Volume 36 — Number 1 — 2023 — ISSN 2190-6858

JLCL
Journal for Language Technology
and Computational Linguistics

Special Issue on Challenges in Computational Linguistics, Empiric Research & Multidisciplinary Potential of German Song Lyrics

Edited by
Roman Schneider and Gertrud Faaß

GSCL
Gesellschaft für Sprachtechnologie & Computerlinguistik

Imprint

Editor
Christian Wartena
Publication supported by the Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)

Board of Directors
Committee and Advisory Board of the GSCL

Current issue
Volume 36 – 2023 – Issue 1

Guest Editors
Roman Schneider and Gertrud Faaß

Address
Christian Wartena
Hochschule Hannover
Expo Plaza 2
D-30539 Hannover
info@jlcl.org

ISSN
2190-6858

Publication
Mostly 2 issues per annum
Publication only electronically on jlcl.org

Contents

Roman Schneider / Gertrud Faaß

Editorial

Popmusik hat sich – im deutschsprachigen Raum ebenso wie weltweit – im Verlauf der zurückliegenden Jahrzehnte von einem ursprünglich jugendkulturellen Phänomen zu einem konstitutiven Bestandteil moderner Kultur und Sprachrealität entwickelt. Ihre vielschichtigen Ausprägungen von Easy Listening über Politsongs, Punk, Rock bis Hip-Hop (um nur einige zu nennen) umgeben uns in den unterschiedlichsten Situationen: beim Radiohören, über Audio-Streaming-Dienste, in Fernsehshows, beim Einkaufen im Supermarkt oder beim Sport. Songs dienen häufig nicht allein der Zerstreuung, sondern vermitteln Botschaften und Gefühle, geben Inspiration oder Orientierung. Auch abseits des kommerziellen "Mainstreams" haben sich ganz unterschiedliche Nischengenres etabliert, die Lebensgefühle und Befindlichkeiten einer pluralistischen Gesellschaft repräsentieren.

Als entsprechend aufschlussreich und wirkmächtig lassen sich die textuellen Inhalte von Songs ansehen, exemplarisch dokumentiert durch Forschungen zu Wechselwirkungen mit Jugendsprache. Songtexte sind sprachwissenschaftlich hochinteressant: Sie kombinieren verschiedene Stile und Register, weisen Merkmale sowohl des schriftlichen als auch des gesprochenen Diskurses auf und können als Repräsentation sprachlicher Vielfalt und Experimentierfreude im Kontinuum zwischen Standard und Nicht-Standard betrachtet werden.

Mit dem öffentlich zugänglichen Songtextkorpus (https://songkorpus.de, vgl. Schneider 2020 sowie Schneider 2022) liegt erstmals eine nachhaltige und breit stratifizierte Datenbasis für die empirische Exploration dieses in der germanistischen Linguistik und Computerlinguistik lange Zeit vergleichsweise wenig beachteten Forschungsgegenstands vor. Sämtliche Korpusinhalte sind unter Rückgriff auf etablierte texttechnologische Standards digitalisiert bzw. kodiert (UTF-8, TEI-P5), mehrfach annotiert (Lemma, POS, Named Entities, Neologismen) und in autorspezifische sowie thematische Archive unterteilt. Das Repository konstituiert damit eine Schatztruhe voll mit interessantem Vokabular, außergewöhnlicher Morphologie und teilweise überraschender Syntax.

Vor diesem Hintergrund versammelt das vorliegende Themenheft inter- und multidisziplinäre Beiträge zur quantitativen und qualitativen Analyse deutschsprachiger Songtexte:

Valentin Werner präsentiert unter der Überschrift *English and German pop song lyrics: Towards a contrastive textology* eine kontrastive korpusbasierte Analyse englischer und deutscher Popsongs. Er konzeptualisiert Songtexte als Textsorte/-register und identifiziert sprachübergreifende Gemeinsamkeiten und Unterschiede.

Jan Langenhorst, Yannick Frommherz und Simon Meier-Vieracker untersuchen im Beitrag *Keyness in song lyrics: Challenges of highly clumpy data* stilistische Phänomene und Keyness nicht nur auf Wortbasis, sondern unter Rückgriff auf Wortklassen-N-Gramme. Dabei spielen u.a. Streuungsphänomene eine Rolle und die Autoren decken auf, dass bei Songtexten „traditionelle“ Methoden der wortorientierten Stilanalyse zu kurz greifen.

Marco Gierke geht der Frage nach: *Ist alte Schule oldschool? Zum ‚Nutzen' von Anglizismen in Deutschraptexten*. Er vergleicht die anglizistische Nomination mit der nativen Entsprechung anhand eines adaptierten Analyse-Frameworks und betrachtet, empirisch unterstützt, den syntaktischen und morphologischen Gebrauch.

Katrin Hein analysiert *Beinahe-ums-Leben-kommen-in-Regenpfützen und Chauvi-Macho-Macker-Stuss – kreative Wortbildungen in Songtexten*. Dabei liegt der Fokus auf solchen Wortbildungen, die häufig nicht den Weg ins Lexikon finden, aber gerade aufgrund ihres okkasionellen Charakters einen erhöhten Grad an Expressivität aufweisen.

Elke Donalies betrachtet *Phraseme im Songkorpus: Etabliertes in Anti-Establishment-Texten* und zeigt anhand einer Stichprobe auf, wie verschiedene Autoren mit bekannten Wortkombinationen spielen, ihre Bedeutung hinterfragen und verändern. Angereichert wird der Überblick mit einem ausführlichen Anhang aufgefundener Phraseme.

Sarah Broll und Roman Schneider implementieren unter dem Titel *Empirische Verortung konzeptioneller Nähe/Mündlichkeit inner- und außerhalb schriftsprachlicher Korpora* ein automatisiertes Verfahren, das mithilfe unkorrelierter Entscheidungsbäume entsprechende Klassifikationen durchführt. Sie zeigen auf, dass Popsongs linguistisch motivierte Merkmale unterschiedlicher Kontinuumsstufen vereinen, und diskutieren überraschende Befunde.

Gertrud Faaß und Helmut Schmid systematisieren in ihrem Beitrag *Segmentierungs- und Annotationsverfahren für die Texte Udo Lindenbergs: Apostrophe und andere Herausforderungen* textsortenspezifische NLP-Fallstricke, erstellen einen Goldstandard, entwickeln ein maßgeschneidertes Segmentierungswerkzeug und trainieren mit hoher Zuverlässigkeit einen passenden POS-/Lemma-Tagger.

Akshay Mendhakar und Mesian Tilmatine evaluieren unter dem Titel *Automatic Authorship Classification for German Lyrics Using Naïve Bayes* automatisierte Klassifikationsverfahren zur Autorenbestimmung für Songtexte und gehen dabei auch auf methodische Grundlagen des verwendeten probabilistischen Frameworks ein.

Wir danken allen Beitragenden, den sachverständigen Gutachtern und Gutachterinnen, der GSCL und dem JLCL-Herausgeber für die wertvolle Unterstützung und wünschen unseren Leserinnen und Lesern viel Freude, Anregungen und produktive Anwendung neuer Erkenntnisse.

Roman Schneider & Gertrud Faaß

Literatur

Schneider, R. (2020). A Corpus Linguistic Perspective on Contemporary German Pop Lyrics with the Multi-Layer Annotated "Songkorpus". In N. Calzolari et al. (Hrsg.), Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC) (S. 842-848). Paris: European Language Resources Association. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-100347

Schneider, R. (2022). Zwischen Schriftlichkeit und Mündlichkeit: Songtexte in der deskriptiven Sprachforschung. Sprachreport, 38 (1), 38-50. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-109499.

Valentin Werner

English and German pop song lyrics: Towards a contrastive textology

Abstract

The present contribution offers a contrastive corpus-based analysis of English and German pop lyrics. It conceptualizes lyrics as a specific text type/register and tries to identify cross-linguistic commonalities and differences. As empirical base, it uses corpora that represent the lyrics of commercially highly successful pop songs in Anglophone and German contexts. Given the similar sociocultural functions and production circumstances of English and German lyrics, the study starts from the assumption that a large-scale linguistic overlap can be traced. While indeed cross-linguistic convergence is found especially for lexical patterns in terms of topic choice, the analysis also reveals a common property of conveying a conversational feel through lexicogrammatical means. However, given the differing typological make-up of the languages contrasted, fine-grained differences emerge as regards the ways conversationality/informality is established in pop lyrics as a performed text type.

Keywords: pop culture, pop cultural linguistics, performed language, media language, music, media linguistics, textlinguistics, register study, typology, LYPOP, Songkorpus

1 Introduction

Pop music is a genre that has taken a foothold in Germany at least since after the Second World War. From a diachronic perspective, it emerges that at first it was very much an imported format with strong roots in Anglophone societies such as the UK and especially the US. Therefore, at the outset German pop performers (or rather imitators) largely sang in English (Diederichsen, 2017), and, traditionally, the percentage of English-language lyrics has been high in the German pop charts, with estimations of more than 70% of all songs for the period 1965–2006, for instance (Achterberg et al., 2011). At the same time, German lyrics have been present in the German pop charts to a considerable degree and despite the global availability and spread of pop cultural artifacts, there is even evidence for an increasing divergence of pop music markets in more recent decades, fostering the recognition of domestic (i.e. German-language) music (Ferreira & Waldfogel, 2013; Bello & Garcia, 2021). Thus, as pop music certainly has developed into a global art form, the question arises whether non-Anglophone songs and lyrics may still show some orientation towards their historical "parent" in the sense of an American or at least Anglophone art form.

Even though such relationships could be explored from various vantage points (e.g. musicology, ethnology, etc.), it is suggested here that it may be worthwhile to apply a linguistic perspective, as lyrics are a central part of almost all successful pop songs. While lyrics traditionally had been sidelined as a subject of linguistic research, the amount and scope of (corpus-

based) work have been growing, especially in the past two decades, as illustrated in the following – by necessity selective – overview (see, e.g., Schneider, 2019, pp. 228–229 or Werner, 2021a, pp. 238–240 for more comprehensive recent summaries).

Besides work on song translation (e.g. Franzon et al., 2021) and early small-scale studies (such as Murphey, 1990), analyses such as Kreyer & Mukherjee (2007), Bértoli-Dutra (2014) or Werner (2012, 2021a) deserve mention in the area of English linguistics as they offer empirical assessments of lyrics as a register or genre, focusing on similarities to and differences of lyrics from other spoken and written text types and on (register-internal) dimensions of variation. Occasionally, regional differences between lyrics are highlighted (e.g., Werner, 2012). Brett & Pinna (2019) are to be credited for developing a contrastive analysis of different pop subgenres based on keywords, while there are also several studies with an explicit language-educational concern (e.g. Bertóli, 2018; Summer, 2018; Werner, 2019b, 2021b, 2021c).

By comparison, linguistic work on German pop lyrics has been scarce. Noteworthy exceptions to this are (i) the publications by Schneider and colleagues based on the *Songkorpus* (see Section 2), in which aspects such as the position of lyrics on a written-spoken/formal-informal cline (Schneider, 2022a; see also Broll & Schneider, this issue), the discursive representation of socially salient topics in lyrics (Schneider et al., 2022), or the occurrence of idioms (Amin et al., 2021) are treated, and (ii) sociolinguistic work on German rap (e.g. Androutsopoulos, 2003; Bohmann, 2010; Wiemeyer & Schaub, 2018), occasionally also involving a contrastive German-English perspective (Lüdtke, 2006).[1]

The foregoing review suggests that there has been increasing linguistic engagement with a focus on either English or German lyrics. However, what is largely lacking to date is a contrastive perspective that takes account of the fact that language usage is highly reflective of cross-cultural similarities and differences. To address this gap, the present study offers a contrastive view of German and English pop lyrics. More specifically, it conceptualizes lyrics as a specific text type/register and tries to identify commonalities and differences of this text type in both languages. It can broadly be situated in the contexts of cross-linguistic register studies (see, e.g., Neumann, 2013, 2016) and contrastive textology (see, e.g., Androutsopoulos, 1999; Androutsopoulos & Scholz, 2002) and employs corpora that represent the lyrics of commercially highly successful pop songs in German and Anglophone contexts (see further Section 2). Given the similar sociocultural functions and production circumstances of English and German lyrics and their historical "parent-child" relation (see above), the starting assumption would be large-scale linguistic overlap. However, due to the differing typological layout of the languages involved, there may also be differences at a micro-level, for instance pertaining to how conversationality/informality is realized in pop lyrics as a performed text type. To

[1] For the sake of completeness, note that corpus-based approaches to German lyrics have been used in other disciplines such as musicology and the psychology of music (see Ruth, 2019 for a pertinent example).

address the aforementioned aspects, the present study will tackle the following research questions:

- Can we find similarities and differences between English and German lyrics as regards topic choice and usage of other content words and phrases?
- How is conversationality/informality established in English and German lyrics, respectively?

The remainder of the present study is structured as follows: Section 2 introduces the notion of "contrastive textology" and thus situates the present work as a cross-linguistic register analysis. It further presents the English and German lyrics corpora used. Section 3 offers a situational analysis of lyrics discourse to establish basic principles regarding the communicative context surrounding this text type. While the preceding sections serve to pave the way for the analysis, Section 4 contains the main results pertaining to lexical and phrasal (Section 4.1) and lexicogrammatical (Section 4.2) aspects, establishing areas of cross-linguistic convergence and divergence. The concluding Section 5 serves to contextualize the overall findings and identifies areas for future research.

2 Approach and data

As mentioned in Section 1, the present study can be viewed as an instantiation of cross-linguistic (English-German) register analysis as conceptualized by Neumann (2016).[2] Specifically, it could be categorized as Neumann's type 3, that is, as a study that "takes register as the main object of research […] with features [as] indicators used to characterize the contrastive registers" (Neumann, 2016, p. 43). Such an approach facilitates the establishment of similarities and differences of the register/text type that is contrastively compared. While she acknowledges that there is no "optimal solution" (Neumann, 2016, p. 43) when it comes to the cross-linguistic comparability of registers/text types, it is assumed that both English and German lyrics are sufficiently similar in terms of their genesis and usage contexts (see Section 3) to permit a contrastive view.

[2] To avoid any potential terminological confusion, note that Neumann's approach is different from multi-dimensional (register) analysis (MDA) as established in Biber (1988, 1989) and the associated cross-linguistic study of universals of register variation (see Biber, 2014). Neumann (2016, p. 42) submits that an MDA approach is too unspecific as it is "only suited for general claims about the range of register variation in contrasted languages [but] does not allow individual comparisons of contrastive register pairs". Indeed, Biber (2014) succeeds in establishing similar dimensions of variation (such as clausal/oral vs. phrasal/literate or narrative vs. non-narrative discourse) across several languages but does not offer contrastive analyses of specific registers. Note also that unfortunately no MDA implementation is available for German yet. For examples of (non-contrastive) MDAs of English lyrics, see Bertóli (2018) or Werner (2021a).

In this regard, it is worth noting that the current study could also be seen as a contribution to what has been termed "contrastive textology" (Androutsopoulos, 1999, p. 256) or "contrastive genre analysis" (Androutsopoulos & Scholz, 2002, p. 3). The basic conjecture here is that "processes of cultural evolution can be described as discourse processes, in the sense that they are manifested in particular discourse practices and objectified in specific text types (genres)" (Androutsopoulos & Scholz, 2002, p. 3). The main implications that follow from this perspective are (i) that relevant genres (such as pop lyrics) share properties cross-linguistically, while (ii) there may also be differences between individual speech communities (English vs. German pop performers), which altogether motivates a contrastive analysis. Importantly for the current purposes, Androutsopoulos (1999, pp. 237–238) further implies that genre convergence is prominent in text types connected to everyday culture and especially in (mass-)mediatized texts that are potentially spread globally. Therefore, he argues for the explicit study of such texts to overcome the bias toward exploring "high registers" (e.g. academic writing, press texts) in order to determine the actual presence of cross-linguistic similarities and differences (e.g. also in terms of the usage of nonstandard features) in genres subject to similar communicative conditions. Pop music lyrics clearly qualify as a relevant text type in this regard (see Section 3).

A contrastive analysis of pop lyrics along the lines theorized in the foregoing passages requires representative corpus data. As lyrics, despite their potentially extensive social impact (see, e.g., Kreyer & Mukherjee, 2007; Bell & Gibson, 2011), commonly are not available as parts of larger monitor corpora, their analysis has to rely on specially compiled corpora. In that regard, a fundamental question that may arise is that of how to define the "pop" in pop lyrics, also with a view to ensure cross-linguistic comparability as discussed above. The approach taken here follows precedence set in previous studies (such as Kreyer & Mukherjee, 2007; Werner, 2012, 2021a; see also Werner, 2018, 2022) and relies on commercially successful material that apparently possesses high appeal to a large audience. At the same time, it is worth noting that "pop" thus defined comprises a multitude of musical (and not necessarily textual) subgenres, which commonly are subjectively defined across various chart platforms, for instance.[3]

The aforementioned commercial appeal can be operationalized through high-ranking chart positions, as is the case for the two corpora used for the present study. For English pop lyrics, it relies on an extended version of LYPOP, a corpus that has been used in several previous analyses with descriptive and applied foci (Werner 2019a, 2019b, 2021a, 2021b, 2021c). The current version of LYPOP contains all the lyrics from the top ten albums of the year-end charts

[3] Van Venrooj & Schmutz (2018) provide an insightful cultural perspective on the inherently fuzzy and culturally determined nature of musical genre boundaries, while Brett & Pinna (2019) provide evidence on the severe constraints operating when trying to assign musical genre designations based on linguistic information from the lyrics.

from the period 2001 to 2021 as determined by the *Official Charts Company* (www.official-charts.com), a British music-industry related organization that compiles chart synopses based on the collection and analysis of record sales and information from streaming services. This corpus comprises 2,387 lyrics (745,287 word tokens/20,330 word types).[4]

For German lyrics, the analysis relies on the "Chart Songs" (CS) section of the *Songkorpus* (www.songkorpus.de). As already introduced above (see Section 1), the *Songkorpus* has served as empirical basis for several studies (Amin et al., 2021; Schneider 2019, 2020, 2022a; Schneider et al. 2022) and is exceptional in that its data can be searched through an online interface. The CS section comprises the lyrics of successful German-language songs from the German top 100 single charts from the period 1970 to 2022 as determined by *Chartsurfer* (www.chartsurfer.de), a website dedicated to providing various chart synopses. The CS section totals 1,962 lyrics (588,081 word tokens/30,005 word types).[5]

Despite minor differences in terms of compilation principles and temporal scope, it is evident that LYPOP and CS may serve as an adequate pair of resources for the contrastive analysis intended as they both represent lyrics of commercially successful pop songs in the respective languages and are of a comparable size.[6]

3 Situational analysis

This section presents a succint overview of the situational characteristics of lyrics with a dual aim: first, to define the properties of the text type under scrutiny and to acknowledge the special character of lyrics as a performed, written-to-be-sung text type (see Kreyer & Mukherjee, 2007; Werner, 2012, 2021d) that has regularly been ignored in linguistic analysis compared to other registers; and second, to provide some more background on the issue of the comparability of English and German pop lyrics, as it is suggested here that they converge in terms of their genesis and usage contexts. The overview is loosely based on previous descriptions such as Biber & Egbert (2018, pp. 178–179) and Werner (2021a, pp. 245–248), who

[4] Word count as determined by *AntConc*. For detailed information on pre- and post-processing the data, please refer to Werner (2021a, p. 241).
[5] Word count as determined by *AntConc*. For detailed information on pre- and post-processing the data, please refer to Schneider (2022a, pp. 44–45).
[6] A first minor contrastive finding that emerges from a look at the numbers is that the English lyrics apparently are more repetitive (type/token ratio of 0.03) than the German ones (type/token ratio of 0.05). As the type/token ratio may vary as a function of text length, it should be clear that it is a crude measure if we wanted to make a statement on lexical density or diversity. While a diachronic analysis could provide valuable insights into the development of lexical diversity of lyrics over time (potentially also using alternative measurements, see, e.g., McCarthy & Jarvis, 2010; Kyle et al., 2021), this is outside of the scope of the present contribution (but see, e.g., Walker, 2016; Meindertsma, 2019).

