An approach to language periodization based on word usage: looking for a pivot point in Latin

V. Generalova, D. Kondakova

Paper title: An approach to language periodization based on word usage: looking for a pivot point in Latin
Paper type: Short project introduction

Abstract:
Questions of periodization are relevant for studying the history of any language. Usually, stages of language evolution are determined on the basis of grammatical (mostly phonetic and morphological) changes. Transformations on the lexical level are often disregarded, despite being clear evidence for language development.

The present study suggests a method for language periodization based on lexical principles for Latin. Although the history of Latin is well-studied and a number of periodization schemes exist, these are still debated [see Adamik 2011 for an overview and critique]. We expect to get a more precise idea of the changes in the Latin language reflected in its lexical layer and determine period(s) during which a major shift can be observed. The corpus of written Latin is large enough to allow for an analysis of a representative selection of material. Since there has already been some work on Latin from the DH perspective, our contribution might be of use for further research in this field.

The method used for the present study is rooted in distributional semantics. The core idea of distributional semantics is that the words which have similar distributions (distributional synonyms) are supposed to have similar meanings. On the basis of this principle, we formulated a hypothesis concerning language change: if at a given point in time T2 a word W has a significantly different set of distributional synonyms than at an earlier point in time T1, a pivot point leading to this change in usage might have taken place between T2 and T1.

This pivot point is expected to be found in a proximity of a point suggested by other periodization methods. By splitting the corpus in two, we generate synonyms of each test word using each half of the corpus separately and then compare these lists. While choosing different years as delimiters, we can find one where the mean difference for all the test words is highest. This year is assumed to be the pivot point.

The distributional vector models were trained on a selected and lemmatized corpus of Latin prose with the word2vec's skip-gram algorithm (see more about the algorithms in [Mikolov 2013]). The sets of synonyms were also evaluated against the results for the model trained on the unsplit corpus. Statistical methods were applied to determine the significance of differences.

At the current stage, the initial hypothesis can be partly confirmed: the same words tend to have different distributional synonyms in different periods of language development. The research is still ongoing and more results are to be obtained.

References

Adamik, B. "The periodization of Latin: an old question revisited." Latin Linguistics in the Early 21st Century: Acts of the 16th International Colloquium on Latin Linguistics, Uppsala. 2011.
Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.