Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 842–848
Marseille, 11–16 May 2020

A Corpus Linguistic Perspective on Contemporary German Pop Lyrics with the Multi-Layer Annotated « Songkorpus »

Roman Schneider
Justus Liebig University
Applied and Computational Linguistics
Otto-Behaghel-Str. 10 D, 35394 Gießen / Germany
roman.schneider@germanistik.uni-giessen.de

Abstract

Song lyrics can be considered as a text genre that has features of both written and spoken discourse, and potentially provides extensive linguistic and cultural information to scientists from various disciplines. However, pop songs play a rather subordinate role in empirical language research so far - most likely due to the absence of scientifically valid and sustainable resources. The present paper introduces a multiply annotated corpus of German lyrics as a publicly available basis for multidisciplinary research. The resource contains three types of data for the investigation and evaluation of quite distinct phenomena: TEI-compliant song lyrics as primary data, linguistically and literary motivated annotations, and extralinguistic metadata. It promotes empirically/statistically grounded analyses of genre-specific features, systemic-structural correlations and tendencies in the texts of contemporary pop music. The corpus has been stratified into thematic and author-specific archives; the paper presents some basic descriptive statistics, as well as the public online frontend with its built-in evaluation forms and live visualisations.

Keywords: Corpus Linguistics, Under-Resourced Language Varieties, Pop Lyrics, German

1. Introduction

Systematically compiled collections of written texts and speech form the most important empirical basis for linguistically motivated research on human language. Extensive data resources exist for standard and close-to-standard language varieties, supplemented by special corpora covering specific language situations and fields of application (Kupietz/Schmidt 2018, Lemnitzer/Zinsmeister 2015, Lüdeling/Kytö 2008, McEnery/Hardie 2012).

Noteworthy against this background is the lack of scientifically valid, sustainably utilizable digital collections of pop lyrics, especially for the German language. Just as pop music has evolved from an originally youth cultural phenomenon in the nineteen-fifties and -sixties into an integral part of modern culture, its textual content has become omnipresent in the realm of everyday language. We are surrounded by pop lyrics, e.g. in the form of in-car radio listening, online streaming services, ambient music for department stores and restaurants, or in the context of television shows. Added to this is a somewhat lyrical claim: song texts can be categorized as "poetry of use" – the German term "Gebrauchslyrik" was introduced 1927 by Bertold Brecht and applied to song texts by Blühdorn (2003). They are "latently poetic, but rarely authentically poetic" (Flender/Rough 1989). Quite frequently, lyrics are not just consumed for the sake of distraction, but intended as a means of conveying messages and feelings, or – on the recipient side – as a medium to find inspiration and explanations at some point in life.

In view of this considerably high "communicative impact factor" (Kreyer/Mukherjee 2007), there is a substantial desideratum regarding the consideration of pop lyrics as a separate genre in corpus linguistics. Though it is true that literary studies have discovered lyrics as a promising subject of investigation, none of the well-established large corpus collections contains lyrics. Correspondingly little studied are empirical aspects such as aesthetics and style (vocabulary, syntax, register, etc.), content (topics, e.g. in the historical and/or political context), emotion and sentiment (categorization, intensity, and distribution), or relationships between form and content. As it is customary for less-researched language varieties, initial testing and validation of statistical measures and NLP methods seem desirable. Here, too, the introduced song corpus can be seen as an important step to fill an existing gap.

2. Related Work

Despite some promising approaches towards the exploration of language characteristics for selected artists (see, e.g., von Ammon/von Petersdorff 2019), empirically grounded research on German pop lyrics on a broad base remains comparatively scarce, not least due to the non-existence of publicly available, reasonably stratified and preprocessed corpora. For the English language, however, there are inspiring examples of corpus-linguistic findings on discourse and language phenomena in lyrics. Werner (2019) contrasts a custom-built corpus of lyrics by US-American rap artists (LYRAP) to a corpus of pop lyrics (LYPOP), exploring the linguistic side of hip-hop discourse. Brett/Pinna (2018) present the Sassari Lyrics (SLY) corpus with 10 million tokens, covering various sub-genres. The BLUR corpus (Miethaner 2005) gathers more than 8,000 digitized American blues songs. Another milestone is provided by Kreyer/Mukherjee (2007) with the Gießen-Bonn Corpus of Popular Music (GBoP), which makes texts of Top 30 albums empirically evaluable; Kreyer (2012) uses this resource in order to check some well-known stereotypes and clichés for truth. Eiter (2017) examines lyrics as a specific genre between spoken and written language, and compares a custom-built 120,000 token song corpus to the balanced COCA (Davies 2015) and BNC (British National Corpus 2007) resources. Katznelson et al. (2010) and Cullen (2009) describe corpus studies of rock, pop, and country lyrics; Watanabe (2018) establishes the American Popular Music Corpus of English (PMCE-US). Bertin-Mahieux et al. (2011) have built a so-called "Million Song Dataset", while Murphey (1992) compiles an early collection of Top 50 chart songs, which is then evaluated quantitatively (eg., regarding type-token ratio) and qualitatively (eg., regarding the use of pronouns).

