Originally published in: Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 – 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg. - München [u.a.]: German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg, 2019. Pp. 229-236

"Konservenglück in Tiefkühl-Town" – Das Songkorpus als empirische Ressource interdisziplinärer Erforschung deutschsprachiger Poptexte

Roman Schneider
Justus-Liebig-Universität
Angewandte Sprachwissenschaft und Computerlinguistik
Otto-Behaghel-Str. 10 D, 35394 Gießen
roman.schneider@germanistik.uni-giessen.de

Abstract

Der Beitrag beschreibt ein mehrfach annotiertes Korpus deutschsprachiger Songtexte als Datenbasis für interdisziplinäre Untersuchungsszenarien. Die Ressource erlaubt empirisch begründete Analysen sprachlicher Phänomene, systemisch-struktureller Wechselbeziehungen und Tendenzen in den Texten moderner Popmusik. Vorgestellt werden Design und Annotationen des in thematische und autorenspezifische Archive stratifizierten Korpus sowie deskriptive Statistiken am Beispiel des Udo-Lindenberg-Archivs.

1 Einleitung

Natürlichsprachliche Korpora als systematisch zusammengestellte Digitalisate von Kommunikationsakten bilden die wichtigste empirische Grundlage linguistisch motivierter Forschung. Für die standardnahe deutsche Gegenwartssprache existieren umfangreiche Korpussammlungen literarischer, journalistischer, juristischer, wissenschaftlicher und anderer weit verbreiteter Textsorten, ergänzt durch diverse Spezialkorpora zur Abdeckung spezifischer Sprachumstände (vgl. Kupietz/Schmidt 2018, Lemnitzer/Zinsmeister 2015, Lüdeling/Kytö 2008).

Bemerkenswert erscheint vor diesem Hintergrund das Fehlen einer wissenschaftlich validen, nachhaltig nutzbaren digitalen Sammlung von Popmusiktexten. So wie sich die Popmusik von einem ursprünglich jugendkulturellen Phänomen in den 1950er-/1960er-Jahren zu einem festen Bestandteil der Alltagskultur entwickelt hat, sind deren textuellen Inhalte in der Sprachrealität inzwischen allgegenwärtig und zunehmend Gegenstand (qualitativer) Forschung (vgl. von Ammon/von Petersdorff 2019). Wir sind von ihnen umgeben, nicht nur beim Radiohören während des Autofahrens, beim Einkaufen im Supermarkt, via Online-Streamingdienst oder in TV-Shows. Hinzu kommt ein durchaus lyrischer Anspruch: Moderne Popsongtexte als „Gebrauchslyrik" (Blühdorn 2003) sind „latent poetisch, aber selten authentisch poetisch" (Flender/Rauhe 1989). Sie dienen oft nicht allein der simplen Zerstreuung, sondern werden genutzt, um Botschaften und Gefühle zu vermitteln oder – auf Rezipientenseite – Inspiration und Erklärungen zu finden.

Angesichts dieses beachtlichen „kommunikativen Impact Factors" (Kreyer/Mukherjee 2007) besteht ein substanzielles Desiderat hinsichtlich der Berücksichtigung des Popmusik-Genres in der Korpuslinguistik. Keine der etablierten Sammlungen enthält Songtexte, entsprechend wenig erforscht sind spezifische Aspekte wie Ästhetik und Stil (Vokabular, Syntax, Register etc.), Inhalt (Thematiken, z. B. im historischen/politischen Kontext), Emotionalität (Kategorisierung, Intensität und Verteilung) oder Beziehungen zwischen Form und Inhalt. Wie für wenig erforschte Sprachgenres üblich, erscheinen initiale Erprobung und Validierung statistischer Maße und Verfahren aufschlussreich, auch hier stößt das Songkorpus in eine bestehende Lücke.