Participants	Addresser/speaker	1. (Team of) lyricist(s) 2. Singer(s) or their persona(s) as animator (*sensu* Goffman, 1979) or authority of the performance (*sensu* Eckstein, 2010) as lyrical *I*
	Addressee	1. Narrow view/intended listener: A (fictional) single or plural unspecified *you* 2. Broader view: The pop music audience (potentially large); also as on-lookers (when consuming lyrics unintentionally)
Relation among participants	Interactiveness	Monologic, no backchanelling (unless live concert)
	Social roles	Hierarchical relationship (star system), high social distance between performer and audience, no personal relationship
	Shared knowledge	1. Some shared world knowledge between singer and audience 2. Potentially some "insider" knowledge between singer and audience
Channel	Mode of communication	Hybrid (written-to-be-sung)
	Permanence	Hybrid (transient in performance as sound waves; permanent in printed/electronic form)
Production/reception circumstances	Planning	Usually carefully planned (e.g. to fit the musical structure), revised and edited, unless spontaneous (e.g. in improvised battle rap)
	Reception	Real-time: Simultaneously heard and understood Printed/electronic text: Complete reader control
Setting	Private/public	Both possible
	Shared time and place of participants	Commonly spatial and temporal distance, unless live performance
Communicative purposes	General purpose	1. Broad view: Entertainment 2. Narrow view: Various purposes (narration, expression of attitudes, persuasion, etc.)
	Factuality	Mixed
	Stance expression	Overt expression of personal attitudes and epistemic stance
Topic		Variable (potentially 'love' as a salient subject)

Table 1: Situational properties of lyrics.

follow the general framework for the description of situational characteristics provided in Biber & Conrad (2019, pp. 39–40). Table 1 summarizes the central aspects, concurrently highlighting the complexity and hybridity of the communicative situation.

4 Results

4.1 Topics and n-grams

A first comparison relates to the topic choice of English vs. German lyrics. To this end, the top 15 content words as determined through the wordlist function of *AntConc* (Anthony, 2022) are contrasted, as presented in Table 2. As the focus here is on the popularity of topics as embodied by the ranking, absolute frequencies are presented only.

	LYPOP			CS		
Rank	**Item**	**Freq**	**Range**	**Item**	**Freq**	**Range**
1	*love*	3,263	657	*Liebe*	1,482	483
2	*baby*	2,610	631	*Leben*	1,540	595
3	*time*	2,371	935	*Nacht*	1,194	503
4	*way*	2,089	782	*Welt*	1,176	501
5	*heart*	1,704	679	*Zeit*	1,072	506
6	*thing(s)*	1,447	717	*Tag*	923	456
7	*life*	1,398	600	*Herz*	860	371
8	*girl*	1,298	398	*Baby*	853	211
9	*man*	1,240	424	*Mann*	645	316
10	*night*	1,195	492	*Augen*	565	305
11	*day*	1,023	476	*Sonne*	514	239
12	*world*	975	423	*Kopf*	438	221
13	*eyes*	909	490	*Glück*	432	250
14	*mind*	873	442	*Mädchen*	423	165
15	*head*	647	338	*Stadt*	379	193

Table 2: Top 15 content words (nouns); equivalents marked in same color.[7]

[7] A methodological note is in order here as regards the most frequent item *love*/*Liebe*. While it is easy to discern between verbal and nominal usages in German through a case-sensitive corpus query, the verb and the noun are exact homonyms in English. To approximate the numbers of verbs and nouns, a random sample of 500 occurrences of *love* was manually annotated to identify the percentage of verbal (44%) and nominal (56%) usages, and absolute numbers, displayed in in Tables 2 and 3, were calculated accordingly.

From the data in Table 2 it emerges that only two items (*love/Liebe*; *man/Mann*) cover the exact same ranks; yet, it is striking that 12 out of 15 items are contained in both the English and the German list. This suggests that the convergence in terms of genesis and usage contexts (see Section 3) results in large-scale overlap as to choice of words (and thus topic choice to a large degree) and that the situational properties apparently have a substantial cross-linguistic effect in this regard. As a side-note, it is worth mentioning that the German list with *Baby* features an Anglicism that also ranks highly in the English data. The occurrence of this word could be interpreted as one instance where an English pattern is adopted to fulfill a similar function as a vague address term to an unspecified fictional addressee (see Section 3), as shown in examples (1) and (2).

(1) I don't like nobody but you, **baby**, I don't care (Ed Sheeran: I don't care)
(2) Das wird unser Tag, **Baby**, wenn wir **aufsteh'n** (Seeed: Aufsteh'n)

	LYPOP			**CS**		
Rank	**Item**	**Freq**	**Range**	**Item**	**Freq**	**Range**
1	*get/ got*	6,492	1,886	*will/ willst*	2,012/ 521	660/ 239
2	*know*	5,330	1,459	*komm'/ kommt*	1,350/ 762	498/ 367
3	*like*	5,056	1,266	*geht/ geh'/ gehen*	1,292/ 1,083/ 616	549/ 409/ 308
4	*go*	2,958	969	*weiß/ weißt*	1,255/ 491	567/ 266
5	*love*	2,564	516	*lass'*	1,159	395
6	*say*	2,422	869	*sag'/ sagt/ sagen*	914/ 532/ 478	378/ 237/ 266
7	*want*	2,240	710	*mach'/ macht/ machen*	761/ 734/ 422	340/ 390/ 213
8	*see*	2,225	941	*seh'/ sehen*	721/ 398	372 /242
9	*let*	2,064	720	*gib'*	524	144
10	*make*	2,017	750	*glaub'*	440	230
11	*come*	1,993	724	*bleibt/ bleib'*	420/ 362	218/ 173
12	*take*	1,896	728	*meinen*	397	216
13	*need*	1,840	611	*steh'*	353	201
14	*feel*	1,838	712	*hör'*	392	188

15	*give*	1,332	498	*liebe/ lieb'*	389/ 353	123/ 124

Table 3: Top 15 content words (verbs); equivalents marked in same color; *be/have/do* and modal usages (e.g. *wanna*) excluded.

A similar picture as for nouns emerges from Table 3, which shows the most frequent content verb forms appearing in the two chart corpora. In this perspective, we find overlap for 10 out of 15 verbs, again with two items (*say/sagen; see/sehen*) covering exactly the same ranks. Table 3 also highlights the presence of clipped forms, specifically apocopes, in the German data. These high-frequency forms, which can be considered as markers of informal face-to-face communication (see Schneider, 2022a), outnumber the full realization in most instances (e.g. *komm', lass', glaub'*, etc.). Further, the German lyrics data yield an extended amount of syncopes (n = 5,717), as illustrated in (2) to (4).

(3) Sie wird heut' Nacht nicht **untergeh'n** und die Welt zählt laut bis zehn (Rammstein: Sonne)[8]
(4) Hallo Zeit, lang nicht mehr **geseh'n** und nein, ich will noch nicht nach Hause **geh'n** (Trettmann feat. Alli Neumann: Zeit steht)

As English lacks the opportunity for apocope and syncope in the one-syllable verbs that are pervasive in the lyrics, it may rely on other linguistic means to index informality/conversationality (see Section 4.2 and Werner, 2021a).

In the next step, the scope will be extended beyond the single-word unit to establish further linguistic features of pop lyrics from a contrastive perspective. To this end, frequent trigrams were extracted from LYPOP and CS. The results are shown in Table 4.

	LYPOP			**CS**		
Rank	**Type**	**Freq**	**Range**	**Type**	**Freq**	**Range**
1	*oh oh oh*	1,789	167	*la la la*	1,490	57
2	*yeah yeah yeah*	667	104	*na na na*	705	39
3	*la la la*	488	32	*oh oh oh*	277	43
4	*na na na*	472	18	*da da da*	250	12
5	*I don't know*	440	189	*le le le*	202	15
6	*I love you*	338	128	*ja ja ja*	182	34
7	*I don't wanna*	308	126	*wenn du mich*	170	55

[8] Note that the use of the syncope in this example is also due to rhyming/rhythm constraints.

8	*do do do*	303	14	*so wie du*	157	37
9	*ah ah ah*	272	25	*ich liebe dich*	154	55
10	*I know you*	262	120	*du bist mein*	122	42
11	*no no no*	262	55	*ich hab' dich*	118	61
12	*I know that*	252	137	*ich will dich*	118	45
13	*you and I*	244	109	*du mit mir*	116	26
14	*the way you*	243	87	*alles was ich*	115	57
15	*I know I*	235	101	*du bist so*	114	37
16	*I need you*	234	78	*ba ba ba*	108	4
17	*and I know*	223	111	*wo bist du*	107	29
18	*and I don't*	221	115	*ich bin der*	101	43
19	*da da da*	219	13	*ich bin ein*	101	40
20	*I want you*	214	73	*uh uh uh*	101	11
21	*you love me*	212	61	*du und ich*	100	57
22	*I want to*	208	95	*yeah yeah yeah*	100	18
23	*you make me*	199	50	*ich will nur*	98	28
24	*in love with*	198	63	*dass du mich*	97	37
25	*I don't want*	194	98	*tut mir leid*	97	45

Table 4: Top 25 trigrams; equivalents marked in same color.

A first striking result emerging from Table 4 is that both lists contain items that have been termed "non-lexical vocables" (Tegge & Parry, 2020, p. 7) or "musical tropes" (Werner, 2012, p. 25). These items do not carry any semantic meaning but are used (repetitively) for vocalization. Thus, they are present for aesthetic reasons and are indicative of the situated nature (language set to music) of lyrics (Werner, 2021a). Therefore, they could be categorized as register markers, defined as "distinctive linguistic constructions that do not occur in other registers" (Biber & Conrad, 2019, p. 54).

There is some overlap as regards the high-frequency items *la la la*, *na na na*, and *oh oh oh*, which possibly could qualify as universal pop lyrics vocables, and thus cross-linguistic register markers. Another potential candidate for a universal vocable is *yeah yeah yeah*. It is less salient in the German data (rank 2 in LYPOP; rank 22 in CS) but exemplifies another Anglicism. By contrast, the lyrics from both languages appear to have diverging inventories of ad-

ditional vocables given their different phonological systems. This is indicated in the transcriptions, with English furthermore featuring *do do do*, *ah ah ah*, and *no no no*, while the German list comprises *da da da*, *le le le*, *ja ja ja*, *ba ba ba*, and *uh uh uh*.

As regards other items in the trigram list, for English the present results tie in with Werner (2021a), who noted a high incidence of mental verbs used to express personal stance, a property typically assigned to conversation. In the German data, some of these also occur (e.g. *I love you/ich liebe dich*; *I want you/ich will dich*), and a high incidence of first and second person singular pronouns can be found for both languages (see also the overlapping combination *you and I/du und ich*), which can be viewed as an indication of informal/conversational usage (see further Section 4.2; Werner, 2012; Schneider, 2022a). In lyrics from both languages, the singer's persona (animator) as an *I/ich* as well as an intended listener as a (fictional) unspecified *you/du* are in focus (see Section 3). By contrast, in the German lyrics there also appear several high-frequency combinations involving the stative verb lemma *sein* ('be'), as in *du bist mein* ('you are my'), *du bist so* ('you are so'), *wo bist du* ('where are you'), *ich bin der* ('I am the'), and *ich bin ein* ('I am a').

4.2 Markers of informality and non-standard features

Previous studies such as Lüdtke (2006), Kreyer & Mukherjee (2007), Werner (2012, 2021a), and Schneider (2022a; see also Broll & Schneider, this issue) have highlighted the hybrid nature of pop (and rap) lyrics in terms of their position on the continua between informal conversation/speech and formal writing (language of immediacy vs. language of distance *sensu* Koch & Oesterreicher, 2012; see also Werner, 2021d). From a production perspective (see Section 3), it is relevant to explore strategies how lyrics (or rather lyricists) attempt to convey a "conversational feel", given (i) the actual production circumstances, (ii) audience expectations as regards the level of (in)formality of lyrics (see, e.g., Squires, 2019), and (iii) the differing structural layout of the two languages contrasted. To this end, it is considered useful to take a more qualitative perspective and to synthesize and assess the claims of previous studies as regards lexicogrammatical items with the help of the present data.

While the presence of clipped forms in German lyrics was already discussed in Section 4.1., related markers of informality also appear in English lyrics, albeit in different contexts. Such clipped forms comprise instances of *g*-dropping, as in (5), contracted modals, as in (6), (7), and (11), apheresis, as in (8) and (9), as well as contractions of the type noun/verb/pronoun + preposition, as in (9) to (11).

(5) Do you ever feel like **goin'** back? You know I spent some time in Hollywood **tryin'** to find **somethin'** (Lewis Capaldi: Hollywood)
(6) I don't ever **wanna** be like them (Stormzy: 21 gun interlude)[9]

[9] See also Table 4.

(7) I **gotta** be cool, relax, get hip and get on my tracks (Queen: Crazy little thing called love)
(8) **'cause** from here it looks the same (Ed Sheeran: Save myself)
(9) What I realised **'bout** who I am is that, you're **kinda** [*kind + of*] taught (Dave: Survivor's guilt)
(10) Stop **tryna** [*tryin(g) + to*] change me (Olly Murs: Stop tryna change me)
(11) I **gotta** [*got + to*] get **outta** [*out + of*] here (Sam Smith: Reminds me of you)

These examples illustrate the value of ellipsis for conveying informality in a genuinely scripted and edited text type, and some of them, such as *g*-dropping, have even been viewed as trademark features of lyrics (Lindsey, 2019).

For German pop lyrics, Schneider (2022a) has identified further contractions besides apocope (see Section 4.1 and (12)), which appear to be particularly salient in the domains of determiners, as in (12) and (13), forms of *sein* ('be'), as in (13) and (14), and possessive pronouns, as in (15) to (17). Also, contractions of infinitive forms (with *-en* in German; see also Section 4.1), as in (14) and (18) and of two words, regularly involving the third person neuter pronoun *es* ('it'), as in (18) and (19), appear.

(12) Denn ich **kauf'** mir **'ne** Villa in Berlin, danach **'ne** Villa in Paris (Katja Krasavice: Onlyfans)
(13) Sicher, Dicker, **is' 'n** Party-Track, leg den hier auf, **is'** deine Party weg (Das Bo: Türlich, türlich)
(14) Du hattest schlechte Zeiten und wir **war'n** auch dabei (Die fantastischen Vier: Troy)
(15) Sie schläft in **mei'm** Hoodie (Civo: Weg von mir)
(16) Da tippt der Kleine mich mit **sei'm** Leuchtfinger an (Willem: Wat)
(17) wir machen Rave in **dei'm** Schlafzimmer (Romero: Sie liebt Techno)
(18) Auch **wenn's** nicht so einfach war wie für dich, versteckt und verkrochen (Curse: Was ist jetzt)
(19) Und obwohl **du's** nicht zeigst, dass es dich grad zerreißt, ich **kann's seh'n**, kann dich seh'n (LEA: Wenn du mich lässt)

Other studies have drawn attention to non-standard features appearing in rap specifically. This is relevant as rap has developed into an important genre within pop music. For English, Werner (2019a) has highlighted the use of items associated with African American English. Such forms also appear in LYPOP, for instance the negator *ain't* and copula absence, illustrated in (20), completive *done*, as in (21), or invariant present tense forms, as in (22).

(20) A local hero, yo, but now he **ain't** unsung, Jimmy brought rum, he **ø** looking for clean cups (Faithless: Reasons)
(21) Asking how I **done** it, man, I did it by the grace (Stromzy: Pop boy)
(22) She **don't** wanna talk 'bout friendships (Dave: Heart attack)

Lüdtke (2006) has emphasized the conversational nature of German rap and lists features such as *am* + base form to express imperfectiveness, as in (23) and (24), *weil* + main sentence, as in (25), as well as dialectal forms, as illustrated in (26) to (28).

(23) Bratan, roll' den Saruch und ich bin **am schweben** (Capital Bra feat. Ufo361: Na na na)
(24) Bro, guck, ich bin **am ballen** (Luciano: La haine)

(25) Danke, mir geht's gut, **weil ich bin high** (Bonez MC: Shotz fired)
(26) Dat is Fettes Brot op Platt inne Disco (Fettes Brot: Nordisch by nature)
(27) Und **wat** da so bei 'rauskommt, ja, **dat** werden **wa ma** seh'n (Culcha Candela: Hamma)
(28) Zuzusehen, dass **net** andauernd Frauen bei dir stehen (Sabrina Setlur: Ich leb' für dich)

While the aforementioned features are present in rap lyrics that form part of CS, it is evident that they are comparatively rare (*dat* n = 77, *wat* n = 207, *net* n = 56), with the standard variants strongly preferred (*das* n = 6,155, *was* n = 3,993, *nicht* n = 7,423)[10] in German lyrics. Naturally, dialectal forms are pervasive in lyrics produced completely in regional dialect (e.g. by bands such as BAP or Spider Murphy Gang), in which they are used to style authenticity.[11] Other vernacular forms that Lüdtke (2006) lists, such as the German *Super(plusquam)perfekt* (e.g. *Ich war gestern dort gewesen*/*Er hat das gesagt gehabt)* and further contracted forms such as *brauchta* (*braucht* + *ihr*) and *kannste* (*kannst* + *du*; cf. *tryna* in example (10)) could not be found at all in the data. These apparently seem to be more characteristic of rap discourse than of pop lyrics discourse in general.[12]

The preceding overview has shown that a multitude of features associated with informal and non-standard usage appear in both English and German pop lyrics and it was suggested that these features are consciously used to convey a "conversational feel". At the same time, it is clear that lyrics largely lack other highly characteristic informal/conversational items, such as false starts or hesitation markers. Just as one case on point, note that the data hardly contain any instances of the latter (CS: *ähm* n = 10, *äh* n = 8; LYPOP: *uhm* n = 6, *uh* n = 43).[13] Given the scripted and edited production of the lyrics as well as the (as a rule) spatial and temporal distance between speaker and audience and the genuinely monologic/non-interactive nature of the discourse (see Section 3), such devices lack a communicative function and thus are absent. This could be related to the concept of the "performance filter" (Werner, 2021d, p. 568) in the sense that only selected items associated with conversationality (or the language of immediacy *sensu* Koch & Oesterreicher, 2012) are (consciously) used to index informality of lyrics discourse and that lyrics therefore are "not as conversational as conversation" (Werner, 2021, p. 256a). The present data suggest that this principle holds cross-linguistically.