Other English-language collections exist for specific genre subdomains, such as the Rock Lyrics Corpus (ROLC; Falk 2013), or a corpus of Billboard 100 songs (Nishina 2017).

Regarding linguistic feature richness, Werner (2012) compares American and British English in pop songs, and outlines its didactic potential (Werner 2018). Werner/Lehl (2015) discuss practical aspects of the appropriateness of lyrics for second language learning, refining some suggestions of Plitsch (1997), who raises the idea of using contemporary lyrics for a more motivating and close-to-reality language teaching. Tegge (2017) checks the lexical coverage of pop lyrics in English language teaching, using two collections of around 1,000 pop songs. Coats (2016) mentions the unique and authentic role of music for the acquisition of lexical fluency, while Terhune (1997) takes a critical look at the undeniable fact that lyrics generally do not conform to close-to-standard syntactic rules. Squires (2018) presents experiments on the influence of nonstandard grammatical forms in pop lyrics on native speakers. Viol (2000) discusses identity-building aspects of British pop lyrics, Motschenbacher (2016) and Van Hoey (2016) compare English Eurovision Song Contest lyrics with more general corpora. Connor (2018) presents rhythmic transcriptions for rap songs, while Olivo (2001) examines rap spelling conventions. In addition to these broadly set up contributions, there exist stylistic analyzes of songwriters, eg. Johnson/Larson (2003) on the use of metaphors in Beatles' lyrics, or Morini (2013) discussing linguistic peculiarities in the lyrics of Kate Bush.

Pop lyrics are sometimes regarded as a reflection of political, economic, or social phenomena (Shukers 1998); Blühdorn (2003) illustrates this through the example of German songwriters Udo Lindenberg and Konstantin Wecker. Both Machin (2010) and Kreyer (2015) evaluate lyrics against the background of discussions about sexuality and gender-appropriate language. Napier/Shamir (2018) take a diachronic perspective and quantify emotional changes in lyrics since 1950. The results show a long-term significant increase of anger, rage and grief (with a brief decline in the mid-eighties). The expression of anxiety continues to increase until the 1980s, albeit with a lower growing rate. The expression of joy in song lyrics remains significantly decreasing over the entire period.

Also for the English language, applications of computational linguistics can be found, such as methods and tools for Text Mining, Sentiment Analysis and Topic modeling. Mahedero et al. (2005) evaluate the suitability of Natural Language Processing tools for the evaluation of pop music texts; Liske (2018) describes the use of the statistics software R for the analysis of lyrics written by artist Prince. Penaranda (2006) uses text mining for empirically based genre assignments, involving linguistic anomalies. Behl/Choudhury (2011) conduct in-depth complex network analyses in order to model specific features (vocabulary limitations, syntactical restrictions, creative word usage, etc.) of Bollywood song lyrics.

3. Corpus Building

Validating quantitative language regularities, e.g. distribution laws such as the Zipf-Mandelbrot law – covering the relationship between frequency rank and frequency of linguistic units –, functional laws such as the Menzerath-Altmann law – dealing with correlations between the length of a linguistic construct and the length of its immediate components – or logistic models like the Piotrowski law for determining the dispersion of new words from a diachronic perspective (see Köhler 2005, Biemann 2007, Schneider 2019), require strict physical integrity of all objects of study. The explanatory power of quantitive regularities only unfolds in the analysis of complete texts, because the measured variables (strophe, verse line, word etc.) are always results of individual text generation processes (Sinclair 2005).

Stratification objective is therefore the comprehensive coverage of complete works, and not just the arbitrary compilation of some lyrics' verses or phrases. As of spring 2020, two corpus archives represent the works of singers/songwriters Udo Lindenberg and Konstantin Wecker, spanning a period of five decades. A third archive (see table 1) includes German-language songs ranked in the German Top 100 single charts (BVMI 2019) since 2001, considering CD sales, internet downloads, and streaming platforms. Further archives are in preparation.

