Digital Lyrics: Multidisciplinary Research on German-language Pop Culture

Songtexte digital: Multidisziplinäre Erforschung deutschsprachiger Popkultur

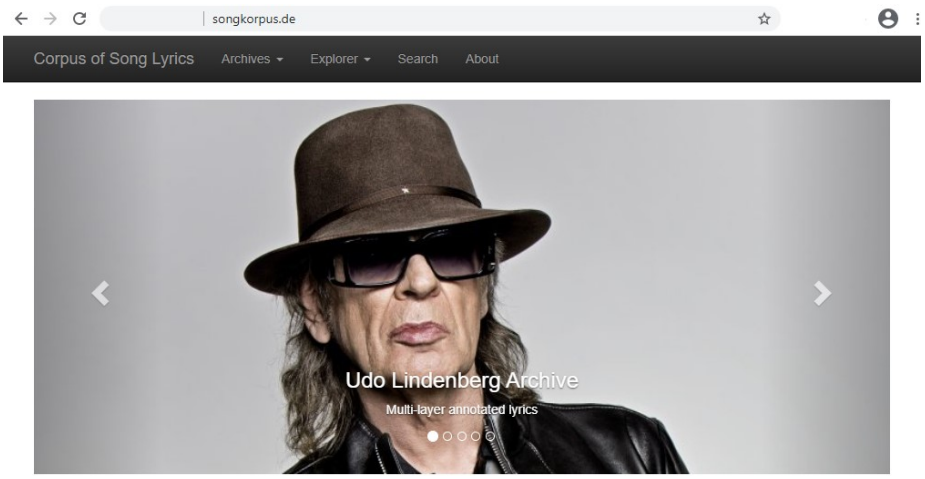
Roman Schneider, Leibniz-Institut für Deutsche Sprache (IDS), Mannheim

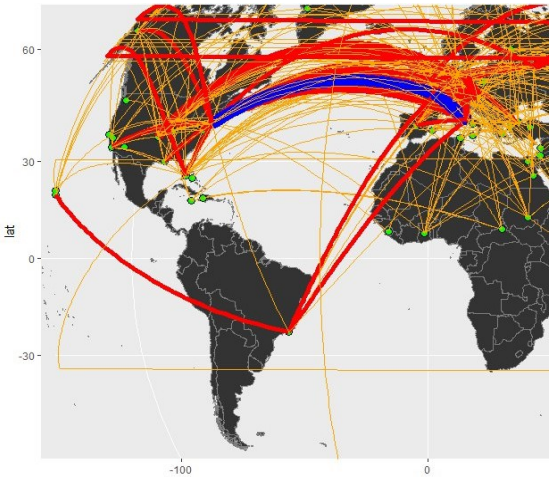
The Corpus Data

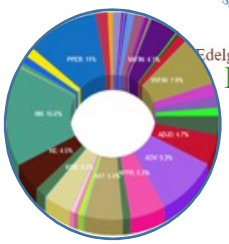
The Song Corpus provides multiply annotated lyrics as a sustainable research basis for linguistics and the broad spectrum of cultural sciences. The resource contains three types of data:

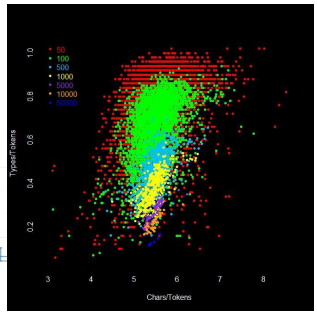
- TEI P5-compliant texts as primary data
- linguistically motivated annotations
- extralinguistic metadata

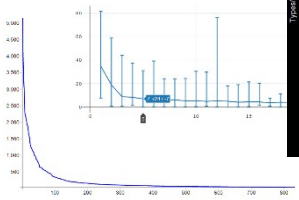
It promotes empirically grounded analyses of genre-specific features, systemic-structural correlations and tendencies in contemporary pop music texts. The corpus has been stratified into thematic and author-specific archives.







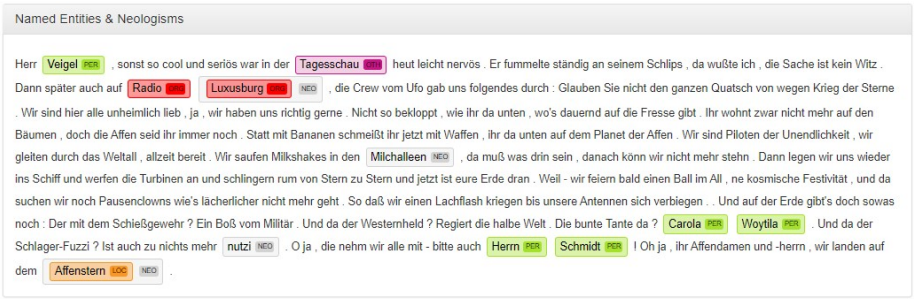




Annotation & Curation

Segmentation has to deal with the challenging fact that lyrics primarily function acoustically. The written form often do not contain punctuation marks, or at least do not use them consistently for the identification of phrases or sentences. Overall, lyrics show a conscious **play on norms on a variety of linguistic levels** (syntactic structure, spelling, semantics, part of speech, word formation, etc.).

In order to assure consistent description quality, the corpus processing takes place as an interplay between automated annotation runs and manual post-editing. The **part of speech** annotation layer broadens the STTS 2.0 tagset for new contractions. Extensible layers for **named entities**, **neologisms**, and **rhyming forms** are added. All annotations are subject to **inter-annotator agreement**.



Exploration

A dedicated website (***songkorpus.de***) offers combined search by token, lemma, and POS, as well as the exploration of statistics and live visualizations. Online queries include frequency analyses on character, word, verse, song, and corpus level. The corpus thus fills a data gap in the continuum between both standard and nonstandard, written and spoken language, that previously prevented empirical answers to syntactic, semantic or pragmatic questions.

Quantitative Phenomena

The corpus allows to test empirical laws and measures systematically across subgenres and time periods, e.g. rank-frequency distributions, correlations between the size of a linguistic construct and its constituents, or type-token ratios.