2 Stand der Kunst

Nachhaltige, empirisch begründete Forschung zu Texten deutschsprachiger Popmusik bleibt bislang aufgrund der Nichtexistenz ausreichend stratifizierter und aufbereiteter Daten ein unerfülltes interdisziplinäres Desiderat. Für das Englische hingegen lassen sich inspirierende Beispiele korpuslinguistischer Forschung zu Diskurs und Sprache in Songtexten finden. So enthält das BLUR-Korpus (Blues Lyrics Collected at the University of Regensburg; Miethaner 2005) mehr als 8.000 digitalisierte Texte und bildet damit eine wert-

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-93189

volle Ressource für die Erforschung amerikanischer Bluessongs. Einen weiteren Meilenstein der Songtextforschung liefern Kreyer/Mukherjee (2007) mit dem von ihnen kompilierten Gießen-Bonn Corpus of Popular Music (GBoP), das englischsprachige Texte von Top-30-Alben empirisch auswertbar macht. Katznelson et al. (2010) und Cullem (2009) beschreiben Korpusanalysen zu amerikanischen Songtexten; Watanabe (2018) begründet das American Popular Music Corpus of English (PMCE-US). Bertin-Mahieux et al. (2011) haben ein „Million-Song-Dataset" aufgebaut, während Murphey (1992) eine frühe Sammlung aus Top-50-Chartsongs kompiliert, quantitativ analysiert (z. B. hinsichtlich des Type-Token-Verhältnisses) und qualitativ auswertet (z. B. hinsichtlich der Verwendung von Pronomina). Weitere englischsprachliche Korpora existieren zu spezifischen Subdomänen, beispielsweise das Rock Lyrics Corpus (ROLC; Falk 2013).

Werner (2012) vergleicht amerikanisches und britisches Englisch in Popsongs und beschreibt Nutzungsaspekte für das Zweitsprachenlernen (Werner/Lehl 2015). Bereits Plitsch (1997) thematisiert den motivierenden Einsatz von Popmusiktexten für den Sprachunterricht, während Terhune (1997) hier insbesondere den syntaktisch oft nicht standardkonformen Aufbau von Songtexten kritisch sieht. Viol (2000) diskutiert identitätsstiftende Phänomene in britischen Popmusiktexten, Motschenbacher (2016) und Van Hoey (2016) vergleichen Eurovision-Song-Contest-Texte mit breiter stratifizierten Korpora. Diskurse von Weiblichkeit und Männlichkeit in Popsongs untersucht Kreyer (2015); Nishina (2017) setzt sprachexterne Faktoren wie Musikgenre und Geschlecht der Interpreten in Bezug zu linguistisch motivierten Analysen (Type Token Ratio, n-Gramme usw.) und kompiliert ein privates Untersuchungskorpus aus Billboard-Songs einer Dekade. Eiter (2017) untersucht Songtexte als Phänomen zwischen gesprochener und geschriebener Sprache. Ergänzend zu solchen übergreifenden Beiträgen finden sich stilistische Analysen einzelner Autoren, etwa von Johnson und Larson (2003) zur Verwendung von Metaphern in Beatles-Texten oder von Morini (2013) zu sprachlichen Eigenheiten in den Songtexten von Kate Bush.

Nicht selten werden Popsongs und ihre Texte als Spiegel gesellschaftlicher Entwicklungen betrachtet (Shukers 1998). Anderson et al. (2003) beschäftigen sich mit Korrelationen aggressiver Handlungen und der Konsumation von als aggressiv klassifizierten Texten. Machin (2010) analysiert Songtexte vor dem Hintergrund aktueller Diskussionen um Sexualität und geschlechtergerechte Sprache. Eine diachrone Perspektive nehmen Napier/Shamir (2018) ein und beziffern mithilfe quantitativer Maße emotionale Veränderungen in Songtexten der zurückliegenden Dekaden seit 1950. Ihre Ergebnisse weisen einen langfristig signifikanten Anstieg der Kategorien Ärger, Wut und Trauer (mit einem kurzzeitigen Rückgang Mitte der 1980er-Jahre) nach. Der Ausdruck von Angst nimmt bis in die 1980er-Jahre hinein ebenfalls kontinuierlich zu, allerdings mit geringerer Steigerungsrate. Deutlich zurückgegangen über den Gesamtzeitraum ist der Ausdruck von Freude.