	Charts	Lindenberg	Wecker	Total
Lyrics	684	316	267	1,263
Tokens	244,276	66,560	63,453	376,157
Verselines	37,934	11,043	9,774	59,085
Strophes	5,903	1,832	2,343	10,166

Table 1: The « Songkorpus » Archives

Pop lyrics, as well as the vast majority of digital resources in linguistics, are typically subject to third parties' rights. Therefore, in order to protect intellectual property, it is necessary to conclude licensing agreements. For artists who have kindly agreed to provide their lyrics for non-commercial, scientific research, the corresponding annotated archives are available in TEI-compliant XML format, and can be downloaded for further exploration. The charts archive, in contrast, is available in a bag-of-words format only, where tab-separated columns contain the number of occurences for a token or lemma per year, and can be used for further processing with statistical tools.

In order to cover various levels of granularity and to ensure interoperability, all lyrics are formatted using structural descriptions according to TEI P5 (TEI Consortium 2019). E.g., the element types <lg> (linegroup) and <l> (line) mark strophes and verse lines; performance directives are annotated with <stage> elements, header element types like <titleStmt>, <publicationStmt>, and <sourceDesc> contain metadata concerning the source or publication of songs.

The initial sentence and word segmentation has to deal with the challenging fact that lyrics primarily have to function acoustically. The written textual form often does not contain punctuation marks, or at least does not use them consistently for the identification of phrases or sentences. As a consequence, fully-automatic detection and annotation of such units produces rather poor results, and must be supported by manual pre-processing. Applying again TEI-standards – namely the element types <add> and <del> –, the original lyrics are transferred into close-to-standard representations, enabling further processing with NLP tools, but also retranslation at all times.

The song bodies are then submitted to the CLARIN infrastructure component WebLicht (Hinrichs et al. 2010). A customized tool chain is worked through, including the IMS tokenizer, TreeTagger with STTS part of speech tagset (Schiller 1999), a named entity recognizer trained on TuebaDZ, and the Berkeley constituent parser. The immediate results confirm the assumption that application of standard-language-oriented categories and procedures to less homogenous language varieties requires specific adaptations (Horbach et al. 2014, Karlova-Bourbonus et al. 2016, Zinsmeister et al. 2014). Lyrics are no exception. Examplary phenomena that deserve systematical treatment are syntactic constructions without subject (*hab noch Sehnsucht*, engl. *still have longing*) and contracted forms, e.g. verb and personal pronoun (*machste*, engl. *you make*), verb and article (*bistn*/*bist'n*, engl. *are an*), or comparison conjunction and article (*wien*/*wie'n*, engl. *like an*). The variety encountered within the lyrics corpus even exceeds the CMC-related extensions discussed in (Westpfahl 2014).

Overall, lyrics often 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, all annotation steps need to be reviewed, so the corpus processing takes place as an interplay between automated annotation runs and manual post-editing. For this purpose, WebLicht results are imported into the web-based curation platform WebAnno (Eckart de Castilho et al. 2016). This allows the application of an extended POS tagset (based on Bartz et al. 2014, Beißwenger et al. 2015, Westpfahl et al. 2017). Its enhanced inventory includes appropriate POS tags for the newly discovered contracted forms mentioned above.

During post-processing, new classes for named entitites are introduced, based on (Benikova et al. 2014). Starting with four established main classes (LOCation, ORGanization, PERson and OTHer), three subclasses are accepted in each case: partitive (e.g. [Bahama-Landebahn]LOC$_{\text{Part}}$), derived (e.g. [Berliner]LOC$_{\text{Deriv}}$ Bär), and fictitious (e.g. [Bodo Ballermann]PER$_{\text{Fict}}$). Nested structures are covered as well (e.g. [Radio [Luxusburg]LOC$_{\text{Fict}}$]ORG$_{\text{Fict}}$). As a fourth main NE class, TIME specifications and intervalls are annotated (e.g. [1990]TIME, [nach 20 Jahren]TIME). Furthermore, a special layer for neologisms and occasionalisms offers the opportunity for handling innovative language and puns. The corresponding tagset comprises "new word" (e.g. *vorherragend* instead of *hervorragend*, engl. *outstanding*), "new meaning" (e.g. *Oberindianer*, addressing not the chief of an indian tribe, but the former GDR head of state), "word combination" (even multilingual, e.g. *howauchever* as a parody of engl. *howsoever*, integrating German *auch*), and "intentional misspelling" (e.g. *Lusthansa* instead of *Lufthansa*). Finally, an annotation layer for linking of rhyming words is added ("initial rhyme", "internal rhyme", "end rhyme"). All annotations – multiple classes, multiple annotators – are subject to inter-annotator reliability, using Fleiss' kappa.