[10] CS also contains 149 occurrences of the apocope form *nich'*.
[11] On the issue of language choice in German pop music, see Larkey (2000) and Diederichsen (2017).
[12] *Kannste* appears in the full version of the Songkorpus, though. Note also that related contracted forms that follow the pattern verb$_{\text{second person singular present tense}}$ + reduced *du*, such as *willste* (*willst* + *du*), *kommste* (*kommst* + *du*) or *darfste* (*darfst* + *du*) appear in CS.
[13] The majority of the instances of *uh* actually is used as a vocable (see Section 4.1). The mirative marker *uh oh* appears 15 times, the positive response marker *uh-huh* (and its spelling variants *a-ha* and *aha*) 128 times in LYPOP. Arguably, many of the occurrences of the latter also qualify as vocables.

5 Conclusion

The present contribution, which can be embedded into the larger context of contrastive textology/cross-linguistic register analysis (see Section 2) was based on the premise that English and German pop lyrics as one instantiation of a performed text type are subject to similar contextual constraints and share a similar sociocultural and communicative purpose (Section 3). Therefore, a starting assumption was large-scale linguistic convergence, which was tested using a corpus-based approach.

Overlap indeed was traceable in the corpus data with regard to (i) salient content topics of the lyrics as well as (ii) usage of content verbs (Section 4.1), which therefore could be considered "pop lyrics universals". Lyrics from both languages also showed considerable congruence in the domain of register markers, notably the presence of (repetitive) non-lexical vocables. While the inventory of high-frequency items (e.g. *na na na*) was found to be similar, due their different phonological systems data from both languages also featured variation for other items (Section 4.1). Another difference at the lexical/phrasal level that emerged from the analysis of high-frequency trigrams pertained to the choice of verbs (mental verbs in English vs. forms of *sein* in German lyrics).

Further, it was hypothesized that there may also be differences at a micro-level, for instance as to how conversationality/informality is realized through lexicogrammatical means. The analysis of markers of informality and non-standardness suggested that both English and German lyrics attempt to convey a "conversational feel" through employing selected sets of relevant features, illustrating the presence of a performance filter as theorized in earlier work. These sets, however, were found to be diverging due to the differing typological structure of English and German, and fine-grained differences, for instance as regards contraction patterns in lexical and modal verbs (Sections 4.1 and 4.2), could be identified.

While the present study has provided a first contrastive insight into cross-linguistic differences and similarities of pop lyrics as a register, the scope of the data would allow several additional routes to be pursued in the future. These comprise a corpus-based contrastive look at cognitive aspects such as conceptual metaphors regarding high-frequency items like *love/Liebe*, where a quick search of the present corpus data yields various patterns (e.g. *love is… a losing game/a losing hand/pain/like a ship/dream/grain of sand/rocket/shadow/brick/ the sea/rain/fate* vs. *Liebe ist… ein Fluch/ein Schild/eine Sucht*, etc.) that could be subject to further scrutiny as regards variation and linguistic creativity (see also Kreyer, 2012; Climent & Coll-Florit, 2021).

Another aspect ignored in the present analysis relates to the presence of Anglicisms in the lyrics. This may be especially worthwhile in view of the fact that German pop lyrics have been viewed as the "child" of an Anglo-American cultural tradition (see Section 1). Relevant examples are given as (29) to (32).

(29) Aha, und alle anderen **Girls** wären gern wie du (Cro: Traum)
(30) Jede Profilneurose bekam 'n **Deal**, und dazu noch 'n Starproduzenten (Absolute Beginner: Es war einmal)
(31) Ohren explodieren langsam, weil diese **Bitch-Niggas** mir viel zu viel reden. Zu viel reden. **Stage** am **shaken** wie Erdbeben. Früher viel **Shit** gemacht, Kugeln auf Boden, als gäb' es ein'n Regen (Pajel: 10 von 10)
(32) Schmeiß' die **Bitch** raus, wenn sie ihren **Mood changt**. Ja, dein **Outfit** kostet so viel wie mein **Shoelace** (Ufo361 & Bonez MC: 7)

Many of these examples are from rap and it could be further discussed whether they are instances of imitation of US rap as a "mother culture" (Androutsopoulos & Scholz, 2002) to create genre-appropriate authenticity. Notably, individual items have been adapted into the linguistic system of German, both as regards their capitalization as nouns (*Girls*, *Deal*, *Bitch-Niggas*, *Stage*, *Shit*, *Mood*, *Outfit*, *Shoelace*) and as regards their adaptation into German morphological patterns (*shaken*, *changt*), at least as indicated by the transcriptions.

On a related note, it may be worthwhile to explore the issues of multilingualism and language mixing/code-switching, as exemplified in (33) and (34).

(33) Heh, du willst **Dollar-Sign**? **Baby**, **valla** nein! (KC Rebell & Summer Cem: Valla nein)
(34) **Never, never go to work**, lieber plantschen und sich anzieh'n fein, **look the girls on the** Po by the tolle **sunshine** (Helge Schneider: Sommer Sonne Kaktus)

Example (33) is trilingual as it involves German, English and Turkish lyrics, arguably with the intent to create (rap) authenticity through establishing multiple cultural connections, crucially not restricted to an US-American one. Alternatively, such examples could be interpreted as exemplifying "glocal" forces in rap (see, e.g., Androutsopoulos, 2003; Barone, in press). The German-English example (34), by contrast, illustrates language mixing for humorous purposes.

Further extensions of the present work are conceivable in terms of (i) developing more detailed quantitative and qualitative contrastive analyses of additional devices associated with informal conversation and strategic use to convey conversationality in written texts, such as discourse markers (e.g. *you know/ich mein'*), for instance (Imo, 2017; Schourup, 1999), and (ii) considering aspects related to language education, for instance how lyrics as highly motivating input material could be exploited to introduce conversational grammar and to develop language awareness (see, e.g., Esa, 2008; Schneider, 2022b; Werner, 2019b, 2021c).

Still, the preceding list is not exhaustive, so there will be many additional options to engage with the language of lyrics from a contrastive and other linguistic perspectives, with the present study serving as a possible point of departure. The availability of relevant corpus material (also going beyond the language pair English-German) will be crucial for such endeavors. As a final general note, it is argued that the present study also has illustrated the potential of "pop

cultural linguistics" (Werner, 2018, 2022) as an emerging research subfield that takes socio-culturally relevant pop cultural artifacts seriously as an object of linguistic study.

6 References

Achterberg, P., Heilbron, J., Houtman, D., & Aupers, S. (2011). A cultural globalization of popular music? American, Dutch, French, and German popular music charts (1965 to 2006). *American Behavioral Scientist*, *55*(5), 589–608. https://doi.org/10.1177/0002764211398081

Amin, M., Fankhauser, P., Kupietz, M., & Schneider, R. (2021). Data-driven identification of idioms in song lyrics. In P. Cook, J. Mitrović, C. Parra Escartín, A. Vaidya, P. Osenova, S. Taslimipoor, & C. Ramisch (Eds.), *Proceedings of the 17th Workshop on Multiword Expressions* (pp. 13–22). ACL. https://aclanthology.org/2021.mwe-1.pdf

Androutsopoulos, J. (1999). Textsortenvergleich und Jugendkultur: Die "Plattenkritik" in deutschen und französischen Jugendmagazinen. In S. Reinart, & M. Schreiber (Eds.), *Sprachvergleich und Übersetzen: Französisch und Deutsch* (pp. 237–260). Romanistischer Verlag.

Androutsopoulos, J. (Ed.). (2003). *HipHop: Globale Kultur – lokale Praktiken*. Transcript.

Androutsopoulos, J., & Scholz, A. (2002). On the recontextualization of hip-hop in European speech communities: A contrastive analysis of rap lyrics. *Philologie im Netz*, *19*, 1–42. http://web.fu-berlin.de/phin/phin19/p19t1.htm

Anthony, L. (2022). *AntConc: A corpus analysis toolkit for researchers, teachers, and learners* (Version 4.0.10). Waseda University. https://www.laurenceanthony.net/software/antconc/

Barone, S. (in press). Under a groove: Rap, hip-hop and their glocalization. In A. Bennett (Ed.), *The Bloomsbury handbook of popular music and youth culture*. Bloomsbury.

Bell, A., & Gibson, A. (2011). Staging language: An introduction to the sociolinguistics of performance. *Journal of Sociolinguistics*, *15*(5), 555–572. https://doi.org/10.1111/j.1467-9841.2011.00517.x

Bello, P., & Garcia, D. (2021). Cultural divergence in popular music: The increasing diversity of music consumption on *Spotify* across countries. *Humanities and Social Sciences Communications*, *8*(1), 1–8. https://doi.org/10.1057/s41599-021-00855-1

Bértoli, P. (2018). Song lyrics: From multi-dimensional analysis to the foreign language classroom. In V. Werner (Ed.), *The language of pop culture* (pp. 210–229). Routledge. https://doi,org/10.4324/9781315168210-10

Bértoli-Dutra, P. (2014). Multi-dimensional analysis of pop songs. In T. Berber Sardinha, & M. Veirano Pinto (Eds.), *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber* (pp. 149–175). Benjamins. https://doi.org/10.1075/scl.60.05ber

Brett, D. & Pinna, A. (2019). Words (don't come easy): The automatic retrieval and analysis of popular song lyrics. In C. Suhr, T. Nevalainen, & I. Taavitsainen (Eds.), *From data to evidence in English language research* (pp. 307–325). Brill. https://doi.org/10.1163/9789004390652_014

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Biber, D. (1989). A typology of English texts. *Linguistics*, *27*(1), 3–43. https://doi.org/10.1515/ling.1989.27.1.3

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*, *14*(1), 7–34. https://doi.org/10.1075/lic.14.1.02bib

Biber, D., & Conrad, S. (2019). *Register, genre and style.* Cambridge University Press. https://doi.org/10.1017/9781108686136

Biber, D., & Egbert, J. (2018). *Register variation online.* Cambridge University Press. https://doi.org/10.1017/9781316388228

Bohmann, A. (2010). "Red mal deutsch, Hundesohn, ich halt nicht viel vom Spitten": Cultural pressures and the language of German hip hop. *Zeitschrift für Anglistik und Amerikanistik*, *58*(3), 203–228. https://doi.org/10.1515/zaa.2010.58.3.203

Climent, S., & Coll-Florit, M. (2021). All you need is love: Metaphors of love in 1946–2016 Billboard year-end number-one songs. *Text & Talk*, *41*(4), 469–491. https://doi.org/10.1515/text-2019-0209

Diederichsen, D. (2017). Singing in German: Pop music and the question of language. In M. Ahlers, & C. Jacke (Eds.), *Perspectives on German popular music* (pp. 190–194). Routledge.

Eckstein, L. (2010). *Reading song lyrics*. Rodopi.

Esa, M. (2008). Musik im Deutschunterricht: Der gezielte Einsatz. *Die Unterrichtspraxis/Teaching German*, *41*(1), 1–14. https://doi.org/10.1111/j.1756-1221.2008.00001.x

Ferreira, F., & Waldfogel, J. (2013). Pop internationalism: Has half a century of world music trade displaced local culture? *The Economic Journal*, *123*(569), 634–664. https://doi.org/10.1111/ecoj.12003

Franzon, J., Klungervik Greenall, A., Kvam, S., & Parianou, A. (Eds.). (2021). *Song translation: Lyrics in context*. Frank & Timme.

Goffman, E. (1979). Footing. *Semiotica*, *25*(1/2), 1–29. https://doi.org/10.1515/semi.1979.25.1-2.1

Imo, W. (2017). Diskursmarker im gesprochenen und geschriebenen Deutsch. In H. Blühdorn, A. Deppermann, H. Helmer, & T. Spranz-Fogasy (Eds.), *Diskursmarker im Deutschen: Reflexionen und Analysen* (pp. 49–72). Verlag für Gesprächsforschung.

Koch, P., & Oesterreicher, W. (2012). Language of immediacy – Language of distance: Orality and literacy from the perspective of language theory and linguistic history. In C. Lange, B. Weber, & G. Wolf (Eds.), *Communicative spaces: Variation, contact, and change* (pp. 441–473). Lang.

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 S. Hoffmann, P. Rayson, & G. Leech (Eds.), *English corpus linguistics: Looking back, moving forward* (pp. 103–115). Brill. https://doi.org/10.1163/9789401207478_008

Kreyer, R., & Mukherjee, J. (2007). The style of pop song lyrics: A corpuslinguistic pilot study. *Anglia*, *125*(1), 31–58.

Kyle, K., Crossley, C. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Larkey, E. (2000). Just for fun? Language choice in German popular music. *Popular Music and Society*, *24*(3), 1–20. https://doi.org/10.1080/03007760008591773

Lindsey, G. (2019). *English after RP: Standard British pronunciation today*. Palgrave Macmillan. https://doi.org/10.1007/978-3-030-04357-5

Lüdtke, S. (2006). *Globalisierung und Lokalisierung von Rapmusik am Beispiel amerikanischer und deutscher Raptexte* [Doctoral dissertation, Gottfried Wilhelm Leibniz Universität]. Hannover. https://www.repo.uni-hannover.de/bitstream/handle/123456789/6795/517913186.pdf

Meindertsma, P. (2019). Changes in lyrical and hit diversity of popular U.S. songs 1956–2016. *Digital Humanities Quarterly*, *13*(4). http://www.digitalhumanities.org/dhq/vol/13/4/000440/000440.html

Murphey. T. (1990). *Song and music in language learning: An analysis of pop song lyrics and the use of song and music in teaching English to speakers of other languages*. Lang.

Neumann, S. (2013). *Contrastive register variation: A quantitative approach to the comparison of English and German*. Mouton de Gruyter. https://doi.org/10.1515/9783110238594

Neumann, S. (2016). Cross-linguistic register studies: Theoretical and methodological considerations. In M.-A. Lefer, & S. Vogeleer (Eds.), *Genre- and register-related discourse features in contrast* (pp. 35–57). Benjamins. https://doi.org/10.1075/bct.87.03neu

Ruth, N. (2019). "Where is the love?" Topics and prosocial behavior in German popular music lyrics from 1954 to 2014. *Musicae Scientiae*, *23*(4), 508–524. https://doi.org/10.1177/1029864918763480

Schneider, R. (2019). "Konservenglück in Tiefkühl-Town": Das Songkorpus als empirische Ressource interdisziplinärer Erforschung deutschsprachiger Poptexte. In: *Proceedings of the 15th Conference on Natural Language Processing* (pp. 229–236). German Society for Computational Linguistics & Language Technology. https://konvens.org/proceedings/2019/Proceedings_of_the_15th_Conference_on_Natural_Language_Processing_KONVENS_2019.pdf

Schneider, R. (2020). A corpus linguistic perspective on contemporary German pop lyrics with the multi-layer annotated "Songkorpus". In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 842–848). ELRA. https://www.aclweb.org/anthology/2020.lrec-1.105.pdf

Schneider, R. (2022a). Zwischen Schriftlichkeit und Mündlichkeit: Songtexte in der deskriptiven Sprachforschung. *Sprachreport*, *1*, 38–50. https://doi.org/10.14618/sr-1-2022_schn

Schneider, R. (2022b). Das Songkorpus: Perspektiven einer korpuslinguistischen Nutzung deutschsprachiger Popmusik für die Fremd- und Zweitsprachenvermittlung. *Korpora Deutsch als Fremdsprache*, *2*(2), 149–153. https://doi.org/10.48694/kordaf.3549

Schneider, R., Lang, C., & Hansen, S. (2022). Das Vokabular von Songtexten im gesellschaftlichen Kontext: Ein diachron-empirischer Beitrag. In H. Kämper, & A. Plewnia (Eds.), *Sprache in Politik und Gesellschaft: Perspektiven und Zugänge* (pp. 295–304). Mouton de Gruyter. https://doi.org/10.1515/9783110774306-017

Schourup, L. (1999). Discourse markers. *Lingua*, *107*(3/4), 227–265. https://doi.org/10.1016/S0024-3841(96)90026-1

Squires, L. (2019). Genre and linguistic expectation shift: Evidence from pop song lyrics. *Language in Society*, *48*(1), 1–30. https://doi.org/10.1017/S0047404518001112

Summer, T. (2018). An analysis of pop songs for teaching English as a foreign language: Bridging the gap between corpus analysis and teaching practice. In V. Werner (Ed.), *The language of pop culture* (pp. 187–209). Routledge. https://doi.org/10.4324/9781315168210-9

Tegge, F., & Parry, K. (2020). The impact of differences in text segmentation on the automated quantitative evaluation of song-lyrics. *PLOS ONE*, *15*(11), e0241979 https://doi.org/10.1371/journal.pone.0241979

Van Venrooj, A., & Schmutz, V. (2018). Categorical ambiguity in cultural fields: The effects of genre fuzziness in popular music. *Poetics*, *66*, 1–18. https://doi.org/10.1016/j.poetic.2018.02.001

Walker, K. (2016). 50 years of pop music. https://www.kaylinpavlik.com/50-years-of-pop-music/

Werner, V. (2012). Love is all around: A corpus-based study of pop music lyrics. *Corpora*, *7*(1), 19–50. https://doi.org/10.3366/cor.2012.0016

Werner, V. (2018). Linguistics and pop culture: Setting the scene(s). In V. Werner (Ed.), *The language of pop culture* (pp. 3–26). Routledge. https://doi.org/10.4324/9781315168210-1

Werner, V. (2019a). Assessing hip-hop discourse: Linguistic realness and styling. *Text & Talk*, *39*(5), 671–698. https://doi.org/10.1515/text-2019-2044

Werner, V. (2019b). Lyrics and language awareness. *Nordic Journal of Modern Language Methodology*, *7*(1), 4–28. https://doi.org/10.46364/njmlm.v7i1.521

Werner, V. (2021a). Catchy and conversational? A register analysis of pop lyrics. *Corpora*, *16*(2), 237–270. https://doi.org/10.3366/cor.2021.0219

Werner, V. (2021b). A register approach toward pop lyrics in EFL education. In E. Seoane, & D. Biber (Eds.), *Corpus-based approaches to register variation* (pp. 209–234). Benjamins. https://doi.org/10.1075/scl.103.08wer

Werner, V. (2021c). Teaching grammar through pop culture. In V. Werner, & F. Tegge (Eds.), *Pop culture in language education: Theory, research, practice* (pp. 85–104). Routledge. https://doi.org/10.4324/9780367808334-6

Werner, V. (2021d). Text-linguistic analysis of performed language: Revisiting and remodeling Koch & Oesterreicher. *Linguistics*, *59*(3), 541–575. https://doi.org/10.1515/ling-2021-0036

Werner, V. (2022). Pop cultural linguistics. In M. Aronoff (Ed.), *The Oxford research encyclopedia of linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.999

Wiemeyer, L., & Schaub, S. (2018). Dimensions of dissatisfaction and dissent in contemporary German rap: Social marginalization, politics, and identity formation. In A. S. Ross, & D. J. Rivers (Eds.), *The sociolinguistics of hip-hop as critical conscience* (pp. 37–67). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-59244-2_3

Correspondence

Valentin Werner
University of Bamberg
Institute of English and American Studies
valentin.werner@uni-bamberg.de

Jan Langenhorst / Yannick Frommherz / Simon Meier-Vieracker

Keyness in song lyrics: Challenges of highly clumpy data

Abstract

Computer-assisted stylistic analyses regularly employ the calculation of keywords. We show that the inclusion of a separate dispersion measure in addition to a frequency measure into keyword analysis (or more generally: keyness analysis), as proposed by Gries (2021), is a necessary extension of said analyses. Using texts from the German *Songkorpus*, we demonstrate that traditional keyword calculations using only frequency measures lead to spurious results. Determining keywords by both measuring a word's frequency and its dispersion in comparison to a reference corpus gives a more realistic view. This is especially relevant for our corpus, since song lyrics turn out to be extraordinarily clumpy data: Words that are very frequent in one artist's subcorpus typically only occur in a few or even just a single one of their songs due to widespread word repetition within songs, e.g., in choruses. Song lyrics in our dataset are shown to not feature words that can be considered key at all. Our contribution is twofold: (1) We demonstrate the utility of Gries' (2021) approach and (2) interpret the (lack of) results in terms of a genre-specific property which is that song lyrics are lexically autonomous works of art.

