Criteria for the Construction of Historical Corpora

Stefanie Eckmann
Ludwig-Maximilians-Universität München
stefanie-eckmann@gmx.de

Tag
Datum
Zeit
Raum

Recently, computational linguistics has developed an increasing interest in historical linguistics (Gulordava and Baroni 2011; Hamilton et al. 2016; Frermann and Lapata 2016; Schlechtweg et al. 2017). Traditional historical linguistics mostly works with a restricted amount of data, while computational linguistic methods can be applied to a large amount of data. There is, however, still the need for a balanced benchmark corpus (Frerman and Lapata 2016, p. 33). The data used for corpus-based and historical computational linguistic research needs to fulfill certain criteria.

In this study, a corpus was built to investigate types of semantic change in diachrony. The construction of the corpus is based on the hypothesis that text genre strongly influences the semantics of individual items. Therefore, there needs to be an even distribution of texts and tokens for each text genre over the time period under investigation. The corpus must be able to handle the following problems: (i), variation in orthography; (ii), regional variation; (iii), distribution of texts and tokens over time; (iv), distribution of texts and tokens for each text genre over time (Eckmann 2017).

The presentation will give an example of how a corpus can be constructed that fulfills the aformentioned criteria. Using German as sample language and the DTA (Deutsches Textarchiv)[1] as database, a corpus with ~10 Mio tokens was constructed to investigate several types of semantic change. It covers the time period from 1600 to 1900, i.e., the development from late Early New High German to Modern New High German (Eckmann 2017).

References: • Eckmann, S. (2017): Towards a reassessment of semantic change in light of philological and corpus-based research. MA Thesis. LMU München. unpublished. • Frermann, L. and Lapata, M. (2016): A Bayesian Model of Diachronic Meaning Change. *TACL* 4, 31–45. • Gulordava, K. and Baroni, M. (2011): A Distributional Similarity Approach to the Detection of Semantic Change in the Google Books Ngram Corpus. *Proceedings of GEMS*, 67–71. • Hamilton, W., Leskovec, J., and Jurafsky, D. (2016): Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *Proceedings of ACL*, 1489–1501. • Schlechtweg, D., Eckmann, S., Santus, E., Schulte im Walde, S. and Hole, D. (2017): German in Flux: Detecting Metaphoric Change via Word Entropy. *Proceedings of CoNLL*.

[1] deutschestextarchiv.de