4. Data Exploration

The curated annotation layers are exported using the WebAnno TSV export format, which is similar to CoNNL file formats, but adds specialized layer information to the header and column representations. Together with the TEI-compliant XML instances, they are stored within an object-relational database system, providing fast and powerful retrieval options. A dedicated website (*songkorpus.de*) offers combined search by various attributes like token, lemma, and POS, as well as the exploration of aggregated statistics and live visualizations.

Lyrics can be considered as a text genre that has features of both written and spoken discourse. Its conceptual textuality may be based on the circumstances that it does not allow non verbal techniques or direct feedback, e.g. asking of clarifying questions. But corpus analyses can be used to identify presumed features of conceptual orality, like character iterations for emulating prosody (*Dann die erste Liebe, Mensch, hab' ich gebrannt, ich war* ***sooo*** *angetörnt, nur noch im Fünfeck rumgerannt*) or reduplications (***Du bist und bist*** *nicht daheim, nur* ***irgend irgendwo*** *drin, wie* ***jeder jeder*** *allein, schon* ***lange lange*** *getrennt*).

Particularly for applied disciplines such as stylometry, the vocabulary richness (Yule 1944) respectively lexical diversity (Carroll 1938) open up an interesting field of investigation. The idea starts from the assumption that measured values like type token ratio are indicators for the individual vocabulary size of an author (Tanaka-Ishii/Aihara 2015). One methodological problem remains the fact that, as a consequence of the Zipf-Mandelbrot law (Mandelbrot 1953), almost all measures (like TTR, STTR) vary depending on corpus size (Tweedie and Baayen 1998, Evert et al. 2017). To approach these issues, the corpus portal (Songkorpus 2020) offers a range of useful metadata, parameters, and statistical graphs.

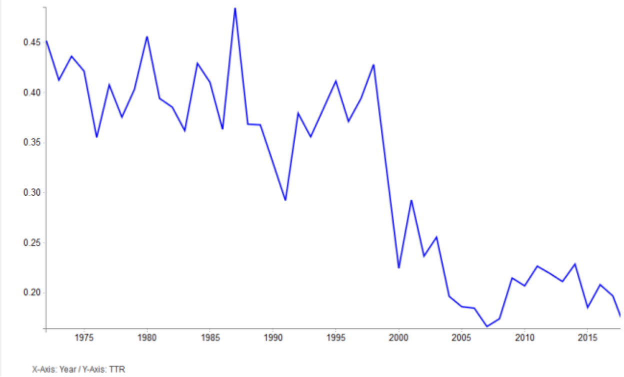

Figure 1: Overall TTR of all corpus archives.

Figure 1 visualizes a significant decrease of the corpus type token ratio since the turn of the millennium. This seems primarily due to the fact that the included charts archive merely covers the period since 2000 – and the Lindenberg and Wecker lyrics, starting in the 1970s, show far higher figures. The overall TTR (~ 0.1) is substantially lower than the measured values for certain years or albums (mostly 0.3 to 0.5), which can be traced back to different sample sizes, but also to the observation that lyrics regularly tackle the same themes and issues. And obviously, the repetition of words and phrases in refrains is responsible for lower TTR values than in custom corpora.

Other statistical evaluations that can be carried out with the online corpus explorer include frequency analyses on character, word, verse, song, and corpus level. A mere look on the most frequent words shows up significant differences to corpus-based word form lists like DeReWo (2014): The top ranks do not start with articles, but with the personal pronoun *ich* (engl. *I*); top bigrams are *ich bin*, *du bist* (engl. *I am*, *you are*) etc., and – to confirm a classic cliché – the highest ranking trigrams of the charts archive read *na na na* and *la la la*. Based on these statistics, the online frontend allows to verify quantitative regularities like Zipf's law or Menzerath's law, computing the correlation between length of songs (in strophes) and length of strophes (in verse lines).