Keywords: German Lyrics, Keyword Analysis, Dispersion, Clumpiness

1 Introduction

The goal of this paper is to show both potentials and limitations of keyness analysis as a contrastive style analysis using a sample of German song lyrics. While keyword analysis, most broadly defined as the identification of "words that are especially characteristic of the texts in a target discourse domain" (Egbert/Biber 2019: 77), is a widely used method to investigate both typical stylistic (Stubbs 2005) and genre-related (Xiao/McEnerey 2005) features of texts, it has rarely been applied to song lyrics. There exist corpus-based studies that take a frequency-oriented look at the characteristic properties of the genre of song lyrics as a whole in contrast to other varieties of text (Werner 2021; Watanabe 2018). Nevertheless, keyword or key-ngram analyses aiming at the detection of stylistic features of, say, artists or subgenres, are hardly available (but see Werner 2022 for a stylistic analysis of lyrics by rap artist Eminem, and Nishina 2017 for a general overview of stylistic features in pop songs). However, since it seems immediately plausible that artists have a characteristic and recognizable lyrical style, it is reasonable to look for measurable stylistic features at this level, too.

Our interest in stylistic features of song lyrics is grounded in a corpus pragmatic approach, which investigates frequent patterns of language use as results of recurring linguistic practices of the authors of the texts in a corpus (Bubenhofer/Scharloth 2012). Since style is a matter of choice from a semiotic repertoire referring to a socially meaningful way in which a linguistic act is carried out (Sandig 2006: 9), it can be studied particularly well in a contrastive manner.

As Bubenhofer and Scharloth (2012: 203) have argued, a corpus linguistic operationalization of style refers to a set of linguistic patterns by which one set of texts is significantly distinguished from another set of texts. This is exactly what is achieved by keyness analysis, which detects linguistic units "whose frequency (or infrequency) in a text or corpus is statistically significant, when compared to the standards set by a reference corpus" (Bondi 2010: 3).

As Culpeper and Demmen (2015: 93) put it in their extensive review of keyword analysis in corpus linguistics, "keywords tend to be of two main types: those relating to the text's 'aboutness' or content, and those which are related to style". While investigating the content or the thematic domains of a (set of) text(s) can be a most interesting task also in the case of song lyrics (Schneider/Lang/Hansen 2022), a stylistic analysis can bring into focus the indexical aspects of linguistic choices. For example, features associated with colloquial style or with dialects (i.e., features with social meanings as mentioned above) may indicate social positioning in certain groups or milieus (Meier-Vieracker 2022: 17–21). How these features, which serve as characteristic style markers (Kreyer and Mukherjee 2007), can be found for specific texts using keyness analysis, is a methodological question of the metrics and the statistical measures (see the extensive review in Gabrielatos 2018). Roughly speaking, measuring statistically significant differences will favour high-frequency items like pronouns which are good candidates for style markers. When measuring effect size, on the other hand, less frequent but exclusive items are favoured.

As we will show in the following sections, standard approaches to keyness analysis run into serious problems with the genre of song lyrics because of its repetitiveness. Although repetition or recurrence itself can be related to style and key items do constitute "chains of repetition in text" (Bondi 2010: 3), the repetitiveness of song lyrics leads to an uneven distribution or *clumpiness* of recurrent items that distort the results. For that reason, we turn to an alternative approach to keyness analysis introduced by Gries (2021) which not only takes frequency into account, but also dispersion. To our knowledge, this paper is the first to implement and apply this method. However, as we will show, even this approach does not lead to interpretable results that can be used in a stylistic analysis of song lyrics. This may have something to do with the rather small dataset. Conversely, the lack of results may also tell us something about the genre of song lyrics in general.

2 Corpus

For our analysis, we use data made available as part of the newly compiled *Songkorpus - Linguistic Corpus of German Song Lyrics* (Schneider 2020).[1] The subcorpus consists of song lyrics performed by seven German artists (singers and bands, see Table 1) of different genres, written between 1969 and 2021. The data allows us to evaluate differences in language use between artists.

[1] Parts of the corpus, including word counts and n-gram lists, are publicly available at https://songkorpus.de.

Artist	Albums	Texts (= songs)	Tokens (share in total corpus)
Udo Lindenberg	48	360	91,216 (21.73%)
Konstantin Wecker	60	283	87,628 (20.87%)
Fettes Brot	16	143	68,718 (16.37%)
Stoppok	17	191	48,030 (11.44%)
Element of Crime	15	114	26,270 (6.26%)
Ulla Meinecke	10	86	21,061 (5.02%)
Hannes Wader	23	210	76,858 (18.31%)
Total	189	1,387	419,781

Table 1: Overview of selected artists in the *Songkorpus*.

3 Traditional approaches to keyness

In order to analyze an artist's language on a lexical level, keyness analysis in which one artist's word frequencies are compared to those of all other artists in a corpus seems to be an appropriate approach. This approach is straightforward both in its calculation (only word frequencies and a short formula are needed) and in its interpretation (words are attracted to or repelled by a corpus to a quantifiable degree). As mentioned above, statistical measures relying on significance are particularly suitable for stylistic analysis, and a widely used measure is Log Likelihood Ratio (*LLR*, Dunning 1993). Using a contingency table, the observed frequencies of a word (or lemma, n-gram etc.) in both a target corpus and a reference corpus are compared to the expected frequencies given an even distribution of the words' frequencies across both corpora. Observed frequencies that deviate from expected frequencies most yield a high LLR value and are interpreted as being most key for a given corpus. Positive keywords are more frequent in the target corpus than expected and can be interpreted as being characteristic or typical for the target corpus texts, while negative keywords occur less frequently than expected and are interpreted as atypical. This type of analysis is a standard method in corpus linguistics and is implemented in many popular tools such as CQPweb (Hardie 2012) or SketchEngine (Kilgarriff et al. 2014).

Calculating keywords for the German singer-songwriter Hannes Wader using this approach in its most basic implementation – neither requiring keywords to have a minimum absolute frequency in the target corpus nor excluding stopwords – yields the results shown in Table 2 (only positive keywords).

Word	Target range	Reference range	Target frequency	Reference frequency	LLR
&	58	214	510	1017	198.61
!	117	434	757	1815	188.68
–	41	62	208	250	176.57
alledem	3	8	57	8	148.32
ciao	1	0	30	0	101.88
na	4	28	82	70	97.03
sah	33	54	76	70	84.26
hatte	24	83	97	115	83.57
Cocaine	2	0	21	0	71.31
Bollmann	1	0	21	0	71.31
kreich	1	0	21	0	71.31
Frubben	1	0	21	0	71.31
sine	1	0	21	0	71.31
Nun	31	29	49	35	66.45
trotz	5	18	38	20	62.41

Table 2: Top 15 Keywords for Hannes Wader compared to all other artists in the *Songkorpus.*

Ignoring the ampersand and the punctuation marks at the top of the list,[2] the keyword with the highest LLR is *alledem* ('all that'). The table also shows the range of each word, i.e., in how many different texts it occurs at least once, for both the target and reference corpus. As can be seen, the 57 occurrences of the word *alledem* in the Hannes Wader subcorpus stem from only three different songs, and the word only occurs in eight different songs in the reference corpus. This is due to the fact that Wader recorded three different versions of *Trotz alledem* ('in spite of all that'), a song based on a 19th century German poem (based on an even older Scottish one):

> Das war 'ne heiße Märzenzeit trotz Regen, Schnee und **alledem**!
> Nun aber, da es Blüten schneit, nun ist es kalt, trotz **alledem**!
> Trotz **alledem** und **alledem** – Trotz Wien, Berlin und **alledem**

This particular keyword derives its keyness from the fact that it is repeated very often in a small number of songs. Same goes for the runner-up *ciao* which occurs 30 times in the Hannes

[2] Punctuation marks are not sung, but set during transcription. Also, transcription conventions differ between the artists. Thus, they are excluded.

Wader corpus and not once in the reference corpus, but the word only ever occurs in one single song, an interpretation of the popular Italian partisan hymn *Bella Ciao*. The word *na*, then, seems to be an interesting candidate for style analysis because it serves as an interjection (e.g., *Na, Willy* or *na gut*) indicating a colloquial style, but also as a non-lexical vocable (*na na na na*). Upon further inspection, the word turns out to only occur in a very small number of songs, but it is not even evenly dispersed across said songs with 96% of its occurrences clustered in one single text where the word is used as a most repetitive non-lexical vocable. To conclude, relying on these LLR values leads to misinterpretation, because single words may seem as typical of an artist while they are in fact typical of certain songs only. Assessing the range values in addition to LLR does certainly add valuable information. However, as seen in the *na* example, it hides how the occurrences of a word are distributed within this range.

Since LLR-based keyword analysis of concrete word forms or lemmas is distorted by the repetitiveness of the genre of song lyrics, a focus on more abstract patterns seems to be promising. Particularly appropriate are part-of-speech ngrams (POS-ngrams) which allow for capturing typical syntactic patterns and contextual embeddings that are especially informative for style analysis (Bubenhofer & Scharloth 2012). For example, a POS-trigram analysis for the singer-songwriter Konstantin Wecker yields the results shown in Table 3:

Ngram	Target range	Reference range	Target frequency	Reference frequeny	LLR
$, KON ADV	109	245	245	387	105.01
NN KON NN	152	401	386	812	84.01
VVPP $, KON	59	82	100	107	76.75
VVINF $, KON	76	122	116	156	65.4
$, KON ART	71	124	120	167	64.11

Table 3: Top five POS-trigrams for Konstantin Wecker compared to all the other artists in the *Songkorpus*.

At first glance, POS-trigrams are more evenly distributed throughout the corpus and should therefore be more informative for style analysis. An interesting finding is the keyness of the POS-trigram NN KON NN which occurs in 152 out of 283 Konstantin Wecker songs (54%) and can thus be seen as a rather common feature of this artist's songs. It is the syntactic form of binomial pairs which typically are (partially) idiomatic expressions with a non-compositional meaning like *milk and honey*. Since binomial pairs usually meet both formal (i.e., phonological) and semantic requirements (Benor and Levy 2006) and make up preassembled wholes in language use, their use can be described as a salient stylistic means (Burger 2015: 55f.). Moreover, they are part of a wide range of sayings and proverbs (Müller 2009). Examples of binomial pairs in the songs of Konstantin Wecker include *Freiheit und Demokratie ('freedom and democracy'), Dämmern und Morgenrot ('twilight and dawn),* and

Brutalität und Gier (brutality and greed'), where the nouns are conceptually linked constituting formulaic patterns. Additionally, there are more creative pairs like *Büro und Illusionen ('office and illusions')* or *Bier und Beifall ('beer and applause')*, which by their very form call for an interpretation that allows for conceptual commonalities to emerge. As a highly recurrent pattern in Wecker's songs, binomial pairs thus seem to be a characteristic and creatively used stylistic feature. Further, in contrast to the *na* example above, a follow-up analysis revealed that the pattern is fairly evenly distributed within the range of songs featuring it. We will revisit this pattern later.

As demonstrated, such a shift in focus to more abstract patterns can indirectly remedy the above-mentioned deficiency of an LLR-based keyness analysis which is that dispersion across texts is not considered at all.[3] While one could also introduce range thresholds (e.g., a word or pattern must appear in at least 30% of all texts), this would be an arbitrary measure which also leaves it unclear whether within this subset a word is dispersed evenly across texts or predominantly occurs in just one of them. As seen, this is especially problematic for song lyrics, as they are particularly repetitive by their nature. Not only choruses are repeated, but also single words or phrases may appear again and again in a given song. Thus, song lyrics are especially susceptible to containing *clumpy* data, i.e., words or patterns which have a low *dispersion*. This makes it difficult to use traditional approaches to keyness analysis.

The problem of (lacking) dispersion in keyness analysis has been discussed before (Egbert and Biber 2019), and most recently, Gries (2021) proposed a new approach to calculating keyness which incorporates a word's dispersion over the corpus as well as its frequency into keyness calculations. This design promises to solve the problem of clumpy data, so we will turn to this approach in the following section.

4 Adding dispersion to the mix

Gries' (2021) newly proposed method turns keyness into a two-dimensional concept with one dimension being a measure that is based on word frequencies and a second one which measures the dispersion of a word over a corpus. The frequency-based measure is calculated using the so-called *Kullback-Leibler Divergence* which determines the divergence of two probability distributions as follows[4]:

$$KLD_{freq} = DKL(P(Corpus|Word) \parallel P(Corpus)) = \left(a \times log_2 \frac{a}{e}\right) + \left(b \times log_2 \frac{b}{f}\right)$$

Where a is a word's relative frequency (i.e., its probability of occurrence) in a target corpus, b is the relative frequency (i.e., its probability of occurrence) of said word in a reference corpus and e and f are the respective proportions of both corpora in the complete dataset (i.e., their respective probabilities of occurrence). Put simply, one asks: 'What is the probability

[3] For a differentiation between dispersion in a corpus linguistic and a statistical sense, see Sönning (2022: 7-8).

[4] Zero is inserted instead of log values where $\frac{a}{e} = 0$ or $\frac{b}{f} = 0$.

that I am looking at corpus A (not corpus B) given that I am looking at word X?'. This probability might diverge from the *overall* probability distribution of looking at corpus A. In our example, if we were to lump both our target and reference corpus together and randomly chose a word, there is a certain probability that we are looking at a word from the Hannes Wader subcorpus given that the word we chose is *alledem*. This probability might diverge from the overall probability of choosing a word from the Hannes Wader subcorpus. The stronger this divergence, the more key to either the target or reference corpus we consider the word. As Gries (2021) shows, this measure decouples the frequency of a word and its association with a corpus to a greater extent, compared to an LLR calculation. The resulting values are normalized to fall within the range of [0, 1] for words that frequency-wise are attracted by the target corpus, and [-1, 0] for words that are repelled by it. 1 would mean the strongest possible attraction and -1 the strongest possible repulsion.

Dispersion is added as a second dimension and measured by again calculating the Kullback-Leibler-Divergence:

$$KLD_{disp} = \sum_{i=1}^{n} p_i \times log_2 \frac{p_i}{q_i}$$

Where n is the number of texts in a given corpus, p_i is the proportion of all occurrences of a given word that occur in text i within the corpus and q_i is said text's proportion within the whole of the corpus. Target and reference corpus are compared by subtracting a word's two KLD_{disp} values for each corpus from one another.[5] This value is again normalized to fall in the range [-1, 1] where a value of 1 would indicate that a word is very dispersed in a target corpus compared to a reference corpus and -1 would mean that a word is very dispersed in the reference corpus while at the same time very clumpy in the target corpus.

For the Hannes Wader subcorpus keyword candidate *alledem*, these two calculations yield a KLD_{freq} value of 0.81 and a difference in KLD_{disp} values of -0.001, respectively. The very high frequency of the word compared to the reference corpus is reflected in a very high KLD_{freq} result. At the same time, the vanishingly small difference in KLD_{disp} values adequately conveys that this word is not well dispersed over the Hannes Wader subcorpus at all (it is minimally more dispersed over the reference texts, even though this difference is negligible). Thus, using Gries' (2021) method, this word would not be considered key regardless of its frequency in the target corpus compared to its frequency in the reference corpus. Due to its multidimensional nature, results of this type of keyness analysis can best be assessed by plotting the frequency and dispersion values against each other. Figure 1 shows the results for Hannes Wader. As one can clearly see, no words obtain high values on both scales and, accordingly, there are no words that can be considered key for the chosen artist. Looking at texts by the German Hip Hop group *Fettes Brot*, which can be expected to have lyrics very different from Hannes Wader, we observe very similar results (Figure 2).

[5] The method is explained at great length using different examples in Gries (2021).

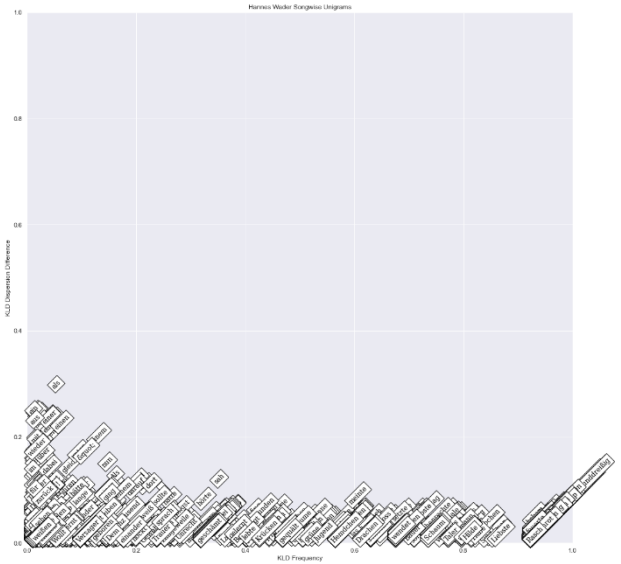


Figure 1: KLD Keyword analysis for Hannes Wader.

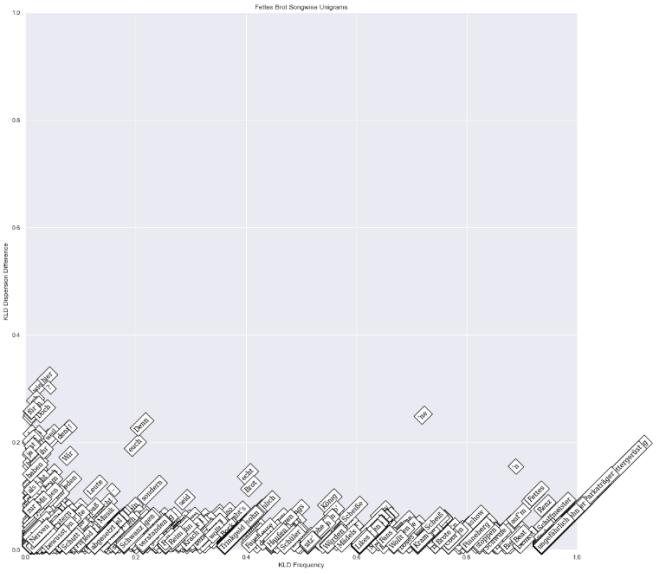


Figure 2: KLD Keyword analysis for Fettes Brot.