In jüngerer Zeit kommen verstärkt computerlinguistische Methoden und Werkzeuge für Text Mining, Sentiment Analysis oder Topic Modeling zum Einsatz. Mahedero et al. (2005) evaluieren die Eignung von Natural Language Processing-Tools für die Auswertung von Popmusiktexten; Liske (2018) beschreibt den Einsatz der Statistikumgebung R für die Analyse von Songtexten des Künstlers Prince. Penaranda (2006) verwendet Text Mining für empirisch begründete Genre-Zuordnungen auf Basis sprachlicher Auffälligkeiten.

3 Korpusdesign und -aufbereitung

Eine Grundvoraussetzung solider empirischer Erforschung sprachimmanenter Phänomenbereiche ist die technisch-physische Integrität der Primärdaten. Insbesondere der Nachweis statistischer Regularitäten hat unter Beachtung strikter Gültigkeitsbedingungen zu erfolgen, zu denen die Gewährleistung intakter Forschungsobjekte zählt (Schneider 2019, 32f.). So lassen sich auf Häufigkeitsverteilungen, Längenmessungen etc. basierende Gesetzmäßigkeiten der Textebene nachweislich nicht unter Zuhilfenahme von willkürlich kompilierten Fragmentsammlungen aus Verszeilen oder Sätzen nachweisen. Zu diesen quantitativen Korrelationen zählen Verteilungsgesetze wie das Zipf-Mandelbrot-Gesetz über den Zusammenhang zwischen Häufigkeitsrang und Frequenz lexikalischer Einheiten, funktionale Gesetze wie das Menzerathsche Gesetz über den Zusammenhang zwischen der Länge eines sprachlichen Konstrukts und der Länge seiner unmittelbaren Komponenten, oder Entwicklungsgesetze wie das Piotrovskiy-Altmann-Gesetz zur Bestimmung der Verwendungshäufigkeiten sprachlicher Einheiten aus diachroner Perspektive (vgl. Köhler 2005, Biemann 2007). Die Erklärungskraft all dieser Korrelationen entfaltet sich erst bei der Analyse zusammenhängender und ungekürzter Texte, da die

Messgrößen (Wort-, Morphem- oder Phoneminventar, Strophen- und Verszeilenlängen usw.) stets das Resultat individueller Textgenerierungsprozesse sind (Sinclair 2005).

Ziel des Korpusaufbaus ist deshalb die möglichst umfassende Abdeckung kompletter Werke. Intern fächert sich das Songkorpus auf in autorenspezifische Archive wie das initiale Udo-Lindenberg-Archiv und themenspezifische Archive, beispielsweise eine als Chart-Song-Archiv firmierende Sammlung sämtlicher deutschsprachigen Top-100-Songtexte der zurückliegenden 20 Jahre.

Besondere Aufmerksamkeit verdient die Nutzungs- und Urheberrechtsproblematik: Grundlage des Schutzes schöpferischer Leistungen in Form von Songtexten ist das Urheberrechtsgesetz (UrhG); nach § 1 UrhG erstreckt sich der Schutz auf Werke der Literatur, Wissenschaft und Kunst. Zwar bestehen seit 2018 durch das Urheberrechts-Wissensgesellschafts-Gesetz großzügigere Regelungen für Forschungs- und Bildungseinrichtungen, trotzdem bleibt für die öffentliche Bereitstellung geschützter Inhalte über Recherche-Schnittstellen eine explizite Autorisierung der Nutzungsrechte erforderlich. Im Rahmen des Songkorpus-Aufbaus werden deshalb für öffentlich zugängliche Archive entsprechende Übertragungsvereinbarungen mit den Rechteinhabern getroffenen.

Zur Gewährleistung der Interoperabilität erfolgt die Kodierung der Songtexte vermittels standardisierter Strukturbeschreibungen gemäß TEI P5 (TEI Consortium 2019), die spezielle Elementtypen für Strophen und Verszeilen bereitstellen. Nach der aufwändigen Segmentierung in Token, Verszeilen, Strophen und Sätze – Songtexte müssen primär akustisch funktionieren und enthalten deshalb selten Interpunktionszeichen zur Identifizierung von Sinneinheiten wie Phrasen und Sätzen – schließt sich eine Anreicherung um Annotationen für interdisziplinäre Fragestellungen an:

- Lemmata
- Wortklassen, Morphologie, Syntax
- Neologismen bzw. originelle Produkte von Wortbildungsprozessen
- Named Entities als Identifizierung von realen und fiktiven Personen, Figuren, Institutionen, Ortsnamen etc.
- Reimformen und Reimschemata

Die Adaption von an standardnaher Sprache orientierten Kategorien und Verfahren an weniger homogene Sprachvarietäten erfordert spezifische Anpassungen (Horbach et al. 2014, Karlova-Bourbonus et al. 2016, Zinsmeister et al. 2014); Songtexte machen hier keine Ausnahme. Exemplarisch seien Konstruktionen ohne Subjekt (*hab noch Sehnsucht*) sowie kontraktierte Formen von Verb und Personalpronomen (*machste*) oder Vergleichskonjunktion und Artikel (*wie'n*) genannt; die im Songkorpus angetroffene Vielfalt übersteigt diesbezüglich noch die in Westpfahl (2014) für den Bereich der Computer Mediated Communication (CMC) diskutierte Liste.

Insgesamt findet sich in den Texten häufig ein bewusstes Spiel mit Normen auf vielfältigen linguistischen Ebenen (Satzstrukturen, Schreibung, Semantik, Wortarten, Wortbildung etc.). Aus diesem Grund erfolgt die Korpusaufbereitung als Wechselspiel zwischen automatisierten Annotationsläufen und manueller Nachbearbeitung. Zunächst wird auf eine für das Songkorpus maßgeschneiderte Toolchain der CLARIN-Infrastrukturkomponente WebLicht (Hinrichs et al. 2010) zurückgegriffen, bestehend aus IMS-Tokenizer, TreeTagger mit STTS-Tagset (Schiller 1999), einem auf TuebaDZ trainierten Named Entity Recognizer sowie dem Berkeley Constituent Parser. Für die Kontrolle und ggf. Korrektur der Resultate erfolgt deren Import in die kollaborative Korpusplattform WebAnno (Eckart de Castilho et al. 2016). Dort kommen dann, neben einem um Phänomene der konzeptionellen Mündlichkeit erweiterten Wortklassen-Tagset (basierend auf Bartz et al. 2014, Beißwenger et al. 2015, Rehbein et al. 2012, Westpfahl et al. 2017) auch Layer und Tagsets für die Auszeichnung von Named Entitites (basierend auf Benikova et al. 2014), Neologismen (z. B. Neuwort, Neubedeutung, Wortkombination) und Reimformen (z. B. Anfangsreim, Binnenreim, Endreim) zum Einsatz. Sämtliche manuellen Bearbeitungsschritte unterliegen während des Kurationsprozesses einer finalen Bewertung unter Zuhilfenahme von Verfahren für die Inter-Annotator-Reliabilität (Kappa-Statistiken).

4 Deskriptive Statistiken und Analysen

Das Udo-Lindenberg-Archiv versammelt mehr als 300 Texte des Pioniers der deutschsprachigen Rock- und Popsongs – und damit sämtliche nichtfremdsprachigen Texte des Autors aus fünf Jahrzehnten sowie einzelne unveröffentlichte Songs.

	Lindenberg-Archiv	*Chart-Song-Archiv*
Songtexte	301	684
Wortformen	62.807	244.276
Verszeilen	10.688	37.734
Strophen	1.769	5.803

Tabelle 1. Archive im Songkorpus (Stand 10/2019).

In den zurückliegenden Jahren wurden für die Komplexität literarischer Texte verschiedene Maße und Methoden vorgestellt; vgl. z. B. Gries (2016), Perkuhn et al. (2012). Ein besonders für angewandte Disziplinen wie die Stilometrie interessanter Untersuchungsbereich betrifft Messungen zum Reichtum des Vokabulars (Yule 1944) bzw. der lexikalischen Vielfalt (Carroll 1938). Die Idee der Wortschatzvarianz geht dabei von der Annahme aus, dass gemessene Werte (Type-Token-Verhältnis als Quotient aus Type-Anzahl und Token-Anzahl) Indikatoren für den Wortschatzumfang eines Autors und mithin charakteristische Eigenschaften sind (Tanaka-Ishii/Aihara 2015). Ein methodisches Problem bleibt der Umstand, dass beinahe alle Ansätze (wie z. B. TTR, STTR) als Konsequenz des Zipf-Mandelbrot-Gesetzes (Mandelbrot 1953) abhängig von der Korpusgröße variieren (Tweedie/Baayen 1998, Evert et al. 2017). Die Online-Plattform des Songkorpus[1] bietet hierzu neben den Primärdaten verschiedene Maße und visualisierte Statistiken an.