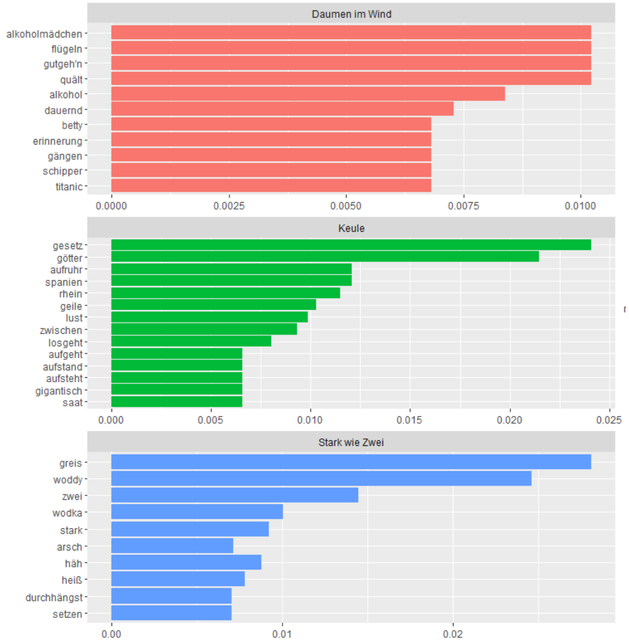

Figure 2: Highest TF-IDF words of selected albums.

Lyrics can be used to calculate the most important keywords for certain artists, time periods, or albums (see figure 2), using weighting schemes like TF-IDF (Manning et al. 2008, 118). Evaluating the neologism annotation layer, innovative word formations may be identified. Many of these new words are examples of portmanteaus: *Laprotz* or *Pumidas* as funny blends of brand names (and the German equivalent of the verb *to splurge* in the former case), and *klaufte* as a mesh of *kaufen* (*to buy*) and *klauen* (*to steal*) – considering that the first evidence of this word in DeReKo (Kupietz et al. 2018) dates back to 1971, its use on the 1976 album *Galaxo Gang* proves Udo Lindenberg as early adopter. Another example even shows a true neologism: *Luxusburg*, a combination of the German word for *luxury* and the country name *Luxemburg*, comes from the 1981 album *Udopia* (a nonce word itself). The only earlier evidence in DeReKo uses its literal sense, and simply describes a luxurious palace. But starting with the mid-80s, the ironic use of *Luxusburg* for the wealthy little state in central Europe can be verified in various newspaper and journal articles.

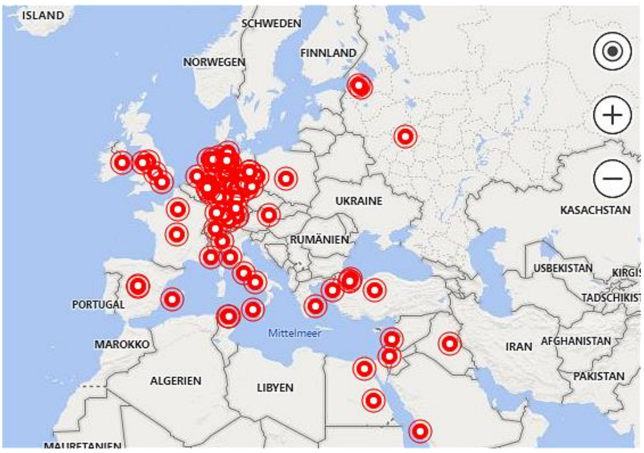

Figure 3: Distribution of location names.

To give one last example: As a special application of named entity recognition, the online corpus explorer offers geovisualization of identified place names that arise in the lyrics. Using their geographical coordinates from a database lookup table, LOC entities are displayed on a map. This allows a quick overview of countries, cities, and even buildings or monuments that are subject of song texts. Figure 3 shows the distribution within Europe, allowing a traceback from locations to corresponding lyrics or artists.

5. Conclusion and Outlook

The multi-layer annotated «Songkorpus» provides a number of potentially valuable primary and meta data for empirical research on contemporary German pop lyrics. It thus fills a public data gap in the continuum between both standard and nonstandard, written and spoken language, that previously prevented comprehensive and statistically founded answers to syntactic, semantic or pragmatic questions for this genre. The multidisciplinary links appear exceedingly promising: besides linguistics and literature, benefiting research areas can be located in the broad spectrum of language didactics, (socio)cultural studies, musicology, or media studies.