This pattern holds true for every single artist from our corpus when compared to the rest of the corpus (Figure 3). Words tend to be scattered along the KLD_{freq} axis with many words obtaining very high values while most words obtain very low values on the KLD_{disp} axis (this can be seen most easily looking at the marginal plots in Figure 3). Words that are far more frequent in the target corpus as compared to the reference corpus do exist for every artist, but they tend to not be more dispersed over the respective artist's repertoire in comparison to the other artists. Song lyrics' word distributions, at least those in our dataset, seem to be extremely clumpy. Gries' (2021) results for the Clinton-Trump corpus (Brown 2016) look somewhat different with words being scattered more across the KLD_{disp} axis rather than concentrating just slightly above zero like in our data. However, a replication of his results including marginal plots (see Figure 4) reveals a generally similar distribution for both axes. There is only one fundamental difference between our results and results from the Clinton-Trump corpus: In contrast to the Clinton-Trump results, there simply aren't any 'real' keywords in our data.

Revisiting the originally promising results for more abstract patterns using the traditional LLR approach (Ch. 2), POS-trigrams actually plot very similarly compared to single words following the KLD method (see Figure 5 for all artists plotted on top of each other). This means that also the pattern *NN KON NN*, which appeared to be typical for Konstantin Wecker when employing an LLR calculation (Table 3), is not an actual POS-key-trigram for Wecker, according to the KLD method (it has a KLD_{freq} value of 0.05 and a difference in KLD_{disp} value of 0.13). The pattern's high frequency, relatively wide range and the fact that it is rather well dispersed within this range in the Wecker corpus do not lead to high values. For frequency, because the probability of looking at the Wecker subcorpus given that we are looking at *NN KON NN* does not strongly diverge from the overall probability of looking at the Wecker subcorpus (the same applying to the reference corpus). For dispersion, because the pattern is only slightly more dispersed in the Konstantin Wecker subcorpus, compared to the reference corpus.

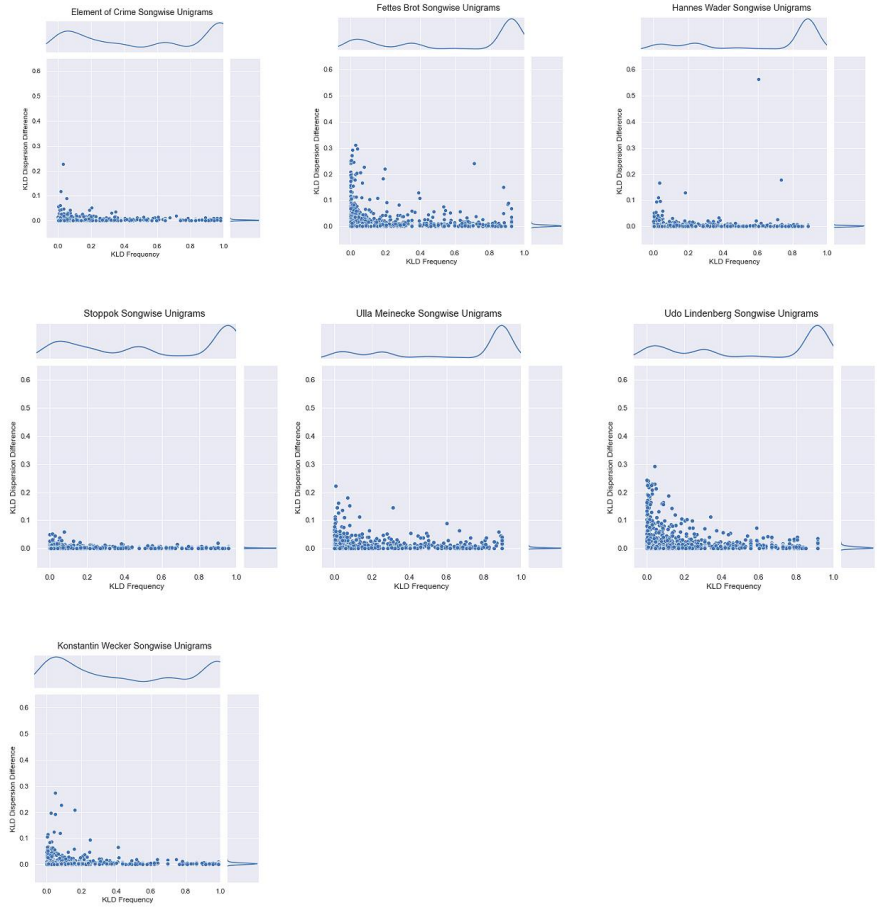


Figure 3: KLD-Keyword analysis for all artists in the corpus (dots represent words).

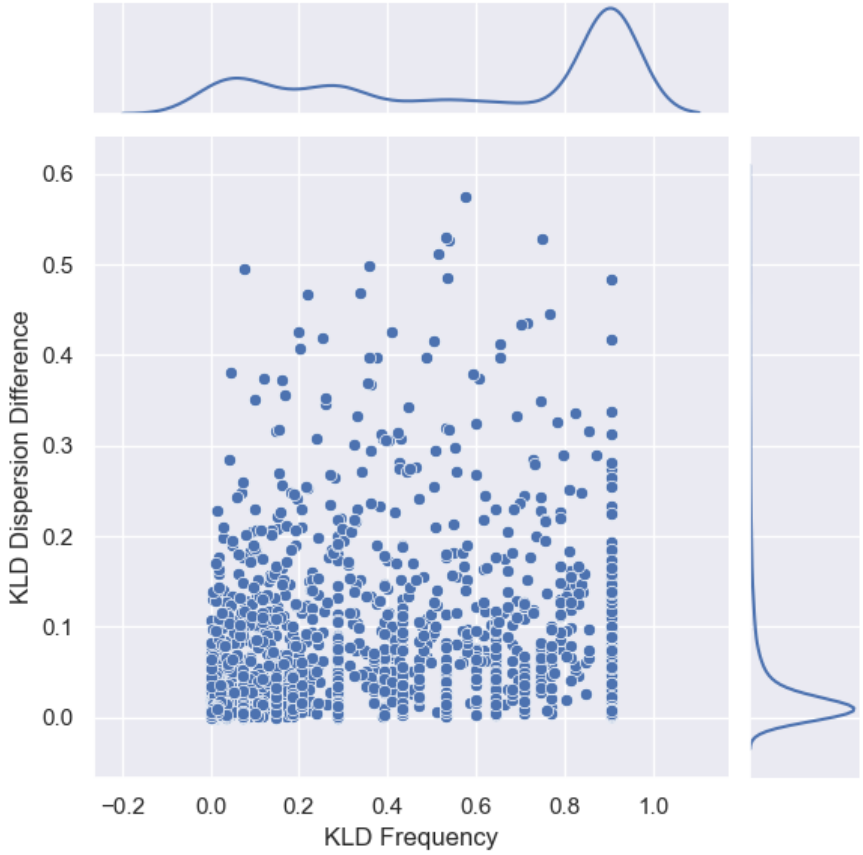


Figure 4: KLD-Keyword analysis replication of Gries' (2021) results including marginal plots for the Clinton-Trump Corpus (Brown 2016). Minor differences due to a different method of tokenization.

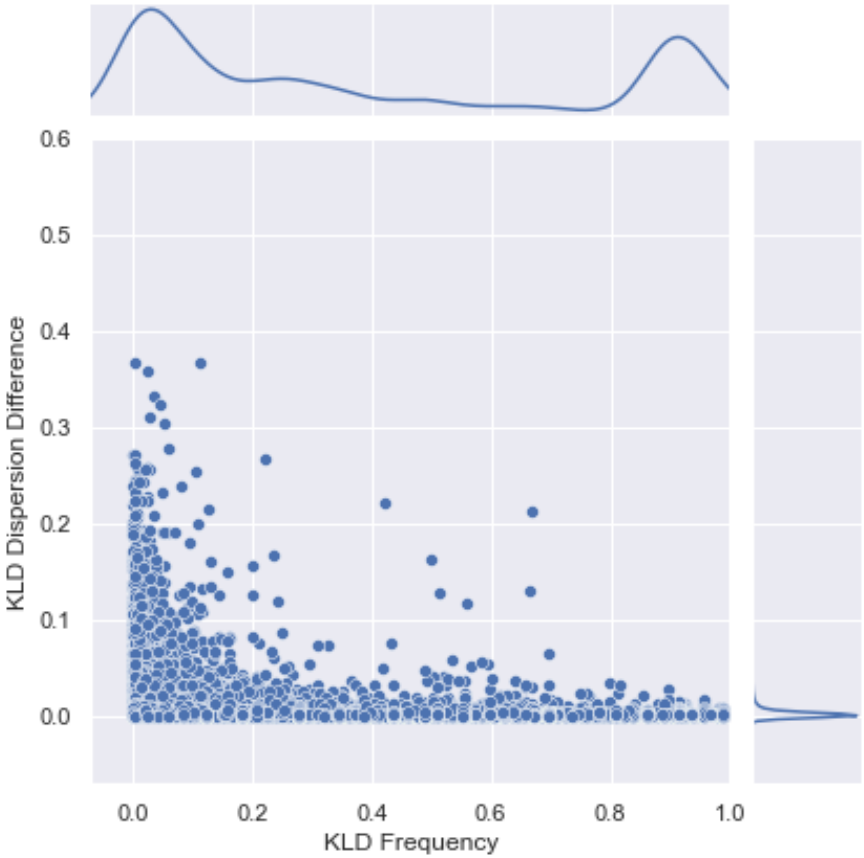


Figure 5: KLD POS Key-trigram analyses for all artists in the corpus (dots represent trigrams).

5 Clumpiness and granularity

When measuring dispersion, it is not always straightforward across *what* an item should be dispersed. The *Songkorpus* is organized into individual song lyrics, but these lyrics in turn are part of albums, which are also annotated. In our analysis thus far, we considered song lyrics as the unit of text, and hence the KLD_{disp} measure compared dispersion values for words across song lyrics. However, one might also conduct the very same analysis on a higher level with albums as the unit of text. This might be a legitimate approach: While individual song lyrics are indisputably the 'real' unit in the sense that they were written as discrete texts by

their respective author, albums also constitute natural units of text in a wider sense. Grouping items on the album level potentially impacts the results of a keyness analysis that incorporates the use of dispersion values. One could assume that a lower number of units in which a word can appear leads to a higher probability of it being dispersed more evenly across these units.

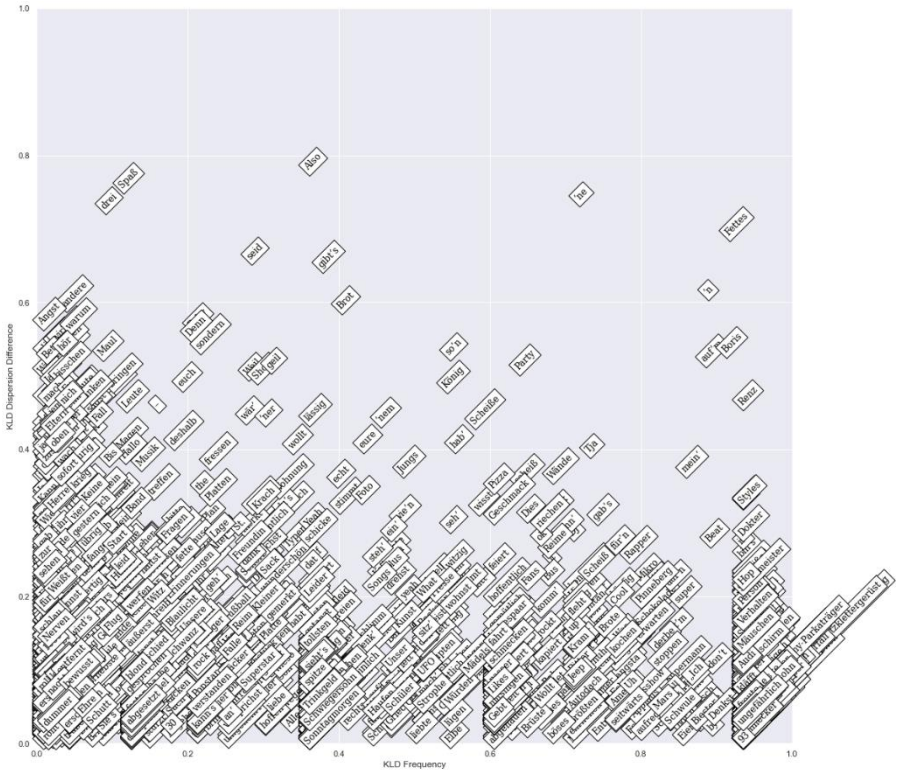


Figure 6: Albumwise KLD Keyword analysis for Fettes Brot.

Keywords computed ‘albumwise’ differ most from ‘songwise’ keywords for Fettes Brot compared to all other artists (see Figure 6). There is a small number of words that are well dispersed over the band’s albums as well as being more frequent in comparison to the rest of the corpus: *Fettes* which is part of the band’s name and tends to get mentioned on many of their releases, as well as parts of the band members’ names such as *Boris*, *Renz*, *König* and colloquial contractions such as *auf’m* (‘on the’).

Figure 7 shows all ‘albumwise’ keywords for all artists compared to the rest of the corpus plotted into the same graph (all words are plotted seven times, once for each artist as reference artist). As can be seen, these results differ from the results for a ‘songwise’ keyness analysis and now look more similar to Gries’ results (see Figure 4). However, there are only very few

keywords for each artist (compared to many for Clinton speeches vs. Trump speeches and vice versa) and most words do not obtain high values on both scales, despite the lower number of text units on which the KLD_{disp} measure is based.

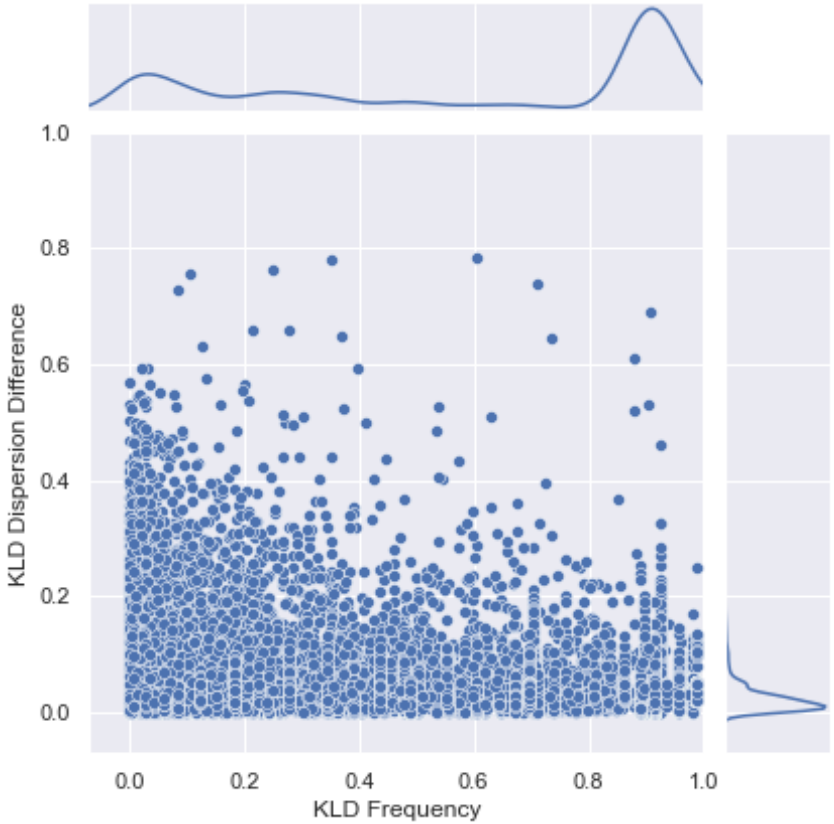


Figure 7: Albumwise KLD Keyword analysis for all artists plotted on top of each other.

The question of what level constitutes the most natural textual unit within a corpus will also arise in settings where researchers deal with corpus data such as novels (whole volumes vs. chapters), newspapers (whole issues vs. sections vs. articles), etc. An adequate level on which dispersion is measured should be chosen wisely in any case.

6 What does clumpiness mean in the case of song lyrics?

Revisiting our initial assumption that one can intuitively distinguish artists from one another by their usage of certain words: this intuition has not been disproven by the analysis presented above. Words occurring only once in an artist's repertoire certainly aren't keywords in a quantitative sense. But they might belong to a larger class of words that, in turn, is typical of a given artist, such as lexical words that represent a certain topic (love, politics, etc.). Studies employing a theme-based approach (e.g., using wordlists) to track words which might be infrequent, but still typical of a certain artist (for human ears) might be better suited for identifying typical patterns.

The extreme clumpiness found in song lyrics, then, can itself be interpreted and compared to other types of data. The only other analysis conducted following Gries' (2021) method – Gries' own case study of keywords in the Clinton-Trump corpus – finds a number of keywords that are both more frequent and dispersed in either of the corpora compared to the other one. If election speeches do contain 'real' keywords and song lyrics do not, this can be seen as informative of the respective genres. On the one hand, during electoral campaigns, politicians try to get their message across in a fairly 'standardized' way, often relying on stump speeches. For example, if education is an important topic in one politician's campaign and a focus on said topic a feasible way of distinguishing oneself from their opponent, then *education* will consistently occur repeatedly in most if not all campaign speeches (leading to high dispersion) while the same most probably cannot be said about their opponent. This likely results in keywords becoming visible in the way described in Gries' (2021) paper. On the other hand, song lyrics are first and foremost individual pieces of art. They do belong to the greater project of an artist's oeuvre, but this type of coherence is apparently not created by word or pattern repetitions across songs.

7 Conclusion

As we could demonstrate in our analysis, word dispersion matters when analyzing keywords. Song lyrics appear to be a case where 'traditional' keyword-oriented style analysis based on mere frequency counts falls short. Our evaluation of Gries' (2021) multidimensional approach to keyness clearly showed its usefulness. Including a measure of a word's or pattern's dispersion over both the target and reference corpus made disappear results that would have given a false impression of typicality. Words that would initially yield high LLR values and could thus be interpreted to be *key* for a given artist were shown to be artifacts of word repetition within songs. A strength of Gries' (2021) method is that the KLD_{disp} measure does not require the use of an arbitrary range threshold and also captures a word's distribution *within* its range of occurrence. Another aspect we have briefly touched on is that when introducing a measure of dispersion as described above, one has to carefully reason about the levels that are 'naturally' present in the data. While the change of level from single songs to whole albums in our corpus did not alter our results in a substantial way, this might be different

for other data. The more one knows about the underlying structure of a given corpus, the better one can control for a possibly clumpy dispersion.