Bild 1. Neologismen im Udo-Lindenberg-Archiv.

Zu den weiteren unmittelbar abfragbaren Daten zählen Frequenzlisten (interessanterweise finden sich hier die Wörter „und“ und „ich“ auf den vordersten Rängen, dann erst gefolgt von Artikeln), Neologismen (vgl. Bild 1), die Überprüfung quantitativer Regularitäten wie dem Zipf'schen Gesetz oder der Korrelation zwischen Strophen- und Verszeilenzahl (vgl. Bild 2) sowie Kollokationsanalysen und n-Gramme (vgl. Bild 3). Außerdem werden Ortsbezeichnungen (Named Entitites) aus den Texten auf einer geografischen Karte verortet.

Bild 4 kontrastiert Worthäufigkeiten im Lindenberg-Archiv und in einem regional und zeitlich ausgewogenen allgemeinsprachlichen Korpus (zu dessen Stratifizierung vgl. Bubenhofer et al. 2013). Dabei gruppieren sich Wörter mit ähnlichen Frequenzen in beiden Sammlungen („akzeptieren“, „besonders“, „in“) nahe der zentralen Trennlinie, während spezifische Wörter (im Lindenberg-Archiv etwa „abgefuckt“, „Freund“, „Welt“) einen größeren Abstand aufweisen.

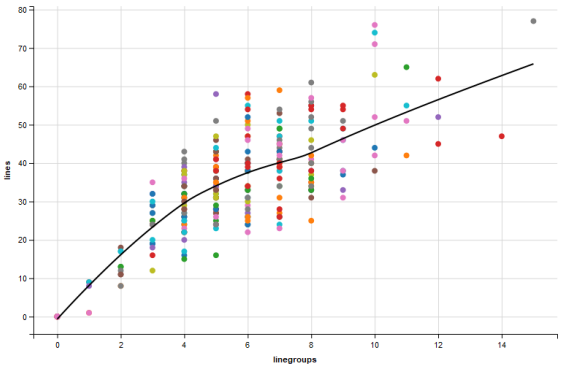

Bild 2. Strophen und Verszeilen ausgewählter Alben.

5 Fazit und Ausblick

Songtexte können als Textgattung betrachtet werden, die als eine Art "Vermündlichung des Lyrischen" Merkmale sowohl des geschriebenen als auch des gesprochenen Diskurses aufweist, sowie als Datenquelle im Kontinuum zwischen Standard und Nonstandard. Vielversprechend erscheinen gezielte Analysen sprachlicher Phänomene, die sich von Entsprechungen in anderen literarischen Schriften, Sach- und Gebrauchstexten oder spontan gesprochener Alltagssprache unterscheiden.

Das Songkorpus komplementiert den Kanon korpuslinguistischer Sammlungen um mehrfach annotierte deutschsprachige Songtexte, mit dem vorgestellten Udo-Lindenberg-Archiv sowie einem Chart-Song-Archiv als initialen Inhalten. Beide werden kontinuierlich aktualisiert und um weitere Archive ergänzt. Die TEI-annotierten Inhalte des Lindenberg-Archivs sind über das Online-Frontend recherchier- und einsehbar und lassen sich für die weiterführende wissenschaftliche Forschung gesammelt herunterladen. Ausgewählte korpuslinguistisch motivierte Auswertungen und Visualisierungen beider Archive können auf Zeichen-, Wort- und Versebene unmittelbar unter *http://songkorpus.de* berechnet werden.

