenci, A. (2017) ‘*Panta rei*: Tracking semantic change with Distributional Semantics in ancient Greek’, *Italian Journal of Computational Linguistics* 3(1), 11-24.

Rodda, M. A., Probert, P. and McGillivray, B. (2019) ‘Vector space models of Ancient Greek word meaning, and a case study on Homer’, *TAL Traitement Automatique des Langues* 60 (3), 63-87.

Sprugnoli, R., Moretti, G. and Passarotti, M. (2020) ‘Building and Comparing Lemma Embeddings for Latin. Classical Latin versus Thomas Aquinas’, *IJCoL. Italian Journal of Computational Linguistics* 6(6-1), 29-45.

Vatri, A. and McGillivray, B. (2018) ‘The Diorisis Ancient Greek Corpus’, *Research Data Journal for the Humanities and Social Sciences* 3(1), 55-65.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) *Efficient estimation of word representations in vector space*. arXiv preprint.