Zeitschrift Korpora Deutsch als Fremdsprache

Jahrgang 2, Nummer 2 (Dezember 2022)
ISSN 2749-4411

DAS *SONGKORPUS* – Perspektiven einer korpuslinguistischen Nutzung deutschsprachiger Popmusik für die Fremd- und Zweitsprachenvermittlung

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

Abstract

Vorgestellt wird das Korpus deutschsprachiger Songtexte als innovative Sprachdatenquelle für interdisziplinäre Untersuchungsszenarien und speziell für den Einsatz im Fremd- und Zweitsprachenunterricht. Die Ressource dokumentiert Eigenschaften konzeptioneller Schriftlichkeit und konzeptioneller Mündlichkeit und erlaubt empirisch begründete Analysen sprachlicher Phänomene bzw. Tendenzen in den Texten moderner Popmusik. Vorgestellt werden Design, Annotationen und Anwendungsbeispiele des in thematische und autorenspezifische Archive stratifizierten Korpus.

Keywords: Korpuslinguistik; Korpusressource; Online-Schnittstelle; Empirik; Songtexte; Oraliteralität

Abstract

We present the Corpus of German Song Lyrics as an innovative language resource for interdisciplinary research scenarios and especially for use in foreign and second language teaching. The resource documents characteristics of conceptual literacy and conceptual orality, and allows for empirical analyzes of linguistic phenomena and tendencies in modern pop lyrics. Design, annotations and application examples of the corpus, which is stratified into thematic and author-specific archives, are presented.

Keywords: corpus linguistics; corpus resource; online interface; empiricism; lyrics; oral literacy

1. Songtexte im Lehr- und Lernkontext

Moderner Fremd- und Zweitsprachenunterricht möchte kommunikative Qualifikationen vermitteln, mit denen sich die Lernenden sprachlich erfolgreich in einem breiten Spektrum authentischer Situationskontexte bewegen können. Entscheidend ist dabei nicht allein die Förderung schriftsprachlicher Kompetenzen, sondern gleichermaßen die Verbesserung produktiver und rezeptiver mündlich-akustischer Fertigkeiten, also des Hörverstehens und des Sprechens, in unterschiedlichsten Alltagsumständen. Realitätsnahen Lehr- und Übungsmaterialien kommt dabei eine zentrale Bedeutung zu. Es gilt zu vermeiden, dass im Unterricht vornehmlich mündlich angewendetes Schriftdeutsch vermittelt wird. Konzeptionelle Distanzsprache sollte nicht als prototypisch für Situationen konzeptioneller Nähe zum Einsatz kommen und Sprachgebrauchsmuster bzw. Wortschatz sollten nicht isoliert, sondern in realen Kontexten gelernt werden. Andernfalls besteht die Gefahr, dass Sprachlernende genauso sprechen wie sie schreiben.

Ganz grundsätzlich helfen hier korpuslinguistische Ressourcen weiter: Sei es, dass datengestützte Befunde den Lehrenden auf relevante Phänomene hinweisen, die seine intuitive Sprachkompetenz erweitern; sei es, dass Lernende mit authentischem Sprachmaterial für konzeptionell variablen Sprachgebrauch versorgt werden können. Besonders hilfreich sind vor diesem Hintergrund naheliegenderweise Inhalte, die sich nicht allein aus den in etablierten Korpussammlungen prominenten Textsorten wie Presse, Gebrauchsliteratur und Belletristik speisen.

Genau an dieser Stelle kommt das *Songkorpus* (vgl. Schneider 2019, 2020) als innovative Datenquelle ins Spiel. Ebenso wie sich Popmusik von einem jugendkulturellen Phänomen zu einem

festen Bestandteil der Alltagskultur entwickelt hat, sind deren Texte mittlerweile allgegenwärtig. Als Ausdruck gelebter Sprache haben sie ein komplexes Binnenleben, sind syntaktisch und lexikalisch hochproduktiv und illustrieren beispielsweise hervorragend authentische Verwendungen von Kollokationen und speziell idiomatischen Redeweisen (vgl. Amin et al. 2021). Nachweisbar sind sie gekennzeichnet durch Merkmale sowohl schriftlicher als auch mündlicher Diskurse (Oraliteralität als schriftliche Fixierung konzeptioneller Mündlichkeit, z.B. durch eine sprachökonomisch angepasste Morphosyntax oder die Iteration von Buchstaben zur Emulation von Prosodie, vgl. Broll / Schneider 2023) sowie durch thematische Bezüge zu gesellschaftlichen Diskussionen (Schneider / Hansen / Lang 2022). Bei aller Mainstream-Nähe bleibt eine Affinität zu Jugendsprache beobachtbar, insbesondere in Subgenres wie Deutschrap/Hiphop. Ihr Einsatz im Sprachunterricht kann damit *Language Awareness* und intrinsische Motivation befördern (vgl. Werner 2020). Nicht zuletzt deshalb nutzen Institutionen wie das Goethe-Institut deutschsprachige Songtexte bereits seit vielen Jahren für Arbeitsblätter auf dem Niveau A2 (Fortgeschrittene Anfänger) und verbinden diese mit Übungsvorschlägen und methodisch-didaktischen Tipps. Das kognitionspsychologische Potenzial sowie der didaktische Einsatz von Musik(-texten) als abwechslungsreicher Impulsgeber im Sprachunterricht sind also nichts Neues (vgl. Allmayer 2009); innovativ wird der korpuslinguistische Integrationsansatz durch die erst damit empirisch fundierte Aufdeckung und Einbeziehung von Sprachmitteln abseits typischer Schrift- bzw. Distanzsprache.

