Detection of Fossilized, Archaic Words

Schlechtweg Dominik
Universität Stuttgart
dominik.schlechtweg@gmx.de

Stefanie Eckmann
LMU München
stefanie-eckmann@gmx.de

Tag
Datum
Zeit
Raum

According to Fritz (2006) obsolescence and loss of words has been investigated to a lesser extent than the occurrence of new words and uses (cf. p. 79). A particular type of loss of words is fossilization in phraseologisms where it may occur that words that are not in the lexicon of the language anymore still occur as part of a bigger lexical unit such as *Fug* in *mit Fug und Recht* (Jang, 2006, cf. p. 22). Fossilized structures are still integral parts of the synchronic grammar and/or lexicon. Similar examples are *gang und gäbe* or *geschweige denn* (Hackstein 2014).

We exploit the observation that archaic words in phraseologisms have a very narrow context, in order to detect them automatically in a large corpus. E.g., *Fug* can only occur in the context of *mit Fug und Recht*. We measure the degree of contextual narrowness of words by drawing from the notion of entropy and using methods from distributional semantics. In hypernym detection, word entropy reflects semantic generality (Santus et al., 2014). It has been successfully applied to detect metaphoric change (Schlechtweg et al., 2017). This measure is now applied to detect fossilized words. The hypothesis is that fossilization results in an extreme drop of word entropy over time and a significantly lower word entropy than other words in the same frequency area at a specific time point. For our investigation, we use the corpus of *Deutsches Textarchiv (erweitert)* (DTA), which is accessible online and downloadable for free.[1] The DTA provides more than 2447 lemmatized and POS-tagged texts (with more than 140M tokens) and covers a time period from the late 15th to the early 20th century.

References: • Fritz, G. (2006): *Historische Semantik*. Metzler. • Hackstein, O. (2014): Persistence phenomena in the evolution of constructions. In: *Evolution of Syntactic Relations*, 89–94. • Jang, A.-Y. (2006): *Lexikalische Archaismen und ihre Verwendung in Pressetexten des heutigen Deutsch*. PhD thesis. • Santus, E., Lenci, A., Lu, Q. and Schulte im Walde, S. (2014): Chasing hypernyms in vector spaces with entropy. *Proceedings of EACL*, 38–42. • Schlechtweg, D., Eckmann, S., Sanuts, E., Schulte im Walde, S. and Hole, D. (2017): German in Flux: Detecting metaphoric change via word entropy. *Proceedings of CoNNL*, 354-367.

[1] http://www.deutschestextarchiv.de/