The virtually complete absence of 'actual' keywords in the *Songkorpus* data might be a surprise. It becomes very plausible, however, when one inspects the data's specific pattern of frequent word repetitions within single songs and few repetitions across songs. The repetition within a song is a typical stylistic device in song lyrics and the fact that word repetition across songs rarely occurs suggests that song lyrics are independent works of art. A limitation of our study is that our subcorpus includes 7 different artists and, depending on how one makes these distinctions, only 2 to 5 genres. The observed results might not hold true for a corpus that is structured differently and features a greater number of artists or artists representing a different set of genres. While our subcorpus contained complete discographies of a small number of artists, the *Songkorpus* archive also contains, e.g., a *Charts Archive* featuring a more diverse set of artists and genres with a smaller number of songs per artist. This dataset could be used for follow-up analyses.

An awareness for the importance of dispersion for keyness analysis seems to generally be on the rise and methods that alleviate the risk of making false assumptions based on frequency-only-methods are being refined. Besides Gries' (2021) and Egbert/Biber's (2019) method, another very promising approach incorporating both frequency and dispersion measures using negative binomial regression has very recently been proposed by Sönning (2022). Available approaches for improving keyness analysis should be evaluated on a greater number of different corpora and their performance should be compared. There exist numerous text genres where one can expect data to be potentially clumpy and an inclusion of dispersion measures might be warranted. For example, newspapers, which are a popular source for general corpora, might have a very particular distribution of certain words across their sections. In these cases, clumpiness might pose less of a problem compared to song lyrics, which we suspect to be an extreme case, but controlling for dispersion should ideally become a standard procedure which should also be included in corpus analysis software.[6]

Keyness analysis in general has proven to be a useful tool for style analysis, partly because it is not based on strong presumptions on the side of the researcher. As is becoming clearer and clearer, however, the available methods of keyword calculation have relied too strongly on a latent assumption of a general correlation between frequency and dispersion. The 'naïve' keyword list calculation using log-likelihood-ratios or similar measures is in many cases an insufficient representation of the occurrence of words or larger patterns in corpus data.

[6] For example, CQPweb v3.2.43 (Hardie 2012) does provide the calculation of dispersion of query results, but the feature is still experimental and buggy.

Data availability

Code for calculating keyness measures, results for all keywords, and code for reproducing the graphs presented in this paper are available at TU Dresden's OpARA platform (https://doi.org/10.25532/OPARA-220).

References

Benor, S., & Levy, R. (2006). The Chicken or the Egg? A Probabilistic Analysis of English Binomials. Language, 82(2), 233–278. https://doi.org/10.1353/lan.2006.0077

Brown, David (2016): Clinton-Trump Corpus. https://www.kaggle.com/datasets/browndw/clintontrump-corpus

Bondi, M. (2010). Perspectives on keywords and keyness: An introduction. In M. Bondi & M. Scott (Eds.), Keyness in texts (Bd. 41, S. 1–18). Amsterdam: Benjamins. https://doi.org/10.1075/scl.41.01bon

Bubenhofer, N., & Scharloth, J. (2012). Datengeleitete Korpuspragmatik. Korpusvergleich als Methode der Stilanalyse. In E. Felder, M. Müller, & F. Vogel (Hrsg.), Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen (S. 195–230). Berlin, New York: de Gruyter.

Culpeper, J., & Demmen, J. (2015). Keywords. In D. Biber & R. Reppen (Hrsg.), The Cambridge Handbook of English Corpus Linguistics (S. 90–105). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139764377.006

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), 61–74.

Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. Corpora, 14(1), 77–104. https://doi.org/10.3366/cor.2019.0162

Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. In Taylor & A. Marchi (Hrsg.), Corpus Approaches To Discourse. a Critical Review (S. 225–258). London: Routledge.

Gries, S. Th. (2021). A new approach to (key) keywords analysis: Using frequency, and now also dispersion. Research in Corpus Linguistics, 9(2), 1–33. https://doi.org/10.32714/ricl.09.02.02

Hardie, A. (2012). CQPweb — combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics, 17(3), 380–409. https://doi.org/10.1075/ijcl.17.3.04har

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., et al. (2014). The Sketch Engine: ten years on. Lexicography, 1(1), 7–36.

Kreyer, R., & Mukherjee, J. (2007). The Style of Pop Song Lyrics: A Corpus-linguistic Pilot Study. Anglia, 125(1), 31–58. https://doi.org/10.1515/ANGL.2007.31

Meier-Vieracker, S. (2022). Between consumers and fans: Writing fan reports as a multifunctional evaluation practice. Journal of Cultural Analytics, 7(2), 4–31. https://doi.org/10.22148/001c.33570

Müller, H.-G. (2009). Adleraug und Luchsenohr. Deutsche Zwillingsformeln und ihr Gebrauch. Frankfurt u.a.: Peter Lang.

Nishina, Y. (2017). A Study of Pop Songs based on the Billboard Corpus. International Journal of Language & Linguistics, 4(2), 125–134.

Sandig, B. (2006). Textstilistik des Deutschen. Berlin, New York: De Gruyter.

Schneider, R. (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 (S. 842–848). Marseille: European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.105

Schneider, R., Lang, C., & Hansen, S. (2022). Das Vokabular von Songtexten im gesellschaftlichen Kontext – ein diachron-empirischer Beitrag. In Sprache in Politik und Gesellschaft (S. 295–304). De Gruyter. https://doi.org/10.1515/9783110774306-017

Stubbs, M. (2005). Conrad in the computer: examples of quantitative stylistic methods. Language and Literature, 14(1), 5–24. https://doi.org/10.1177/0963947005048873

Sönning, Lukas (2022): Count regression models for keyness analysis. PsyArXiv. https://psyarxiv.com/25mwj/. Preprint.

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

Werner, V. (2021). Catchy and conversational? A register analysis of pop lyrics. Corpora, 16(2), 237–270. https://doi.org/10.3366/cor.2021.0219

Werner, V. (2022): “Guess who’s back, back again”. In: Stylistic Approaches to Pop Culture. New York: Routledge. S. 176–204. https://doi.org/10.4324/9781003147718-9

Xiao, Z., & McEnery, A. (2005). Two Approaches to Genre Analysis: Three Genres in Modern American English. Journal of English Linguistics, 33(1), 62–82. https://doi.org/10.1177/0075424204273957

Correspondence

Jan Langenhorst
TU Dresden
Institute of German Studies and Media Cultures
jan.langenhorst@tu-dresden.de

Yannick Frommherz
TU Dresden
Institute of German Studies and Media Cultures
yannick.frommherz@tu-dresden.de

Simon Meier-Vieracker
TU Dresden
Institute of German Studies and Media Cultures
simon.meier-vieracker@tu-dresden.de

Marco Gierke

Ist *alte Schule oldschool*? Zum ‚Nutzen' von Anglizismen in Deutschraptexten

Abstract

Der vorliegende Beitrag vergleicht die Verwendung der anglizistischen Nomination *old school* und der nativen Entsprechung *Alte Schule* im Hip-Hop-Subkorpus des Songkorpus (Schneider 2020). Dieser Vergleich erfolgt auf zwei Ebenen: Zum einen wird die diskursspezifische Verwendung anhand eines adaptierten Analyse-Frameworks für Hip-Hop-Texte von Androutsopoulos und Scholz (2002) untersucht, zum anderen wird der syntaktische und morphologische Gebrauch in den Deutschraptexten analysiert. Dabei zeigt sich, dass es jeweils spezifische Verwendungstendenzen auf diskursiver Ebene gibt, die wesentlichsten Unterschiede aber in der syntaktischen und morphologischen Verwendung auftreten, allen voran in der höheren Produktivität der anglizistischen Nomination. Es wird dafür argumentiert, dass sich dies unter anderem auf sprachstrukturelle bzw. wortformale Spezifika des Englischen zurückführen lässt, wie den nicht vorhandenen Flexionssuffixen der Adjektive. Damit werden die in der Anglizismenforschung etablierten Überlegungen zu Verwendungsgründen um eine simple, aber gegebenenfalls folgenreiche Beobachtung ergänzt, die sich vor allem bei den sprachökonomischen Ansätzen einordnen lässt. Schließlich wird darüber auf diskursiver Ebene wiederum auch ein Bezug zu terminologischen Vorteilen hergeleitet: Trotz flexibler Verwendung wird das schriftliche Abbild bei Wortbildungen geschont (*Oldschoolstyle*, *Oldschool-Aufnahmen*, *Oldschooler*), was für die Wiedererkennbarkeit des Diskurselements – neben der zusätzlichen Auszeichnung durch die Eigenschaft ‚fremdsprachig' – zuträglich sein könnte.

Keywords: Anglizismus, Hip-Hop, Deutschrap, Code-Switching, Crossing, Fachsprache, Nomination, Graphematik, Morphologie, Sprachwandel, Sprachökonomie

1 Einleitung

„[…] ich bin ohne das Oldschool-Feeling. Ich bin von der alten Schule." [Advanced Chemistry (1995): "Alte Schule"]

Der Begriff *Anglizismus* ist vermutlich älter als *Fremdwort* und seit dem 18. Jahrhundert Bestandteil sprachkritischer bzw. -puristischer Diskurse (vgl. Spitzmüller 2005: 166).[1] Bezeichnungen wie *Luxusentlehnung* illustrieren dabei – auch im fachwissenschaftlichen Kontext – die häufig diskutierte Nutzenfrage, die vor allem bei lexikalischen Entlehnungen gestellt wird (vgl. Altleitner 2007: 14 f.; Scholz 2004). Für den Anglizismus wird dann zumeist

[1] Für einen Abriss der Anfänge des Begriffs Anglizismus im sprachkritischen Diskurs siehe Spitzmüller 2005: 166–176.

entweder ein natives semantisches Äquivalent gesucht oder aber ihm wird eine wörtliche Übersetzung direkt gegenübergestellt. So verweist Altleitner (2007: 162) auf Braselmann, die den Unterschied zwischen *Backshop* und *Bäckerei* herausarbeitet (2002: 300) und der Verein Deutsche Sprache bietet mit dem aktivistischen Anglizismenindex [2] „deutschsprachige Entsprechung[en]", um „entbehrlichen Anglizismen" zu ‚begegnen' (Abruf 12.01.23). Neben der ohnehin zu problematisierenden Übersetzbarkeit (vgl. Altleitner 2007: 160) sollen im Rahmen der Fremdwortdebatte in der vorliegenden Untersuchung zwei wesentliche Stoßrichtungen näher beleuchtet werden: Zum einen soll der pragmatische Aspekt betont werden, zum anderen sollen aber auch konkrete syntaktische/morphologische Vorteile des Englischen für das Deutsche zur Diskussion gestellt werden. Diese sind zwar recht offensichtlich, in der Forschung zu Verwendungsgründen von Anglizismen m. W. n. aber noch nicht in dieser Deutlichkeit berücksichtigt worden. Hierzu wird das Vergleichspaar OLD_SCHOOL – ALTE SCHULE[3] im Hip-Hop-Subkorpus des Songkorpus (Schneider 2022) gebrauchsbezogen untersucht. Dieses Vergleichspaar bietet sich aus drei wesentlichen Gründen für eine solche Untersuchung an. Erstens handelt es sich hierbei um eine formal direkte Entsprechung, wodurch im Vergleich weniger Übersetzungsfaktoren zu interferieren drohen. Zweitens können beide Formen als im Deutschen etabliert bezeichnet werden – so ist OLD_SCHOOL in einschlägigen deutschsprachigen Wörterbüchern verzeichnet (vgl. bspw. Duden 2020), wodurch potenziell Zugriffspotential für Sprachproduzierende unterstellt werden kann. Drittens – und hier ergibt sich die Verknüpfung mit dem Songkorpus – kann innerhalb einer Domäne (Hip-Hop) verglichen werden, wodurch eine direktere Konkurrenz angenommen werden kann, die womöglich pointierter Gebrauchsunterschiede innerhalb dieser Domäne exponiert und zugleich den terminologischen Charakter im Hip-Hop-Diskurs herausstellt. Dass OLD_SCHOOL natürlich nicht (mehr) nur in diesem Kontext verwendet wird, zeigt sich exemplarisch auch Belegen wie dem folgenden:

(1) DA TUT SICH WAS IN SACHEN TEINT
Ansonsten – ganz old-school – drei Make-up-Abstufungen nebeneinander auf den Bereich zwischen Kinn und Hals zeichnen. Die Nuance, die dem Teint am nächsten kommt, ist Ihre. [DeReKo; BRG16/FEB.00051 BRIGITTE]

Neben dem syntaktischen/morphologischen Gebrauch wird für den pragmatischen auf ein Analysemodell von Androutsopoulos und Scholz (2002) zurückgegriffen, das zur sprachvergleichenden Analyse ganzer Raptexte entwickelt worden ist und hier für die Untersuchung eines einzelnen Vergleichspaars innerhalb einer Sprache erprobt wird.

[2] https://vds-ev.de/arbeitsgruppen/deutsch-in-der-oeffentlichkeit/ag-anglizismenindex/

[3] Zur ökonomischen und sachgerechten Zusammenfassung wird diese Notation stellvertretend für alle gängigen Varianten (u. A. <oldschool>, <old-school>, <old school>, <Oldschool>, <Old School>; <alte Schule>, <Alte Schule>) genutzt. Die Wahl einer dieser angeführten Varianten würde bereits semantische/grammatische Implikationen mit sich führen. Hierzu mehr in Abschnitt 3.3.

2 Warum Deutschraptexte?

Für diese Untersuchung ist das Hip-Hop-Subkorpus auf zwei Ebenen von besonderem Interesse: (1) Generell sind Raptexte durch die kulturelle Genese des Hip-Hops in Deutschland ein reichhaltiger Nährboden zur Untersuchung von Anglizismen und (2) für das vorliegende Untersuchungspaar an sich. Für die erste Ebene ist zunächst die Unterscheidung von Hip-Hop und Rap relevant. Hip-Hop bezeichnet ein ganzes (Sub-)Kulturgefüge, dass sich aus den vier Säulen DJing, Graffiti, Breakdance und Rap zusammensetzt (vgl. Güler Saied 2012: 17–35). Raptexte als „hip-hop's major means of verbal expression" (Androutsopoulos & Scholz 2002: 2) sind dabei, obwohl *Rap* häufig mit *Hip-Hop* synonym gebraucht wird, historisch tatsächlich als letztes dieser vier Elemente in der Hip-Hop-Kultur entstanden (vgl. Verlan 2003: 140). Dabei liegen die Ursprünge der Hip-Hop-Kultur in den USA, wodurch zunächst auch in Deutschland bis Ende der 1980er Jahre Rap fast ausschließlich englischsprachig auftritt.[4] Auch wenn sich dann allmählich eine ‚deutsche' Hip-Hop-Kultur herausgebildet hat, blieb die Sprache der ‚Mutterkultur' nach wie vor relevant. Zudem findet sich das musikalisch konstitutive Element des Samplings auch in Raptexten wieder: „Ständig werden Old School-Praktiken verdrängt, um als Zitat wieder aufzutauchen, kanonisiert und zugleich rekontextualisiert" (Streeck 2002: 538). Damit ist der (kulturelle) Verweis bzw. die Referenz ein wesentliches, auch durch Sprachwahl ausdrückbares Merkmal des Raps – hierzu mehr in den Abschnitten 5.2 und 4.1. Insofern ist dieser gekoppelte Kultur- und Sprachkontakt, also die Entlehnung über pop- und hier spezifischer musikkulturelle Güter speziell aus dem US-amerikanischen Raum, repräsentativ für das Aufkommen von Anglizismen jüngerer Vergangenheit (vgl. Busse 2008: 39). Warum ist nun gerade der Vergleich von OLD_SCHOOL und ALTE SCHULE im Kontext des Hip-Hop-Diskurses interessant? Wenngleich OLD_SCHOOL in Wörterbüchern sowohl adjektivisch (attributiv, adverbial, prädikativ) als auch substantivisch verzeichnet ist (vgl. Duden 2020), ist in den einschlägigen Anglizismen-Wörterbüchern (Dictionary of European Anglicisms [DEA], Anglizismen-Wörterbuch [AWb]) wie auch in etymologischen Wörterbüchern für das Deutsche (Kluge/Seebold 2011) kein Eintrag zu weiteren Hintergründen zu finden.[5] In der deutschsprachigen Version der Online-Enzyklopädie Wikipedia [6] wird die Verbreitung des Begriffs vor allem der Musikrichtung Hip-Hop zugeschrieben (Abruf 14.09.22). Dies geschieht jedoch ohne weitere Quellenangaben und die Diskussionsseite zum Artikel gibt Einblick in die Umstrittenheit dieser These. Dennoch: Sowohl die Ergebnisse der durchgeführten Korpusanalyse als auch andere Quellen[7] konturieren OLD_SCHOOL / ALTE SCHULE als relevantes Diskurselement und plausibilisieren die Untersuchung gerade in diesem Korpus. Auch Androutsopoulos und Scholz verweisen auf

[4] Für einen historischen Überblick die Entwicklung in Deutschlande siehe Verlan 2003 und Güler Saied 2012: 55–114.

[5] In DEA und Kluge/Seebold 2011 werden nur *old-timer* bzw. *Oldtimer* thematisiert. Die Recherche in zahlreichen Musik-Lexika war ebenfalls erfolglos.

[6] https://de.wikipedia.org/wiki/Old_School

[7] Bspw. trug ein Sampler von MZEE Records den Namen *Alte Schule*, auf dem sich zu diesem Diskurselement positioniert wird (vgl. Margara 2018: 6).

die Verankerung von OLD_SCHOOL im entsprechenden Diskurs: „Hip-hop culture generally involves a kind of ‚generational conflict', the poles of which are often referred to as ‚old school' versus ‚new school'." (2003: 491) Daher kann in Bezug auf die Domäne Hip-Hop womöglich von ‚echten' funktionalen Äquivalenten ausgegangen werden, was bei vielen Vergleichen mit bemühten Übersetzungen von Anglizismen nicht der Fall ist (vgl. Eichinger 2008: 77f.) – somit werden gegebenenfalls andere Vergleichsdimensionen zugänglich.

3 Einordnungen

Zur Verortung dieser Untersuchung gilt es zunächst, die zentralen Termini zu konturieren. Dies betrifft den zugrunde gelegten Anglizismen-Begriff, die theoretische Verortung der Einbindung dieser Anglizismen sowie die Benennung der untersuchten Einheiten.