2. Stratifikation, Umfang und Annotation des Korpus

Das *Songkorpus* fächert sich in themen- und autorenspezifische Archive auf. Es enthält zum einen fortlaufend die erfolgreichsten deutschsprachigen Hitparaden-Songs ab 1970; Chartplatzierungen dienen dabei analog zu Auflagenzahlen bei Zeitungen oder Bestsellerlisten in der Belletristik als Kriterium der Wirkmächtigkeit. Weiterhin dazu gehören Archive speziell für ostdeutsche Lieder (*DDR-Archiv* 1970-1990), die *Neue Deutsche Welle* (1978-1985) sowie Hiphop (zeitlich eingegrenzt auf die Jahre 2000-2020). Und schließlich umfasst das Korpus Künstler-Archive mit den kompletten Werken ausgewählter Sänger und Bands. Inhalte der letztgenannten Kategorie werden unter Abschluss von Nutzungsvereinbarungen mit den Rechteinhabern kuratiert und dürfen auf dieser Basis umfassend annotiert für die wissenschaftliche Nutzung zur Verfügung gestellt werden. Insgesamt deckt die kontinuierlich anwachsende Datenbasis mit derzeit ca. 8.000 Songtexten bzw. über 2 Millionen Worttokens mehr als ein halbes Jahrhundert populäre deutschsprachige Musik ab und unterstützt damit nachhaltig synchrone und diachrone Perspektiven (vgl. Schneider 2022).

Sämtliche Songtexte sind in einem einheitlichen TEI P5-kompatiblen XML-Format kodiert, das neben dem Primärtext – segmentiert in Strophen und Verszeilen – Meta-Angaben zu Autor/Künstler, Erscheinungsjahr, Album etc. integriert. Die o.g. Künstler-Archive weisen in vielen Fällen explizite Korrektur-Annotationen auf, d.h. mit *add*-/*delete*-/*update*-Tags gekennzeichnete Postkorrekturen oder hinzugefügte Interpunktionszeichen. Tokenbezogen annotiert sind zudem folgende linguistisch motivierte Informationstypen:

1. Grundform (Lemma)
2. Wortklasse (unter Nutzung eines erweiterten STTS-Tagsets)
3. *Named Entities* (Orte, Personen, Organisationen u.Ä.)
4. Syntaktische Konstituenten
5. Neologismen und Okkasionalismen
6. Reimstrukturen (nur in ausgewählten Texten)

Die Annotationslayer 1 bis 4 werden für alle Archive in einem ersten Arbeitsschritt automatisiert unter Nutzung des CLARIN-Dienstes WebLicht[1] angelegt und für die Künstler-Archive anschließend manuell kontrolliert. Die Annotationslayer 5 und 6 werden ausschließlich für Künstler-Archive manuell ergänzt. Sämtliche Annotationslayer liegen in standardisierten XML-Containerformaten vor, wahlweise im WebLicht Text Corpus Format (TCF) oder als Apache UIMA CAS XMI, aus denen sich bei Bedarf z.B. beliebig separierte CoNLL-Formate generieren lassen.

3. Nutzung des Korpus

Im *Songkorpus* kann über das öffentliche Internetportal songkorpus.de online recherchiert werden; außerdem stehen unter der Webadresse songkorpus.de/data/ mehrere für lokale Auswertungen herunterladbare Datensets zur Verfügung:

1. Komplettes Korpus als *Bag-Of-Words* (Tabulator-separierte Token- bzw. Lemmalisten mit pro Jahr berechneten Frequenzangaben)
2. N-Gramme (alle Bi-, Tri-, Tetra-, Penta- und Hexagramme mit korpusbasierten Frequenz-, Assoziations- und Kontextmaßen)
3. Wortvektoren (mehrdimensionale GloVe-Vektordarstellungen aller Korpuswörter)
4. Verschiedene taskspezifische Datensets (z.B. idiomatische Analysen, Auswertungen zu gesellschaftspolitisch relevantem Vokabular)

Weitere Inhalte und Formate sind für akademische Zwecke auf Anfrage erhältlich.