The machine-readable corpus data can be explored online, or downloaded for further statistical processing. Unlike existing free lyrics sites that remain a legal grey area, «Songkorpus» offers a consistent, sustainable, legally sound solution for academic research. The processing pipeline involves automatic tagging – that will increasingly use specially trained NLP tools, as data is accumulated – and manual verification steps.

Further plans are to extend the chronological and sub-genre span by adding more collected works of individual artists, and cross-sectional archives. In view of the largely text-based nature of already included singer/songwriter compositions, it would be a worthwhile option to compare its linguistic characteristics and thematic tendencies with, for instance, German hip-hop discourse.

6. Bibliographical References

von Ammon, F. and von Petersdorff, D. (Eds.). 2019. Lyrik/lyrics. Songtexte als Gegenstand der Literaturwissenschaft. Wallstein Verlag, Göttingen.

Bartz, T., Beißwenger, M. ans Storrer, A. 2014. Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. In Journal for Language Technology and Computational Linguistics 28 (1), pages 157–198.

Behl, A. and Choudhury, M. 2011. A corpus linguistic study of bollywood song lyrics in the framework of complex network theory. In Proceedings International Conference on Natural Language Processing. Macmillan Publishers, India.

Beißwenger, M., Bartz, T., Storrer, A. and Westpfahl, S. 2015. Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation. Empirikom shared task on automatic linguistic annotation of internet-based communication (EmpiriST 2015). URL : http://sites.google.com/site/empirist2015/

Benikova, D., Biemann, C. and Reznicek, M. 2014. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik.

Bertin-Mahieux, T., Ellis, D., Whitman, B. and Lamere, P. 2011. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference.

Biemann, C. 2007. A Random Text Model for the Generation of Statistical Language Invariants. In Proceedings of HLT-NAACL-07. Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics. Rochester, NY, USA.

Blühdorn, A. 2003. Pop and Poetry – Pleasure and Protest: Udo Lindenberg, Konstantin Wecker and the Tradition of German Cabaret. In German Linguistic and Cultural Studies, Bd 13.

Brett, D. and Pinna, A. 2018. Words (don’t come easy): The Automatic Retrieval and Analysis of Popular Song Lyrics. Leiden/NL, Brill Publishers, pages 307–325. DOI: https://doi.org/10.1163/9789004390652_014

BVMI (Bundesverband Musikindustrie e.V.). 2019. Systembeschreibung der Offiziellen Deutschen Charts. Version 4.7. URL : https://www.musikindustrie.de/markt-bestseller/offizielle-deutsche-charts/systembeschreibung

Carroll, J. B. 1938. Diversity of Vocabulary and the Harmonic Series Law of Word-frequency Distribution. In The Psychological Record. 2, 16, pages 379–386.

Coats, G. 2016. Analyzing song lyrics as an authentic language learning opportunity. In Report of the Central States Conference on the Teaching of Foreign Languages, Vol. 1, pages 1–22.

Connor, M. 2018. The musical artistry of rap. Jefferson: McFarland.

Cullen, B. 2009. A Corpus Analysis of Pop Song Lyrics. New Directions. Nagoya Institute of Technology.

Eckart de Castilho, R., Mújdricza-Maydt, É. Muhie Yimam, S., Hartmann, S., Gurevych, I., Frank, A. and Biemann, C. 2016. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In Proceedings of the LT4DH workshop at COLING 2016, Osaka.

Eiter, A. 2017. ‘Haters gonna Hate’: A Corpus Linguistic Analysis of the Use of Non-Standard English in Pop Songs. University of Innsbruck, Department of English Studies. DOI: 10.13140/RG.2.2.31181.33763

Evert, S., Wankerl, S. and Nöth, E. 2017. Reliable measures of syntactic and lexical complexity: The case of Iris Murdoch. In Proceedings of the Corpus Linguistics 2017 Conference, Birmingham, UK.

Falk, J. 2013. We Will Rock You: A Diachronic Corpus-based Analysis of Linguistic Features in Rock Lyrics. Växjö: Linnaeus University.

Flender, R. and Rauhe, H. 1989. Popmusik: Aspekte ihrer Geschichte, Funktionen, Wirkung und Ästhetik. Darmstadt: Wissenschaftliche Buchgesellschaft.

Hinrichs, M. Zastrow, T. and Hinrichs. E. 2010. WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010), Malta.

Horbach, A., Steffen, D., Thater, S. and Pinkal, M. 2014. Improving the performance of standard part-of-speech taggers for computer-mediated communication. In Proceedings of KONVENS 2014, Hildesheim, Germany.