Forschungsthemen, die durch die neue Ressource befördert werden, umfassen z.B.: (a) Topic Modeling, Identifizierung prominenter Themen für ausgewählte Zeiträume und Autoren (b) Parallelitäten zwischen Personen-, Orts- oder Institutionsbezeichnern und prominenten Themen im öffentlichen Diskurs (c) Sentiment Analysis zur Beschreibung von Emotionalität in Songtexten oder

[1] http://songkorpus.de unter dem Menüpunkt „Explorer“

Musikgenres (d) Einfluss sprachexterner Faktoren (z. B. individuelle Veröffentlichungsproduktivität) auf die lexikalische Vielfalt (e) Stilistische Analysen, Identifizierung von „style markers“ wie Verwendungshäufigkeit bestimmter Personalpronomen (f) Textähnlichkeitsmessungen (g) Reimformen und Reimschemata (h) Identifizierung autoren-/zeitspezifischer Formulierungsmuster und symbolischer Elemente/Metaphern (i) empirische Annäherungen an Phänomene wie Ironie und Wortwitz (j) Variationsstudien zu dialektalen Songtexten (k) Empirische Aussagen zur Standardkonformität und Verortung im Kontinuum zwischen Schrift- und Umgangssprache.

Bild 3. Prominente Bigramme ausgewählter Alben.

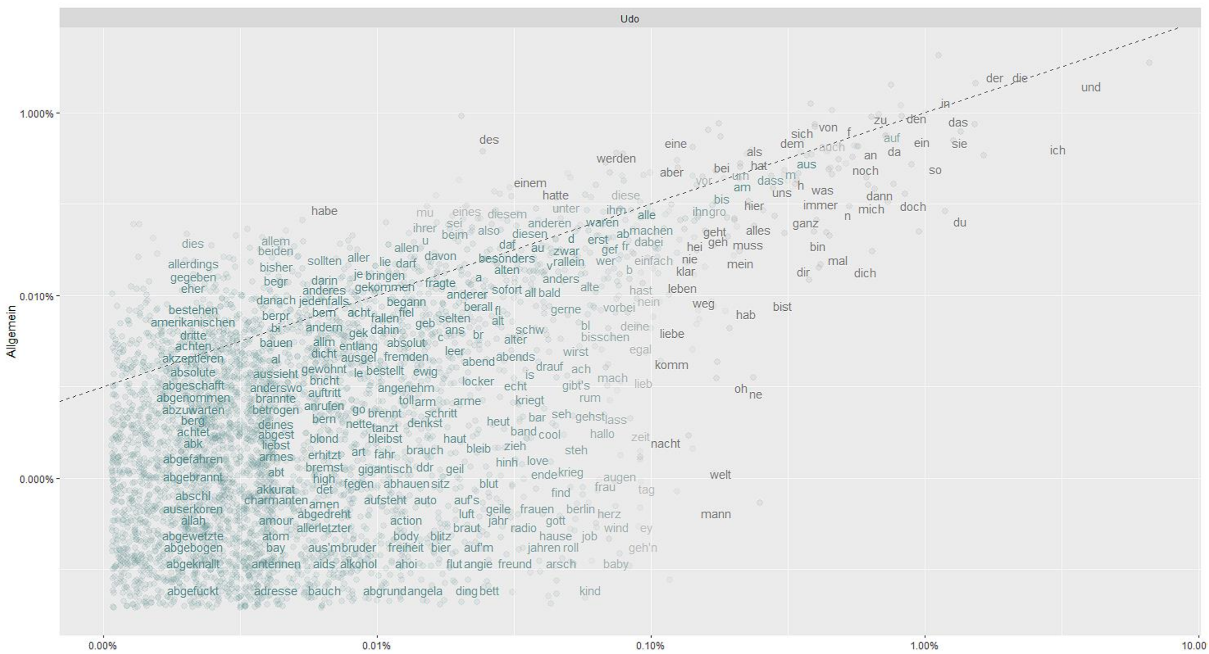

Bild 4. Wortfrequenzvergleich.

Das Songtextkorpus schließt damit eine Datenlücke, die bislang die empirisch fundierte Beantwortung syntaktischer, semantischer oder pragmatischer Fragestellungen für diese Textsorte erschwert. Die interdisziplinären Anknüpfungspunkte erscheinen vielfältig und vielversprechend: Neben Linguistik und Literaturwissenschaft lassen sich profitierende Forschungsbereiche im breiten Spektrum der Kulturwissenschaften sowie der Musik-, Medien- oder Geschichtswissenschaft verorten.