Unmittelbar online können im sogenannten *Explorer* Live-Berechnungen und -Visualisierungen durchgeführt werden, und zwar auf Buchstaben-, Wort-, Vers- oder Songlevel. Das Spektrum reicht von der Auswertung auffälliger Lexik (z.B. Komposita, Neologismen, Palindrome, *Strechwords*) über Verteilungsanalysen (z.B. Vokale, Wortarten, *Named Entities*) und Syntaxbäume bis hin zur statistischen Verifizierung quantitativer Gesetze (z.B. Zipf, Menzerath-Altmann); vgl. exemplarisch Abbildung 1.

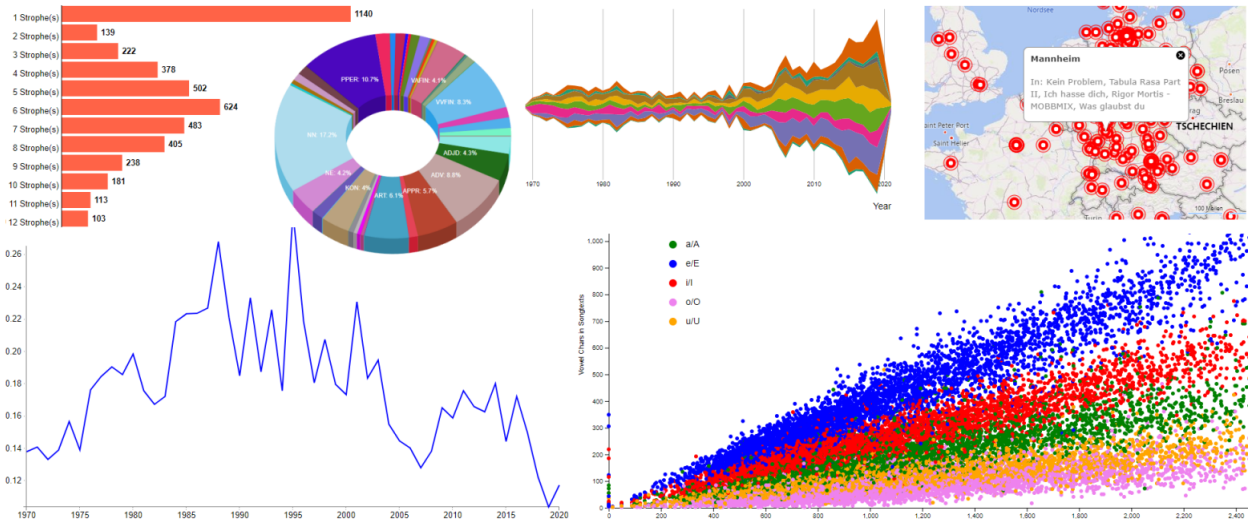

Abbildung 1
Empirische Visualisierungen im Songkorpus-Explorer

[1] Vgl. weblicht.sfs.uni-tuebingen.de (28.10.2022).

Parallel dazu lassen sich über die Online-Rechercheschnittstelle kombinierte Anordnungen von Wortformen, Lemmata und Wortart-Annotationen incl. Spezifizierung von Wortabständen abfragen. Die Schnittstelle lässt sich formularorientiert bedienen, das Erlernen einer speziellen Abfragesprache ist nicht erforderlich. Abbildung 2 demonstriert eine Belegsuche nach Verbzweitstellungen nach subordinierendem Konnektor *weil*, also nach einem besonders in gesprochener Sprache prominenten Phänomen. Dynamisch eingeblendete Tooltipps erleichtern hier die Auswahl von STTS-Wortartenbezeichnern. Fundstellen werden im KWIC-Format (*Keyword In Context*) präsentiert. Im vorliegenden Beispiel illustrieren die Referenzbelege einen umgangssprachlichen Duktus, teilweise auch eine jugendsprachliche Verortung, und lassen sich entweder auszugsweise in Lehrmaterialien übernehmen oder kombiniert als Datenbasis für weitergehende Erforschungen des Phänomens einsetzen.