Johnson, M.L. and Larson, S. 2003. ‘Something in the Way She Moves’: Metaphors of musical motion. In Metaphor and Symbol 18(2), pages 63–84.

Karlova-Bourbonus, N., Grumt Suárez, K. and Lobin, H. 2016. Compilation and Annotation of the Discourse-structured Blog Corpus for German. In Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, Ljubljana.

Katznelson, N., Gelman, J., Lindblom, K. and Caput, M. 2010. American Song Lyrics: A Corpus-Based Research Project Featuring Twenty Years in Rock, Pop, Country and Hip-Hop. San Francisco, CA: San Francisco State University.

Köhler, R. 2005. Korpuslinguistik. Zu wissenschaftstheoretischen Grundlagen und methodologischen Perspektiven. In LDV-Forum, 20 (2), pages 1–16.

Kreyer, R. 2012. Love is like a stove – It burns you when it's hot: A corpus-linguistic view on the (non-)creative use of love-related metaphors in pop songs. In Hoffmann, S., Rayson, P. and Leech, G. (Eds.). English Corpus Linguistics: Looking Back. Moving Forward, pages 103–115.

Kreyer, R. 2015. “Funky fresh dressed to impress”: A corpus-linguistic view on gender roles in pop songs. In International Journal of Corpus Linguistics, 20 (2), pages 174–204.

Kreyer, R. and Mukherjee, J. 2007. The Style of Pop Song Lyrics: A Corpus-linguistic Pilot Study. In Anglia - Zeitschrift für englische Philologie, 125 (1), pages 31–58.

Kupietz, M. and Schmidt, T. 2018. Korpuslinguistik. Germanistische Sprachwissenschaft um 2020. Band 5. Berlin: Walter de Gruyter.

Lemnitzer, L. and Zinsmeister, H. 2015. Korpuslinguistik. Eine Einführung. Tübingen: Narr.

Liske, D. 2018. Lyric Analysis with NLP & Machine Learning with R. DataCamp.

Lüdeling, A. and Kytö, M. (Eds.). 2008. Corpus Linguistics. An International Handbook. Handbücher zur Sprach- und Kommunikationswissenschaft 29 (1-2). Berlin: de Gruyter.

Machin, D. 2010. Analysing Popular Music: Image, Sound, Text. Los Angeles, CA: Sage.

Mahedero, J., Martínez, A., Cano, P., Koppenberger, M. and Gouyon, F. 2005. Natural language processing of lyrics. In Proceedings of the 13th annual ACM international conference on Multimedia (MULTIMEDIA '05). ACM, New York, NY, pages 475–478.

Mandelbrot, B. 1953. An information theory of the statistical structure of language. In Jackson, W. (Ed.). Communication Theory. New York: Academic Press, pages 503–512.

Manning, C.D., Raghavan, P. and Schütze, H. 2008. Introduction to Information Retrieval, Cambridge University Press.

McEnery, T. and Hardie, A. 2012. Corpus Linguistics: Method, theory and practice. Cambridge: Cambridge University Press.

Miethaner, U. 2005. I can look through muddy water: Analyzing Earlier African American English in Blues Lyrics (BLUR). Regensburger Arbeiten zur Anglistik und Amerikanistik 47. Frankfurt am Main: Peter Lang.

Morini, M. 2013. Towards a musical stylistics: movement in Kate Bush's "Running up that Hill". In Language and Literature 22 (4), pages 283–297.

Motschenbacher, H. 2016. A corpus linguistic study of the situatedness of English pop song lyrics. In Corpora 11.1, pages 1–28.

Murphey, T. 1992. The Discourse of Pop Songs. In TESOL Quarterly 26, pages 770–774.

Napier, K. and Shamir, L. 2018. Quantitative Sentiment Analysis of Lyrics in Popular Music. In Journal of Popular Music Studies, Vol. 30 No. 4, December 2018, pages 161–176.

Nishina, Y. 2017. A Study of Pop Songs based on the Billboard Corpus. In International Journal of Language and Linguistics, Vol. 4, No. 2, June 2017, pages 125–134.

Olivo, W. 2001. Phat lines: Spelling conventions in rap music. In Written Language & Literacy, 4(1), pages 67–85.

Penaranda, J. 2006. Text Mining von Songtexten. Diplomarbeit. Technische Universität Wien.

Plitsch, A. 1997. Music + Song = Authentic Listening in the Language Classroom. In Der Fremdsprachliche Unterricht Englisch 31 (1), pages 4–13.