3.1 Anglizismus

In der sprachwissenschaftlichen Auseinandersetzung mit den Konsequenzen des Sprachkontakts mit dem Englischen wurden zahlreiche Klassifikationen und Definitionen entwickelt, was sich bereits am Umfang der klassifikatorischen terminologischen Einordnungsbemühungen bei Arbeiten zu Anglizismen zeigt (vgl. u. A. Altleitner 2007; Busse 1993; Eisenberg 2018; Fiedler 2014; Langner 1995; Onysko 2007; Peycheva 2014; Siekmeyer 2007; Spitzmüller 2005; Wetzler 2006; Yang 1990). So stellt bspw. Wetzler fest, dass die „terminologische Einordnung englisch-deutscher Entlehnungsvorgänge […] immer wieder Thema der Forschung [war]" (2006: 45). Hinsichtlich des Begriffs *Anglizismus* merkt Altleitner an, dass es in „der einschlägigen Literatur zur Entlehnung […] keine durchgehend einheitliche Terminologie zu diesem Begriff [gibt]" (2007: 29) und Fiedler weist darauf hin, dass es „[s]elbst bei Arbeiten zu lexikalischen Entlehnungen […] Unterschiede hinsichtlich des untersuchten Materials in Abhängigkeit von der zugrunde gelegten Definition von Anglizismus und der Einordnung in Klassifikationsmodelle [gibt]" (2014: 31). Eine umfassende Positionierung in diesem Kontext ist nicht das Ziel dieser Untersuchung,[8] es wird lediglich an geeigneten Stellen auf diesen klassifikatorisch-terminologischen Diskurs Bezug genommen. Zunächst ist hervorzuheben, dass sich Fremdwörter im Verständnis dieser Untersuchung durch ihre vom nativen System abweichenden formalen Eigenschaften auszeichnen, was sie für Mitglieder einer Sprachgemeinschaft als solche erkennbar macht. Die Herkunft ist damit nicht entscheidend, sondern nur „synchron erkennbare Fremdheitsmerkmale" (Seiffert 2005: 221). Dies hat den Vorteil, dass die Unterscheidung ‚echter' Fremdwörter und sogenannter ‚Pseudofremdwörter' (bspw. *Dressman*) entfällt, der Nachvollzug ist meist ohnehin schwierig und „für das Deutsche selbst bleibt das […] meistens bedeutungslos" (Eisenberg 2018: 29). Für *Anglizismus* wird die offene Definition Eisenbergs als Arbeitsgrundlage herangezogen:

> Fremdwörter sind Wörter des Deutschen, auch wenn sie ganz oder teilweise aus anderen Sprachen übernommen sind. Ein Fremdwort aus dem Englischen bezeichnet man als Anglizismus

[8] Für einen umfassenden historisch-quantitativen Überblick siehe Spitzmüller 2005: 161–182.

> und bringt damit zum Ausdruck, dass es sich nicht um ein Wort des Englischen handelt, sondern um eines, das ganz oder in Teilen aus dem Englischen stammt. (2013: 2)

Im Untersuchungskorpus liegt ein Anglizismus demnach nur dann vor, wenn er in einen deutschen Kontext integriert ist und auch nur dann wird er berücksichtigt. Zur Schlüssigkeit sei das „teilweise" noch dahingehend erweitert, dass formale Eigenschaften des Englischen bereits genügen. Dieser Fremdwortdefinition entsprechend wird auch nicht nach der genauen Herkunft differenziert. U. A. Busse folgend

> „wird *Anglizismus* […] als Oberbegriff für alle sprachlichen Beeinflussungen aus dem angloamerikanischen Sprachraum aufgefasst, egal ob sie aus Großbritannien, den USA, Kanada usw. stammen oder nur mittelbar darauf zurück gehen." (2008: 41 f.: Hervorhebung im Original)

Ob nun im Deutschen im Rahmen von Analogiebildung produziert oder tatsächlich aus dem Englischen stammend, im Resultat sind beide Wege als ‚sprachliche Beeinflussung' des Englischen interpretierbar.

3.2 Sprachmischung

In Eisenbergs angeführter Definition werden Anglizismen als „Wörter des Deutschen" bestimmt. Gerade im Kontext von intendierter Sprachvermischung bei Raptexten als Illustration kreativer Sprachvirtuosität – auch hierzu später mehr – stellt sich dabei die Frage, die auch Eisenberg aufwirft, wenn er Fremdwörter von Zitatwörtern unterscheidet (vgl. 2018: 3): Wird OLD_SCHOOL in den Raptexten ‚im Deutschen' verwendet oder findet ein Wechsel ins Englische statt?[9] Dieser Problematik widmet sich auch Onysko in seiner Untersuchung zu Anglizismen im Nachrichtenmagazin Der Spiegel ausführlicher und bringt das Konzept des ‚Codeswitchings' ein. In Auseinandersetzung mit diversen bestehenden Differenzierungsansätzen unterscheidet er zwischen „multi-element syntactic units (codeswitches)" und „single lexical items (e. g. borrowings, compounds, derivations)" (2007 :38). Das Problem dieser Grenzziehung liegt für OLD_SCHOOL auf der Hand: Strukturell könnte es zu komplex sein. Anders als das Schriftsystem des Deutschen markiert das Englische Komposita nicht konsequent durch Zusammenschreibung, zumal dies nicht mal ausschlaggebend sein muss. Die Frage, ob es sich um ein Kompositum oder eine syntaktische Phrase handelt, ist nicht eindeutig zu klären (vgl. Eisenberg 2013: 318). Auch Onysko stellt im Verlauf seiner Untersuchung fest: „As yet, the criteria to differentiate borrowing from codeswitching postulated in research have failed to provide a clear delineation of two types of language influence" (2007: 272). Angesichts seiner Untersuchungsdaten weist er darauf hin, dass auch *Single Lex-*

[9] Vgl. hierzu auch Spitzmüller 2005: 173–176.

ical Items Codeswitching sein können und *Multi-Element Syntactic Units* entsprechend *Borrowings* [10] (vgl. ebd. 272–286). Verbindungen aus Adjektiv und Substantiv (wie OLD_SCHOOL) werden jedoch von ihm explizit als häufigstes Muster phrasaler Anglizismen angeführt und strukturell für den Übernahmeprozess verortet: „Such constructions are syntactic groups in English and, as anglicisms, compound nouns in German." (ebd. 282) Was potenzielle Erweiterungen betrifft, muss vermutlich im Einzelfall entschieden werden. Daneben diskutiert Androutsopoulos (2003) aus soziolinguistischer Perspektive in diesem Zusammenhang *(Language) Crossing* (Rampton 1998) und verortet es dabei im Zusammenhang mit weiteren Konzepten zur Erfassung von Sprachmischung wie *Double Voicing* (Bakhtin 1971) und *Bricolage* (Schlobinski/Kohl/Ludewigt 1993).[11] Vom *Codeswitching* unterscheide sich das *Crossing* im Wesentlichen dadurch, dass hierbei klar sei, „dass sich der Sprecher nicht seiner ‚eigenen' Sprachen, sondern eines fremden Codes bedient" (Androutsopoulos 2003: 90) und dies mit einer „latente[n] politische[n] Relevanz" einhergehe (ebd.: 89). Er stellt heraus: „Kreuzungen verweisen stets auf (stereotypische) Werte und Eigenschaften der ethnischen Gruppen, denen die Sprache oder Varietät eigen ist." (ebd.) So erscheint dies ein besonders passender theoretischer Anknüpfungspunkt für die Analyse von Raptexten als Teile des Hip-Hop-Diskurses zu sein. Relevanter als die genaue Zuordnung zu einem dieser Konzepte ist für die Untersuchung der Verwendung von OLD_SCHOOL in Deutschraptexten der bei diesen Konzepten stärkere Einbezug sozialer Identität, die über die Verwendung von Einheiten anderer Sprachen spezifisch modelliert werde (vgl. Androutsopoulos 2003: 93). Dass dabei „‚Prestige' ein wichtiges Stichwort in der soziolinguistischen Diskussion um Crossing" (ebd.: 92) ist und auch bei der Frage nach Gründen für die Verwendung speziell von Anglizismen häufig angeführt wird (vgl. bspw. Altleitner 2007: 166), veranschaulicht diese zu berücksichtigende soziolinguistische bzw. pragmatische Dimension und wird später eingehender erläutert.

3.3 Graphematische Variation und Benennung der Untersuchungseinheiten

Das zuvor bereits angedeutete Problem der Getrennt- und Zusammenschreibung für die grammatische Interpretation der untersuchten Einheiten führt zwangsläufig auch zur Frage der weiteren Benennung. Onysko bezeichnet analoge Einheiten zu OLD_SCHOOL (bspw. *bad guy*, *happy end*, *political correctness*) im Englischen – wie bereits erwähnt – zunächst als syntaktische Gruppen und für das Deutsche als Substantivkomposita (vgl. Onysko 2007: 282). Dies geschieht unter der Kapitelüberschrift „phrasal anglicisms" (ebd.), was für das Deutsche ebenfalls phrasalen Status suggeriert, bezeichnet *Anglizismus* doch wie zuvor definiert ursprünglich Englisches als Bestandteil des Deutschen. Das bei der Grenzziehung zwischen *Borrowing* und *Codeswitching* genutzte und problematisierte strukturelle Kriterium (*Single*

[10] Aufgrund der bereits angedeuteten ohnehin reichhaltigen terminologischen Diskurslage wird auf eine Übersetzung sicherheitshalber verzichtet und nur gemäßigt an die deutsche Orthografie angeglichen.

[11] Für einen vergleichenden Überblick dieser Konzepte siehe Androutsopoulos 2003: 84–93.

Lexical Item vs. *Multi-Element Syntactic Unit*) zeigt sich hier nochmals besonders nachdrücklich: Mit den für das Englische kodifizierten Schreibweisen <old school> und <old-school> deutet sich an, was sich für das Deutsche als protypisches Problem für solche Fälle herausstellt, da sie bereits im Englischen unterschiedlich kodifiziert werden (vgl. Langner 1995: 168). Augst bemerkt hierzu: „Wenn es aber, wie im Englischen, generell drei gleichberechtigte Schreibmöglichkeiten für Komposita gibt – Getrenntschreibung, Bindestrichschreibung, Zusammenschreibung –, so ist dies mit dem generellen deutschen System inkompatibel, das nur die Bindestrich- und die Zusammenschreibung kennt." (1992: 46) Für das Deutsche finden sich im Duden die Schreibweise <oldschool> für das Adjektiv (eine verdeutlichende Schreibung mit Bindestrich ist prinzipiell auch möglich) und die Schreibweisen <Old School> und <Oldschool> für das Substantiv (vgl. Duden 2020: 836) (auch hier ist die Schreibung mit Bindestrich prinzipiell möglich). Gewonnen ist damit für die syntaktische/morphologische Interpretation im Deutschen für das vorliegende Korpus allerdings wenig, da diese Formen im Zuge der Nutzung im Deutschen ggf. bewusst unverändert als Zitatwörter getrenntgeschrieben werden. Es kann hier nicht das Anliegen sein, die bestehende Forschung zur Getrennt- und Zusammenschreibung für den vorliegenden Phänomenbereich vollständig zu diskutieren. Einen kurzen aktuellen Forschungsüberblick gibt Fuhrhop (2022: 107–128).[12] Hier geht es darum, dies soweit zu problematisieren, wie es für die Benennung der behandelten Einheiten notwendig ist. Wesentlich für OLD_SCHOOL ist, dass „die Getrennt- und Zusammenschreibung über zwei grundlegende Prinzipien geregelt [ist], einem morphologischen (das Wortbildungsprinzip) und einem syntaktischen Prinzip (das Relationsprinzip)" (ebd.: 107). Anhand des Beispielpaars *er isst Schweinebraten – er wird Schweine braten* verdeutlicht Fuhrhop die Kontextrelevanz dafür, ob es sich um ein Wort bzw. Kompositum oder ein Syntagma handelt (vgl. ebd.). Die Prinzipien fasst sie wie folgt zusammen:

> **Das Wortbildungsprinzip:** „Verbindungen" aus zwei oder mehr Stämmen werden zusammengeschrieben, wenn sie aufgrund einer Wortbildung miteinander verbunden sind.
> **Das Relationsprinzip:** Einheiten, die syntaktisch analysierbar sind, das heißt insbesondere die, die in syntaktischer Relation zu anderen Einheiten in einem Satz stehen, sind syntaktisch selbstständig und werden entsprechend syntaktisch vollständig geschrieben (umgeben von Spatien). (ebd.: 114; Hervorhebung im Original)

Für <*Oldschool*> – <*Old School*> trifft nun das zu, was Fuhrhop als „potentiell wortfähig" beschreibt, sie „bestehen potentiell aus dem gleichen Material" (ebd.). An sich stelle die Komposition den unproblematischen Teil der Getrennt- und Zusammenschreibung dar, da dies an der Form erkennbar oder syntaktisch entzifferbar sei. (vgl. ebd.: 116) Warum dies für die Untersuchungseinheit nicht der Fall ist, wird im Vergleich zum nativen ‚Pendant' ersichtlich: Warum stellt sich die Frage nach der Getrennt- und Zusammenschreibung bei *Alte Schule* nicht? Hier weist das Flexionssuffix des Adjektivs auf dessen Verwendung als Attribut hin, es ist syntaktisch analysierbar und wird deshalb nach dem Relationsprinzip

[12] Die wesentlichen ausführlichen Monografien hierzu sind Jacobs 2005 und Fuhrhop 2007.

getrenntgeschrieben.[13] Im Englischen aber gibt es eine solche syntagmenmarkierende Flexion des attributiven Adjektivs nicht, was nach Zifonun auch einer der Gründe sei, warum die „Prämodifikation im Englischen besonders ‚stark' ausgebildet" ist (2010: 179). Bei der Diskussion um die Verwendungsgründe für Anglizismen wird auf diese Einsicht später zurückgegriffen werden. Das Problem fehlender Adjektivflexion bei der Unterscheidung von Kompositum und Syntagma ist auch für den nativen Bereich identifiziert worden, nämlich bei Stadtadjektiven, was Fuhrhop u. A. an *Schweizer Käse* – *Schweizerkäse* verdeutlicht (vgl. Fuhrhop 2007: 126–128).[14] Sie stellt diesbezüglich fest: „Der formale Unterschied zwischen *Schweizerkäse* und *Schweizer Käse* ist lediglich der Akzent bzw. die Zusammenschreibung." (ebd.: 126) Das hier angeführte graphematische Kriterium der Zusammenschreibung für Anglizismen generell und für das vorliegende Korpus speziell wurde bereits problematisiert, auf das intonatorische besteht kein Zugriff bzw. die Analyse wäre zu aufwendig. Über diese formalen Kriterien hinaus führt sie aber auch ‚Begriffsbildung' an:

> Allerdings finden wir auch andere Begriffsbildungen wie *Schwarzes Brett*. Hier wird im Zuge der Rechtschreibreform über die Groß- und Kleinschreibung debattiert. Es wird aber nicht debattiert, ob *Schwarzes Brett* / *schwarzes Brett* zusammengeschrieben werden soll und damit als ein Wort ausgegeben wird. Das wäre auch absurd, wegen der Flexion kann diese Verbindung kein Wort sein. (ebd.: 127)

Wie bereits erwähnt fehlt diese Flexion im Englischen, prinzipiell ordnet sie Fälle wie OLD_SCHOOL aber „wie *Schwarzes Brett* als ‚feste Begriffe'" ein (Fuhrhop 2022: 160). In der einschlägigen Monografie Müllers zur Groß- und Kleinschreibung im Deutschen findet sich eine Auseinandersetzung mit solchen Fällen unter der Bezeichnung ‚feste Verbindungen' (vgl. 2016: 139–142), womit er der Terminologie des amtlichen Regelwerks zur deutschen Rechtschreibung folgt (§ 63). Schwerpunktmäßig diskutiert Müller die festen Verbindungen aus diskursiver Perspektive und weist auf die Nähe zu Eigennamen hin:

> Die Ursache für die Ausweitung des Eigennamenkonzeptes auf ‚feste Verbindungen' liegt im Bedürfnis der Sprachgemeinschaft zur Bildung komplexer, auf spezielle Sachverhalte bezogener Terme, die häufig aus Adjektiv-Nomen-Verbindungen bestehen und durch Majuskelschreibung gekennzeichnet werden. […] Die Verwendung der Majuskel bei Adjektiv-Nomen-Verbindungen lässt sich als Hinweis an den Leser interpretieren, das Adjektiv nicht als

[13] Adjektive mit Flexionssuffix innerhalb von Komposita tauchen m. W. n. nur in mindestens dreigliedrigen Komposita auf, in denen ein entsprechendes Syntagma als Determinans integriert ist und dann mit Bindestrichen im Rahmen einer sogenannten ‚Durchkopplung' realisiert wird (bspw. <Rote-Kreuz-Schwester>). Die einzig mir bekannte Ausnahme ist die in Duden angegebene Schreibung <Sauregurkenzeit> (vgl. Duden 2020: 985).

[14] Angeführt wird dies hier nur bezogen auf Substantivkomposita. Es sei darauf hingewiesen, dass dieses Problem auch für Adjektivkomposita identifiziert worden ist, hier aber anhand verschiedener Tests erfolgreicher gelöst werden konnte (vgl. Fuhrhop 2007: 104–109).

> Attribut misszuverstehen und demnach den Gesamtkomplex und nicht nur das Nomen als Diskursinstanz zu betrachten. (2016: 140–142)

Mit dem Hinweis auf im Rahmen des Erwerbs der Rechtschreibkompetenz dokumentierte Fehlschreibungen wie <Rotesmeer> oder <Sächsischeschweiz> weist er dabei wiederum auf den bei Fuhrhop herausgearbeiteten Bezug zur Getrennt- und Zusammenschreibung hin (vgl. ebd.: 141). Er kommt zu dem Schluss, dass

> „[d]er Diskurs und nicht das Lexikon […] die Ebene [bildet], auf der eine sinnvolle Entscheidung über den Majuskelgebrauch getroffen werden kann. […] Während die Verwendung des *Gelben Trikots* als Begriff Texten des Themas ‚Profiradsport' vorbehalten bleibt, können Bekleidungsfachgeschäfte nach wie vor nur *gelbe Trikots* verkaufen." (ebd.)

Was hier für den Profiradsport und *Gelbes Trikot* erläutert wird, kann angesichts der zuvor dargestellten Einbettung im Hip-Hop-Diskurs für OLD_SCHOOL auch angenommen werden, in der Regel kann also von einer ‚festen Verbindung' bzw. einem ‚festen Begriff' ausgegangen werden. Für die spätere Analyse wird dennoch einzeln beurteilt werden, ob nicht ggf. doch ein nicht-terminologischer Gebrauch wie in *she goes to a very old school* vorliegt. Dass es hier schwerfällt, eine entsprechende Einbettung von OLD_SCHOOL im Deutschen zu konstruieren, deutet bereits auf einen terminologischen Vorteil hin. Allein die Eigenschaft ‚fremdsprachig' markiert im Kontrast zum nativen Kontext die enge Zusammengehörigkeit von englischem Adjektiv und Substantiv und rückt so beides, wie das fehlende Flexionssuffix des Adjektivs, für die Rezeption näher zusammen. Über diese ‚Fremdsprachenklammer' werden vermutlich auch noch stärker syntaktisch analysierbare Fälle wie *Meet and Greet* ohne weiteren Diskursüberblick im Kontrast zum nativen Kontext insgesamt terminologisch interpretiert. Damit ist die Eigenschaft ‚fremdsprachig' funktional bei den von Buchmann angeführten ‚linear-suprasegmentalen Mitteln' einordbar. Bezüglich der Bindestrichschreibung stellt sie fest:

> Der Bindestrich kann […] hier beispielsweise durch eine Kursivsetzung <*Rund um die Uhr*-Bereitschaft> oder eine andere Schrifttype <Rund um die Uhr-Bereitschaft [ersetzt werden]. Wie auch schon für die durch Anführungszeichen graphisch markierten Schreibungen gezeigt […], ist der durchkoppelnde Bindestrich im Syntagma fakultativ und kann durch linear-suprasegmentale Mittel ersetzt werden. Der Bindestrich vor dem Determinatum hingegen ist obligatorisch" (2015: 257)

Dies trifft natürlich nur für den von Buchmüller untersuchten Schreibgebrauch zu, nicht für die aktuell gültige amtliche Regelung der deutschen Rechtschreibung. Hier findet sich:

> **§44** Man setzt einen Bindestrich zwischen allen Bestandteilen mehrteiliger Zusammensetzungen, in denen eine Wortgruppe oder eine Zusammensetzung mit Bindestrich auftritt, sowie in unübersichtlichen Zusammensetzungen aus gleichrangigen, nebengeordneten Adjektiven.

