Originally published in: The Twelfth International Corpus Linguistics Conference 2023. Lancaster University, Monday 3rd - Thursday 6th July, 2023.
Book of abstracts. – Lancaster: Lancaster University, 2023. P. 155.

[17] *A Distributional Comparison between FOLK and DeReKo.*

Marc Kupietz (Leibniz-Institut für Deutsche Sprache), Peter Fankhauser (Leibniz-Institut für Deutsche Sprache) and Josef Ruppenhofer (Leibniz-Institut für Deutsche Sprache).

Abstract. The goal of this study is to compare language use in spoken vs. written contemporary German. To this end, we contrast a orthographically normalized variant of version 2.18 of the Forschungs- und Lehrkorpus Gesprochenes Deutsch FOLK with 3.4M tokens (Schmidt 2017) with the DeReKo-2017-II edition German Reference Corpus with 33.0G tokens (Kupietz et al. 2018). We train a structured skipgram word2vec model (Ling et al. 2015) for DeReKo and retrain it on FOLK to arrive at comparable word embeddings for spoken language. On this basis we compare relative frequencies of words in DeReKo and FOLK, and analyse word embeddings for maximum displacement between their written and spoken variants together with their paradigmatic and syntagmatic neighbourhood. First results indicate that the top words w.r.t. difference in relative frequency (including particles and personal pronouns) are also maximally displaced and accordingly change their syntagmatic neighbourhood. However, their paradigmatic neighbourhood remains fairly stable, i.e., displacement occurs for whole paradigmatic clusters rather than on individual words.

References

Kupietz, Marc/Lüngen, Harald/Kamocki, Paweł/Witt, Andreas (2018): The German ReferenceCorpus DeReKo: New Developments – New Opportunities. In: Proceedings of the 11th InternationalConference on Language Resources and Evaluation (LREC 2018). Miyazaki/Paris:European Language Resources Association (ELRA), pp. 4353-4360.

Ling, Wang / Dyer, C. / Black, A. / Trancoso, I. (2015): Two/too simple adaptations of word2vec for syntax problems. In Proc. of NAACL.

Schmidt, Thomas (2017): Construction and Dissemination of a Corpus of Spoken Interaction –Tools and Workflows in the FOLK project. Journal for Language Technology and Computational Linguistics (JLCL 31/1), pp. 127–154.

Keywords: distributional semantics, word embeddings, spoken vs. written