Literatur

Frieder von Ammon, Dirk von Petersdorff (Hg.). 2019. *Lyrik/ lyrics. Songtexte als Gegenstand der Literaturwissenschaft.* Wallstein Verlag, Göttingen.

Craig A. Anderson, Nicholas L. Carnagey, Janie Eubanks. 2003. *Exposure to violent media: The effects of songs with violent lyrics on aggressive thoughts and feelings*. In: Journal of Personality and Social Psychology, 84(5), 960–971.

Annette Blühdorn. 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.

Noah Bubenhofer, Marek Konopka, Roman Schneider. 2013. *Präliminarien einer Korpusgrammatik*. Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP) 4. Tübingen: Narr.

Thomas Bartz, Michael Beißwenger, Angelika Storrer. 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): 157–198.

Michael Beißwenger, Thomas Bartz, Angelika Storrer, Swantje Westpfahl. 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). http://sites.google.com/site/empirist2015/

Darina Benikova, Christian Biemann, Marc Reznicek. 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. http://www.lrec-conf.org/proceedings/lrec2014/pdf/276_Paper.pdf

Thierry Bertin-Mahieux, Daniel Ellis, Brian Whitman, Paul Lamere. 2011. *The Million Song Dataset*. In: Proceedings of the 12th International Society for Music Information Retrieval Conference.

Chris Biemann. 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. http://wortschatz.uni-leipzig.de/~cbiemann/pub/2007/biemannRandomText-HLTNAACL07main.pdf

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

Brian Cullen. 2009. *A Corpus Analysis of Pop Song Lyrics. New Directions*. Nagoya Institute of Technology.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Annette Frank, Chris Biemann. 2016. *A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures*. In: Proceedings of the LT4DH workshop at COLING 2016, Osaka. https://www.clarin-d.net/images/lt4dh/pdf/LT4DH11.pdf

Alexander Eiter. 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

Stefan Evert, Sebastian Wankerl, Elmar Nöth. 2017. *Reliable measures of syntactic and lexical complexity: The case of Iris Murdoch*. In: Proceedings of the Corpus Linguistics 2017 Conference, Birmingham, UK. http://purl.org/stefan.evert/PUB/EvertWankerlNoeth2017.pdf

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

Reinhard Flender, Hermann Rauhe. 1989. *Popmusik: Aspekte ihrer Geschichte, Funktionen, Wirkung und Ästhetik*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Stefan Th. Gries. 2016. *Quantitative Corpus Linguistics with R.* 2nd rev. & ext. Edition. London & New York: Routledge, Taylor & Francis Group.

Marie Hinrichs, Thomas Zastrow, Erhard Hinrichs. 2010. *WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure*. In: Proceedings of the Seventh conference on International Language Resources and Evaluation. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2010*),* Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/270_Paper.pdf

Andrea Horbach, Diana Steffen, Stefan Thater, Manfred Pinkal. 2014. *Improving the performance of standard part-of-speech taggers for computer-mediated communication.* In: Proceedings of KONVENS 2014, Hildesheim, Germany.

Mark L. Johnson, Steve Larson. 2003. *'Something in the Way She Moves': Metaphors of musical motion*. In: Metaphor and Symbol 18(2): 63–84

Natalie Karlova-Bourbonus, Holger Grumt Suárez, Henning Lobin. 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.

Noah Katznelson, Joseph Gelman, Katrin Lindblom, Marie Caput. 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.

Reinhard Köhler. 2005. *Korpuslinguistik. Zu wissenschaftstheoretischen Grundlagen und methodologischen Perspektiven*. In: LDV-Forum, Band 20/2: 1–16. https://jlcl.org/content/2-allissues/22-Heft2-2005/Reinhard_Koehler.pdf

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

Rolf Kreyer, Joybrato Mukherjee. 2007. *The Style of Pop Song Lyrics: A Corpus-linguistic Pilot Study*. In: Anglia - Zeitschrift für englische Philologie, Band 125, Heft 1: 31–58. DOI: 10.1515/ANGL.2007.31

Mark Kupietz, Thomas Schmidt. 2018. *Korpuslinguistik*. Germanistische Sprachwissenschaft um 2020. Band 5. Berlin: Walter de Gruyter.

