Empirical Research Between Standard and Non-Standard: The German Song Corpus

Donnerstag,
05.03.2020
10:30–11:15
ESA1 W Foyer

Keoma G. Kögler, Roman Schneider
Justus-Liebig-Universität Gießen
keoma.g.koegler@germanistik.uni-giessen.de , roman.schneider@germanistik.uni-giessen.de

CLP

We introduce a corpus with multi-layer annotation of German pop lyrics as a basis for interdisciplinary research. Lyrics can be considered a text genre with features of both written and spoken discourse, as well as a language variety in the continuum between standard and non-standard. Given the fact that pop music has developed from an originally youth cultural phenomenon into an integral part of modern culture, its textual content also has a high "communicative impact factor" (Kreyer/Mukherjee 2007), and apparently shows a remarkable amount of variation and creativity. Nevertheless, empirically based research on German pop lyrics remains an interdisciplinary desideratum so far – possibly due to the non-existence of reasonably stratified and preprocessed data resources and obstacles concerning copyright.

The publicly available Song Corpus (Schneider 2019) fills this gap. Its TEIP5-encoded content is divided into author-specific and thematic archives, such as the Udo Lindenberg Archive and a Charts 2000-2018 Archive, containing nearly half a million tokens within > 1,000 lyrics. Every text is automatically processed via a customized WebLicht tool chain, and manually corrected using the collaborative corpus platform WebAnno. The POS annotation layer broadens the STTS 2.0 tagset for contractions such as *auf'm* – short for *auf dem* (engl. *on the*) – and *willst's* – short for *willst es* (engl. *want it*). Furthermore, extensible layers for named entities (differing between real and fictional person and location names), neologisms (using a rather broad definition of what is new or innovative), and rhyming forms (e.g. end rhyme, beginning rhyme, internal rhyme) are added. All manual processing steps are subject to inter-annotator agreement; the resulting corpus database is explorable online (www.songkorpus.de), allowing fine-grained retrieval and the ad hoc calculation of linguistically motivated statistics.

The corpus collection provides a sustainable empirical basis for linguistics and the broad spectrum of cultural studies. Potential research studies include: (i) empirical statements on linguistic variety (ii) sentiment analysis for selected time periods or musical genres (iii) influence of external factors on lexical diversity (iv) stylistic analyses, identification of lexical or syntactical style markers (v) dialect lyrics (vi) text similarity (vii) comparison of rhyming forms and schemes (viii) empirical approaches to phenomena such as irony and wit (iv) topic modeling for selected authors/genres/time periods (x) identification of author-/genre-/time-specific formulation patterns (xi) symbolic elements and metaphors (xii) identification of parallelisms between the notion of persons, places, or institutions, and prominent topics in public discourse.

References: Kreyer, R. & Mukherjee, J. (2007). The Style of Pop Song Lyrics: A Corpus-linguistic Pilot Study. Anglia – Zeitschrift für englische Philologie 125(1), 31–58. Schneider, R. (2019). "Konservenglück in Tiefkühl-Town" – Das Songkorpus als empirische Ressource interdisziplinärer Erforschung deutschsprachiger Poptexte. Proceedings of the Conference for the Processing of Natural Language (KONVENS), Erlangen.