Schiller, A., Teufel, S. and Stöckert, C. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Technical paper, University of Stuttgart: Institut für Maschinelle Sprachverarbeitung (IMS). URL : http://www.sfs.uni-tuebingen.de/resources /stts-1999.pdf

Schneider, R. 2019. Mehrfach annotierte Textkorpora. Strukturierte Speicherung und Abfrage. Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP) 8. Tübingen: Narr.

Shuker, R. 1998. Key Concepts in Popular Music. London: Routledge.

Sinclair, J. 2005. Corpus and Text: Basic Principles. In Martin Wynne (Ed.): Developing Linguistic Corpora: A Guide to Good Practice. Oxford: Oxbow Books, pages 1–16.

Squires, L. 2018. Genre and linguistic expectation shift: Evidence from pop song lyrics. In Language in Society. 48, pages 1–30.

Tanaka-Ishii, K. and Aihara, S. 2015. Computational Constancy Measures of Text. Yule's K and Rényi's Entropy. In Computational Linguistics 41 (3), pages 481–502.

Tegge, F. 2017. The lexical coverage of popular songs in English language teaching. In System, No. 67, pages 87–98.

TEI Consortium. 2019. TEI P5: Guidelines for Electronic Text Encoding and Interchange 3.5.0. URL : http://www.tei-c.org/Guidelines/P5/

Terhune, T. 1997. Pop Songs: Myths and Realities. In The English Connection 1 (1), pages 8–12.

Tweedie, F.J. and Baayen, H. 1998. How variable may a constant be? In Computers and the Humanities 32, pages 323–352.

Van Hoey, T. 2016. 'Love love peace peace': A corpus study of the Eurovision Song Contest. Graduate Institute of Linguistics, National Taiwan University.

Viol, C.-U. 2000. A Crack in the Union Jack? National Identity in British Popular Music. In Diller, H., Otto, E. and Stratmann, G. (Eds.). Youth Identities: Teens and Twens in British Culture. Heidelberg: Winter, pages 81–106.

Watanabe, A. 2018. A Style of Song Lyrics: The Case of Really. In Zephyr (2018), 30, pages 12–27.

Werner, V. 2012. Love is all around: a corpus-based study of pop lyrics. In Corpora 7 (1), pages 19-50.

Werner, V. (Ed.), 2018. The language of pop culture. Routledge Studies in Linguistics 17. New York: Routledge.

Werner, V. 2019. Assessing hip-hop discourse: Linguistic realness and styling. In Text&Talk. An Interdisciplinary Journal of Language, Discourse & Communication Studies; 39(5), pages 671–698.

Werner, V. and Lehl, M. 2015. Pop lyrics and language pedagogy: A corpus-linguistic approach. In Formato, F. and Hardie, A. (Eds.) : Corpus Linguistics 2015, Lancaster: UCREL, pages 341–343.

Westpfahl, S. 2014. STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data. In Proceedings of LAW VIII – The 8th Linguistic Annotation Workshop. Association for Computational Linguistics (ACL Anthology W14-49), pages 1–10.

Westpfahl, S., Schmidt, T., Jonietz, J. and Borlinghaus, A. 2017. STTS 2.0. Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS). Arbeitspapier. Mannheim: Institut für Deutsche Sprache. URL : https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/6063

Yule, G.U. 1944. The Statistical Study of Literary Vocabulary. Cambridge University Press, Cambridge.

Zinsmeister, H., Heid, U., Beck, K. 2014. Adapting a part-of-speech tagset to non-standard text: The case of STTS. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik.

7. Language Resource References

The British National Corpus. 2007. Version 3 (BNC XML Edition). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk

Davies, M. 2015, Corpus of Contemporary American English (COCA), Harvard Dataverse, V2. URL : https://doi.org/10.7910/DVN/AMUDUW

DeReWo. 2014. Corpus-Based Lemma and Word Form Lists. URL: https://www.ids-mannheim.de/kl/projekte/methoden/derewo.html

Kupietz, M., Lüngen, H., Kamocki, P., Witt, A. 2018. The German Reference Corpus DeReKo: New Developments – New Opportunities. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki: European Language Resources Association (ELRA), pages 4353–4360. URL : https://www.ids-mannheim.de/kl/projekte/korpora.html

Songkorpus. 2020. Corpus of German Song Lyrics. Version 1.1. URL: http://songkorpus.de