Lothar Lemnitzer, Heike Zinsmeister. 2015. *Korpuslinguistik. Eine Einführung*. Tübingen: Narr.

Debbie Liske. 2018. *Lyric Analysis with NLP & Machine Learning with R*. DataCamp. https://www.datacamp.com/community/tutorials/R-nlp-machine-learning

Anke Lüdeling, Merja Kytö (Hgg.). 2008. *Corpus Linguistics. An International Handbook*. Handbücher zur Sprach- und Kommunikationswissenschaft 29 (1-2). Berlin: de Gruyter.

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

Jose Mahedero, Álvaro Martínez, Pedro Cano, Markus Koppenberger, Fabien Gouyon. 2005. *Natural language processing of lyrics*. In: Proceedings of the 13th annual ACM international conference on Multimedia (MULTIMEDIA '05). ACM, New York, NY: 475–478. DOI: https://doi.org/10.1145/1101149.1101255

Benoît Mandelbrot. 1953. *An information theory of the statistical structure of language*. In: W. Jackson (Hg.): Communication Theory. New York: Academic Press: 503–512.

Ulrich Miethaner. 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.

Massimiliano Morini. 2013. *Towards a musical stylistics: movement in Kate Bush's "Running up that Hill"*. In: Language and Literature 22 (4): 283–97.

Heiko Motschenbacher. 2016. *A corpus linguistic study of the situatedness of English pop song lyrics*. In: Corpora 11.1: 1–28

Tim Murphey. 1992. *The Discourse of Pop Songs*. In: TESOL Quarterly 26: 770–774.

Kathleen Napier, Lior Shamir. 2018. *Quantitative Sentiment Analysis of Lyrics in Popular Music.* In: Journal of Popular Music Studies, Vol. 30 No. 4, December 2018: 161–176. DOI: 10.1525/jpms.2018.300411

Yasunori Nishina. 2017. *A Study of Pop Songs based on the Billboard Corpus*. In: International Journal of Language and Linguistics 4 (2) 2017: 125–134.

Jerome Penaranda. 2006. *Text Mining von Songtexten*. Diplomarbeit. Technische Universität Wien.

Rainer Perkuhn, Holger Keibel, Marc Kupietz. 2012. *Korpuslinguistik*. Paderborn: Fink.

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

Ines Rehbein, Sören Schalowski, Heike Wiese. 2012. *Erweiterung des STTS für gesprochene Sprache*. STTS Workshop am IMS Stuttgart.

Anne Schiller, Simone Teufel, Christine Stöckert. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. University of Stuttgart: Institut für Maschinelle Sprachverarbeitung (IMS).

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

Roy Shuker. 1998. *Key Concepts in Popular Music*. London: Routledge.

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

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

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

Todd Terhune. 1997. *Pop Songs: Myths and Realities.* In: The English Connection 1 (1): 8–12.

Fiona J. Tweedie, Harald Baayen. 1998. *How variable may a constant be?* In: Computers and the Humanities 32: 323–352.

Thomas Van Hoey. 2016. *'Love love peace peace': a corpus study of the Eurovision Song Contest*. Graduate Institute of Linguistics, National Taiwan University.

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

Ayano Watanabe. 2018. *A Style of Song Lyrics: The Case of Really*. In: Zephyr (2018), 30: 12–27. https://doi.org/10.14989/233019

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

Valentin Werner, Maria Lehl. 2015. *Pop lyrics and language pedagogy: A corpus-linguistic approach.* In: Formato, F.; Hardie, A. (Hg.) (2015): Corpus Linguistics. Lancaster: UCREL: 341–343.

Swantje Westpfahl. 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): 1–10. http://www.aclweb.org/anthology/W14-4901

Swantje Westpfahl, Thomas Schmidt, Jasmin Jonietz, Anton Borlinghaus. 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. urn:nbn:de:bsz:mh39-60634

George Udny Yule. 1944. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge.

Heike Zinsmeister, Ulrich Heid, Kathrin Beck. 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. http://www.lrec-conf.org/proceedings/lrec2014/pdf/721_Paper .pdf