

Digital Lyrics: Multidisciplinary Research on German-language Pop Culture

Songtexte digital: Multidisziplinäre Erforschung deutschsprachiger Popkultur

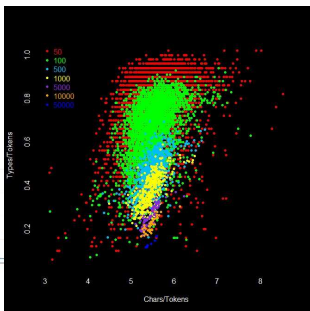
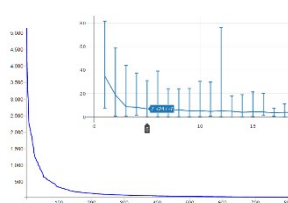
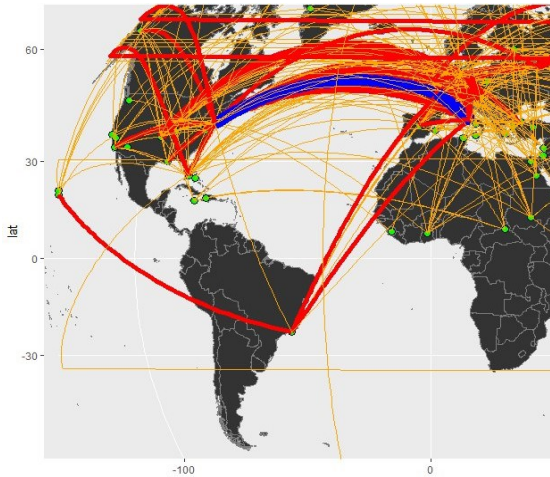
Roman Schneider, Leibniz-Institut für Deutsche Sprache (IDS), Mannheim

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:

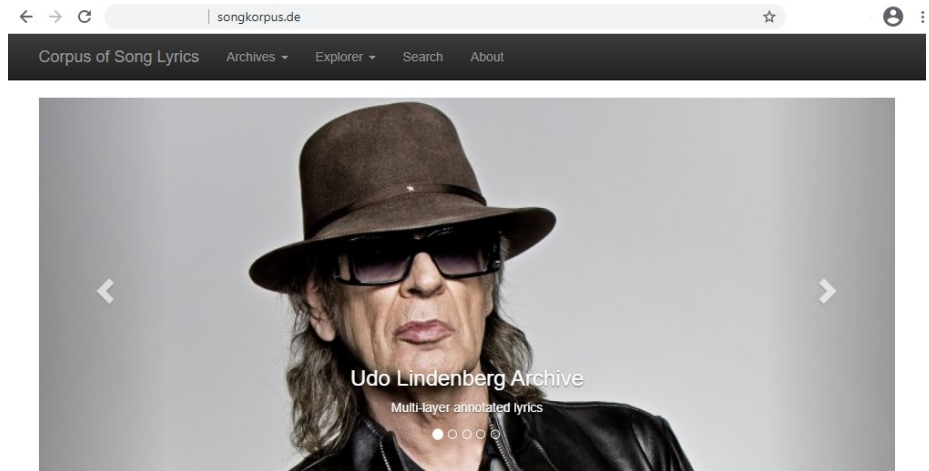
- TEI P5-compliant texts as primary data
- linguistically motivated annotations
- extralinguistic metadata

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.



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.



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**.

Named Entities & Neologisms

Herr **Veigel** **PSR**, sonst so cool und seriös war in der **Tagesschau** **NEO** heute leicht nervös. Er fummelte ständig an seinem Schlips, da wußte ich, die Sache ist kein Witz. Dann später auch auf **Radio** **NEO**, **Luxemburg** **NEO**, die Crew vom Ufo gab uns folgendes durch: Glauben Sie nicht den ganzen Quatsch von wegen Krieg der Sterne. Wir sind hier alle unheimlich lieb, ja, wir haben uns richtig gerne. Nicht so bekloppt, wie ihr da unten, wo's dauernd auf die Fresse gibt. Ihr wohnt zwar nicht mehr auf den Bäumen, doch die Affen seid ihr immer noch. Statt mit Bananen schmelßt ihr jetzt mit Waffen, ihr da unten auf dem Planet der Affen. Wir sind Piloten der Unendlichkeit, wir gleiten durch das Weltall, allzeit bereit. Wir saufen Milkhakes in den **Milchalleen** **NEO**, da muß was drin sein, danach könn wir nicht mehr stehn. Dann legen wir uns wieder ins Schiff und werfen die Turbinen an und schlingern rum von Stern zu Stern und jetzt ist eure Erde dran. Weil - wir feiern bald einen Ball im All, ne kosmische Festivität, und da suchen wir noch Pausencloowns wie's lächerlicher nicht mehr geht. So daß wir einen Lachflash kriegen bis unsere Antennen sich verbiegen. Und auf der Erde gib's doch sowas noch: Der mit dem Schießgewehr? Ein Boß vom Militär. Und da der Westermheld? Regiert die halbe Welt. Die bunte Tante da? **Carola** **PSR**, **Woytila** **PSR**. Und da der Schlager-Fuzzi? Ist auch zu nichts mehr **nutzi** **NEO**. O ja, die nehme wir alle mit - bitte auch **Herrn** **PSR**, **Schmidt** **PSR**. Oh ja, ihr Affendamen und -herren, wir landen auf dem **Affenstern** **NEO**.

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.