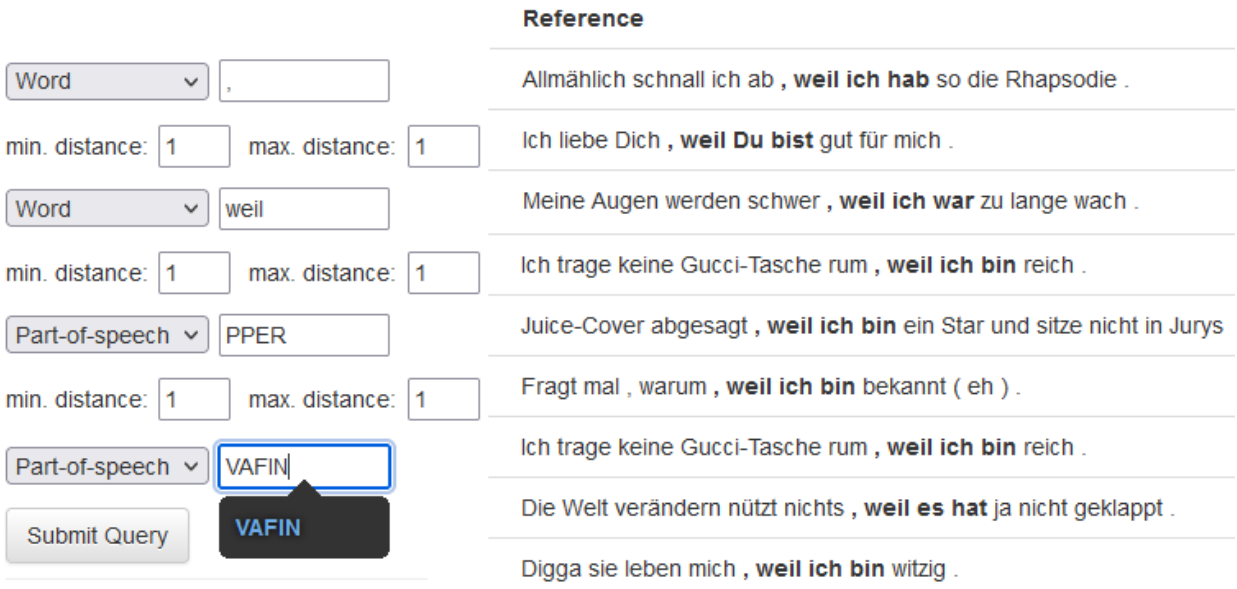

Abbildung 2
Online-Suchformular und Ergebnispräsentation (Auszug)

Literatur und Ressourcen

Allmayer, Sandra (2009): *Grammatikvermittlung mit Popsongs im Fremdsprachenunterricht*. Saarbrücken: Südwestdeutscher Verlag für Hochschulschriften.

Amin, Miriam / Fankhauser, Peter / Kupietz, Marc / Schneider, Roman / (2021): Data-driven Identification of Idioms in Song Lyrics. In: Cook, Paul / Mitrović, Jelena / Parra Escartín, Carla / Vaidya, Ashwini/ Osenova, Petya / Taslimipoor, Shiva / Ramisch, Carlos (eds.): *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*. Stroudsburg: Association for Computational Linguistics, 13-22.

Broll, Sarah / Schneider, Roman (erscheint 2023): Empirische Verortung konzeptioneller Mündlichkeit inner- und außerhalb schriftsprachlicher Korpora. In: *Journal for Language Technology and Computational Linguistics (JLCL)*.

Schneider, Roman / Hansen, Sandra / Lang, Christian (2022): Das Vokabular von Songtexten im gesellschaftlichen Kontext – ein diachron-empirischer Beitrag. In: Kämper, Heidrun / Plewnia, Albrecht (Hrsg.): *Sprache in Politik und Gesellschaft: Perspektiven und Zugänge*. Berlin, Boston: de Gruyter, 295-304.

Schneider, Roman (2022): Zwischen Schriftlichkeit und Mündlichkeit: Songtexte in der deskriptiven Sprachforschung. In: *Sprachreport* 1/2022, 38-50.

Schneider, Roman (2020): A Corpus Linguistic Perspective on Contemporary German Pop Lyrics with the Multi-Layer Annotated "Songkorpus". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*. Marseille: European Language Resources Association (ELRA), 835-841.

Schneider, Roman (2019): "Konservenglück in Tiefkühl-Town" - Das Songkorpus als empirische Ressource interdisziplinärer Erforschung deutschsprachiger Poptexte. In: *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*. Erlangen: German Society for Computational Linguistics & Language Technology (GSCL), 229-236.

Werner, Valentin (2020): Lyrics and Language Awareness. In: *Nordic Journal of Modern Language Methodology*. 7.1. DOI: 10.46364/njmlm.v7i1.521.

Biographische Notiz: PD Dr habil. Roman Schneider ist Leiter des Programmbereichs Sprachinformationssysteme am Leibniz-Institut für Deutsche Sprache (IDS) in Mannheim und beschäftigt sich dort mit der Reflexion und Anwendung empirischer Methoden für die Grammatik- und Orthografieforschung sowie mit der Verschränkung aktueller sprachwissenschaftlicher Erkenntnisse mit zielgruppenspezifischen Digitalformaten.

Kontaktanschrift:
PD Dr. Roman Schneider
Leibniz-Institut für Deutsche Sprache
R 5 6-13
68161 Mannheim
schneider@ids-mannheim.de