Nils Diewald/Marc Kupietz/Harald Lüngen

TOKENIZING ON SCALE

Preprocessing large text corpora on the lexical and sentence level

Abstract When comparing different tools in the field of natural language processing (NLP), the quality of their results usually has first priority. This is also true for tokenization. In the context of large and diverse corpora for linguistic research purposes, however, other criteria also play a role – not least sufficient speed to process the data in an acceptable amount of time. In this paper we evaluate several state-of-the-art tokenization tools for German – including our own – with regard to theses criteria. We conclude that while not all tools are applicable in this setting, no compromises regarding quality need to be made.

Keywords Corpora; tokenization; software

Contact information

Nils Diewald
Leibniz-Institut für Deutsche Sprache
diewald@ids-mannheim.de

Marc Kupietz
Leibniz-Institut für Deutsche Sprache
kupietz@ids-mannheim.de

Harald Lüngen
Leibniz-Institut für Deutsche Sprache
luengen@ids-mannheim.de

This page is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
DOI: *https://doi.org/10.14618/t86d-e088*