Bezogen auf die Getrennt- und Zusammenschreibung bei OLD_SCHOOL wird die prinzipielle Offenheit zwischen Syntagma und Kompositum durch das fehlende Flexionssuffix

beim Adjektiv bei solchen ‚Bildungen' im amtlichen Regelwerk insofern beibehalten, als dass hier das zuvor erwähnte intonatorische Kriterium entscheidend für die Option zur Zusammenschreibung wirkt:

> **§ 37 E4** Aus dem Englischen stammende Bildungen aus Adjektiv + Substantiv können zusammengeschrieben werden, wenn der Hauptakzent auf dem ersten Bestandteil liegt, also *Hotdog* oder *Hot Dog*, *Softdrink* oder *Soft Drink*, aber nur *High Society*, *Electronic Banking* oder *New Economy*.

Zeigt sich also die für zweisilbige Komposita im Deutschen typische trochäische Betonung (vgl. Eisenberg 2013: 141), kann die hier für das deutsche Kompositum vorgesehene Zusammenschreibung realisiert werden, für alle anderen Fälle ist dies blockiert. Damit kann insgesamt festgestellt werden, dass OLD_SCHOOL für die voraussichtlich meisten Fälle im vorliegenden Korpus als ‚feste Verbindung' bzw. ‚fester Begriff' zu vermuten ist. Der Getrennt- und Zusammenschreibung wird für die Analyse keine Bedeutung für die Unterscheidung zwischen Kompositum und Syntagma beigemessen, was wie gezeigt sowohl aus nativer Sicht als auch als Konsequenz der Eigenschaft ‚fremdsprachig' (durch die zusätzliche Möglichkeit der Getrenntschreibung im Englischen) durch Transferenzschreibungen plausibilisiert werden kann. Die Großschreibung wird für die Identifizierung von Substantiven als recht sicher angenommen, die Kleinschreibung aber schließt eine Einstufung als Substantiv nicht aus, da auch dies durch die Eigenschaft ‚fremdsprachig' bedingt sein könnte. Um in der weiteren Benennung die hinsichtlich der Unterscheidung von Kompositum und Syntagma möglicherweise tendenziösen Termini ‚Verbindung' und ‚Begriff' zu vermeiden und auch den Fall von Zusammenschreibungen, die natürlich kein Syntagma sein können sowie das native ‚Pendant' ALTE SCHULE einzuschließen, wird der Terminologie Fiedlers gefolgt und *Nomination* zur Benennung genutzt (vgl. Fiedler 2014: 41). Hier führt sie die Beispiele „*Hot Dog; Blind Date; Bad Bank; fauler Kredit*" (ebd.) an, bei denen sich das vorliegende Vergleichspaar gut einordnen lässt. Dies ist zudem auch etwas spezifischer als bspw. die Bezeichnung *phraseologische Termini*, unter denen Burger diese neben anderen beschreibt (vgl. 2015: 50), knüpft gleichzeitig aber an die hier relevante fachsprachliche/terminologische Verwendung an, ohne die die Großschreibung des Adjektivs orthografisch ja gar nicht möglich wäre (vgl. § 63(2.2)). Auch Burger beschreibt:

> Das Besondere dieser Gruppe von Ausdrücken besteht darin, dass sie genauso funktionieren wie jeder (Wort-)Terminus. Das heißt, sie sind in ihrer Bedeutung weitestgehend festgelegt („normiert"), und diese Festlegung gilt primär nur innerhalb des fachlichen Subsystems der Sprache. (2015: 50)

Dass es im Untersuchungskorpus Belege mit Großschreibung des Adjektivs auch der nativen Nomination ALTE SCHULE gibt, verweist auf den terminologischen Gebrauch und damit wiederum auch auf die ‚Fachsprache' Hip-Hop. Zudem sind diese Nominationen besonders typisch für fachsprachliche Kommunikation, da „mit dieser für das Deutsche […] charakteristischen grammatischen Konstruktionsweise dem erhöhten Benennungsbedarf im Rahmen fachlicher Kommunikation leicht Genüge getan werden [kann]." (Roelcke 2020: 112) Egal ob

Kompositum oder Syntagma, „Komposita (und hier vor allem Determinativkomposita) gestatten (wie auch Mehrwortbenennungen) eine ausdrückliche Spezifikation von Bezeichnungen auf der Wortebene." (ebd.) Auch die hier von Roelcke für die Komposita angeführte häufige Bildung von Antonymen lässt sich im Untersuchungskorpus mit NEW_SCHOOL für die anglizistische Nomination gut belegen und stützt diese Bezeichnungsentscheidung. Schließlich beinhaltet der Nominationsbegriff auch die kommunikativ-pragmatische Definition, die Fleischer (1996: 150) nach Bellmann als „Referenz plus Pragmatik" (1989: 31) definiert, was vor allem im Hinblick auf den Vergleich der beiden Nominationen im pragmatischen Gebrauch sehr gut zu passen scheint.

4 Verwendungsgründe für Anglizismen

Um den Nutzen von Anglizismen in Deutschraptexten angemessen nachvollziehen zu können, ist ein kurzer Überblick über bestehende Hypothesen hierzu sinnvoll.

4.1 Bestehende Ansätze

Das mit dem Gebrauch von Anglizismen postuliert transportierte Prestige ist eine der klassischsten Vermutungen zu den Gebrauchsgründen und findet sich bspw. bei Fiedler (2014). In der aktuellen Forschung findet dies zunehmend weniger Beachtung und wird aus sprachpuristischer Perspektive zudem häufig als Kritikpunkt funktionalisiert (vgl. Spitzmüller 2005: 281–289). Als ebenfalls abnehmend angeführt werden kann das sogenannte Lokalkolorit, also die subtile Kommunikation einer gesamtkulturellen ‚Stimmung' – Wenngleich dies bei der Betrachtung Hip-Hop-spezifischer Gründe in ähnlicher Form erneut aufgegriffen wird. Beispiele und Erläuterungen hierzu finden sich etwa in den Untersuchungen von Yang (1990) oder Meder (2005). Ebenfalls stilistisch kann das Argument der Ausdrucksvariation interpretiert werden, das sich neben Yang und Meder u. a. auch bei Altleitner (2007) findet. Das wohl präsenteste Argument der Sprachökonomie von Anglizismen findet sich in fast allen Arbeiten, die sich mit Verwendungsgründen auseinandersetzen, wenngleich es unterschiedlich zugeschnitten wird. So wird teilweise eher auf die Präzision Bezug genommen, die sich mit anderen angeführten Erklärungen wie der Bedeutungsdifferenzierung oder der Auseinandersetzung mit Bezeichnungslücken (vgl. Altleitner 2007: 162 f.; Fiedler 2014: 40) zusammenbringen lässt. Fremdwörter böten sich hier auch deshalb an, weil sie semantisch nicht ‚vorbelastet' seien und somit problemlos für solche Zwecke genutzt werden können. Gerade für Fachtermini hebt dies Altleitner nochmal hervor (2007: 155). Zudem wird auch häufig auf die ‚Kürze' Bezug genommen, so verweist sie (ebd.: 156) bspw. auf eine ältere Arbeit Pfitzners (1978), in der die große Anzahl oft einsilbiger kurzer Wörter des Englischen angeführt wird, welche diese aus sprachökonomischer Sicht vielleicht eher als ihre deutschsprachigen Äquivalente für den Gebrauch qualifiziere. Bei ihrer Untersuchung anhand des Nachrichtenmagazins *Der Spiegel* geht Yang mit Blick auf bestehende Forschungsliteratur sogar soweit, diesen Aspekt „als die wichtigste Entlehnungsmotivation für die Anglizismen" (1990: 123) herauszustellen. Dieses aus der Forschungsliteratur entnommene Fazit zieht sie auch angesichts ihrer eigenen Untersuchungsergebnisse gerade für den journalistischen Bereich (vgl.

ebd.: 124). Für das vorliegende Vergleichspaar OLD_SCHOOL / ALTE SCHULE besonders interessant ist das von Altleitner vorgebrachte Argument der Unübersetzbarkeit, bei dem sie sich auf Carstensen (1965; 1984), Adorno (1979) und Fink (1979) bezieht (vgl. Altleitner 2007: 160–162). Als wesentliches Hindernis der Übersetzbarkeit wird neben der formalen Äquivalenz vor allem die semantische angeführt, dass also der Übersetzungsversuch „in allen denkbaren Kommunikationssituationen dasselbe bedeutet wie der fremdsprachliche Ausdruck" (ebd.: 160. Spezifischer sprachintern verweist Fiedler schließlich noch auf die Sprachverwandtschaft des Deutschen und des Englischen, wobei sie sich auf Busse und Solms (2002) bezieht (vgl. 2014: 40).[15] Diese stellen die diachrone Entwicklung beginnend von der Frühzeit für das Deutsche und das Englische bis in die Gegenwart vergleichend dar und gelangen zum Schluss, dass „[t]ypologisch betrachtet […] sowohl das Englische als auch das Deutsche im Verlauf ihrer Geschichte die meisten Flexionsendungen verloren" haben (Busse & Solms 2002: 136). Synchron beschäftigt sich Barz mit dem Einfluss des Englischen auf die deutsche Wortbildung und gelangt zur ebenfalls sprachintern fokussierten Interpretation, dass sich zahlreiche Übereinstimmungen in den Grundmustern der Wortbildung beider Sprachen finden lassen, was „eine ideale Voraussetzung für eine Einflussnahme der englischen Wortbildung auf die deutsche" sei (2008: 47). Dabei stellt sie jedoch zur Diskussion, inwieweit es sich hierbei wirklich immer um einen Einfluss des Englischen auf das Deutsche handele, oder inwieweit auch Parallelentwicklungen in beiden Sprachen vorliegen könnten (vgl. ebd.).

4.2 Hip-Hop-spezifische Verwendungsgründe

In seiner Auseinandersetzung mit dem engen Verhältnis von Identität und Sprache im Rap für die Hip-Hop-Kultur beschreibt Streeck:

> Etwas strenger und enger könnte man Rap auch als eine evolvierende Diskursinstitution fassen, als Familie von Sprechakten, deren Beherrschung einem gestattet, an kulturellen (Re-)Konstruktionen von Wirklichkeit teilzuhaben. Rap ist das Ganze der Sprachspiele, mit denen sich eine spezifische, aber offene Gemeinschaft organisiert hat und gemeinsame Wirklichkeiten – geteiltes Bewusstsein – erzeugt. (2002: 538)

Um sich mit der Verwendung von Sprache im Hip-Hop-Kontext auseinanderzusetzen, ist das Bewusstsein über den gemeinschafts- wie genrekonstitutiven Aspekt der Sprache hierbei unabdingbar, da er auf das Individuum bezogen die zwangsläufige Verortung in diesem Geflecht indiziert, die mit seiner sprachlichen Teilhabe einhergeht. So wird sprachlich eine „Identitäten-Montage" (ebd.) vorgenommen, die das „Verhältnis zwischen sozialen Stilen des Sprechens, Gemeinschaft und Individuierung" (ebd.) zu navigieren hat. Mittel dieser diskursiven ‚Verortung' unterteilt Streeck dabei u. a. in *Representing*, *Dedication* und *Signifying*. Während *Representing* eher die lokale Zugehörigkeit ausdrückt, umfasst *Dedication* eher die soziale Zuordnung, indem hier bspw. Vorbilder benannt werden (vgl. ebd.: 543). Beides trägt

[15] Dieser Aspekt wird u. a. auch von Munske, gerade im Vergleich zum postulierten ‚Prestige' des Englischen, betont (2004: 166).

zur Authentizität der jeweiligen Individuen bei, die vielfach als wesentlichste Währung in der Hip-Hop-Kultur betrachtet wird und sich in Ausdrücken wie *Street Credibility, keeping it real* oder dem Vorwurf des *Sell-outs* niederschlägt. Das *Signifying* stellt neben den Authentizitätsaspekt im Rahmen lokaler/sozialer Verankerung von *Representing* und *Dedication* den Kompetenzaspekt und bezeichnet den Umgang mit Worten und ihren Bedeutungen in Form von (häufig diskreditierenden) Anspielungen usw., die bei konventioneller Interpretation harm- oder sinnlos sind und vom ‚Opfer' bestenfalls nicht verstanden werden (vgl. ebd. 544–546).[16] Ebenfalls zur Kompetenz gehört – bei Übersetzung wenig überraschend – der *Skill*, den Streeck als den „wichtigste[n], authentische[n] Identitätsbestandteil" beschreibt (ebd. 547) und als sprachliche Virtuosität paraphrasiert werden kann. Gemeint ist hier vor dem Hintergrund der Ursprünge im Battle-Rap und der improvisierten Freestyles „die Fähigkeit, spontan, aber kunstvoll auf unvorhersehbare sprachliche Stimuli zu reagieren" (ebd.: 546) Auf die Raptexte im Korpus bezogen bleibt der Aspekt sprachlicher Virtuosität als Kommunikation des eigenen *Skills* vorhanden, von Improvisationen wird nicht ausgegangen. Der eigene *Skill* kann schließlich auch über die in Abschnitt 3.2 beschriebenen Mittel der ‚Sprachmischung' signalisiert werden, wobei gleichzeitig die Indexikalität der anderen jeweils gewählten anderen Sprache im Kontext von *Representing* und *Dedication* genutzt werden kann, „they index social positioning" (Androutsopoulos 2007: 1). Bezogen auf englische Elemente in Raptexten anderer Sprache wie dem Deutschen fazitieren Androutsopoulos und Scholz:

> In conclusion, our classification and analysis show that English elements are an essential part of non English rap discourse. Some English elements have first and foremost a referential function with respect to the culture's major roles and activities. In addition, rappers can stylize themselves as 'underground' or 'subcultural experts' […]. Extensive English usage in rap songs is both a connection with rap's origins, and a demonstration of rappers' communicative skills. (2002: 25)

Vor diesem Hintergrund kann das in Abschnitt 4.1 angeführte Lokalkolorit im Hip-Hop-Kontext als Verwendungsgrund womöglich reaktiviert und erweitert auch als ‚Sozialkolorit' verstanden werden. Zusätzlich kann die kompetente Verwendung von Anglizismen zur *Skill*-Demonstration beitragen.

5 Dimensionen des Vergleichs

Ziel dieser Untersuchung ist es, durch den Vergleich des ‚Nutzens' i. S. d. Verwendung von OLD_SCHOOL und ALTE SCHULE Gemeinsamkeiten und vor allem Unterschiede zu identifizieren, um so zu Einsichten zum Nutzen der anglizistischen Nomination i. S. d. Mehrwerts zu gelangen. Dabei wird zwischen pragmatischer Verwendung im Hip-Hop-Diskurs und grammatischer bzw. morphologischer Verwendung unterschieden.

[16] Die ist nach Streeck auf die Kommunikationsgeschichte der Sklaverei zurückzuführen (2002: 545 f.).

5.1 Pragmatische Verwendung

Einen vielversprechenden Anknüpfungspunkt für die pragmatische Verwendung bietet die Untersuchung von Androutsopoulos und Scholz (2002). Die Autoren untersuchen hier anhand eines Korpus von 50 zufällig ausgewählten Raptexten pro Sprache den ‚Rekontextualisierungsprozess' in den jeweiligen Sprach- und damit Hip-Hop-Gemeinschaften für Frankreich, Deutschland, Spanien, Italien und Griechenland. Raptexte sehen sie dabei als „hip-hop's major means of verbal expression (other means of expression being writing, DJing and breakdancing), their analysis provides access to the study of hip-hop's appropriation in Europe." (ebd.: 2) Dies zeigt das Interesse am Gesamtdiskurs, wie auch in der Framework-Übersicht ersichtlich:

(i)	socio-cultural frame	(1)	social base of hip-hop culture in each country
		(2)	market and media infrastructure
(ii)	rap discourse	(1)	song topics
		(2)	**genre-typical verbal actions (speech act patterns)**
		(3)	**cultural references in rap songs**
(iii)	linguistic patterns	(1)	language variation
		(2)	**rhetorical patterns**
		(3)	English elements in non-English lyrics

Tabelle 1: Analyseframework zur sprachvergleichenden Analyse von Raptexten, nach Androutsopoulos & Scholz 2002: 4. Hervorhebung M. G.

Die Unterschiede zwischen dem Vergleichsinteresse der Autoren und der vorliegenden Untersuchung muss sich natürlich auch im Framework niederschlagen. So arbeitet die vorliegende Untersuchung nicht sprachvergleichend, sondern nur innerhalb einer Sprache. Zudem werden nicht ganze Raptexte miteinander verglichen, sondern lediglich die Verwendungen von OLD_SCHOOL und ALTE SCHULE innerhalb deutschsprachiger Raptexte als Gesamtheit. Damit entfällt der (i) *Socio-cultural-Frame* für die Analyse gänzlich, bei (ii) *Rap-Discourse* die *Song-Topics*. Von besonderem Interesse sind hingegen die *Speech-Act-Patterns* sowie die *Cultural References*. Bei den (iii) *Linguistic Patterns* ergibt sich (1) *Language-Variation* bereits aus dem Untersuchungsgegenstand der anglizistischen Nomination. Dies trifft offensichtlich auch für (3) *English Elements in non-english Lyrics* zu. Interessant ist hier noch die Kategorisierung, die Androutsopoulos und Scholz vornehmen. Von den dort angeführten Gruppen lässt sich OLD_SCHOOL nicht nur gut bei der (i) *Cultural Terminology* – den „culture specific key-words" (2002: 104) – zuordnen, sondern wird dort auch explizit genannt (vgl. ebd.). Die (2) *Rhetorical Patterns* werden bei der Analyse berücksichtigt. Nachfolgend werden die für die Analyse relevanten Kategorien erläutert.

5.1.1 Speech-Act-Patterns

Bei der sogenannten (1) *Self-referential Speech* und (2) *Listener-directed Speech* schränken Androutsopoulos und Scholz beides auf bestimmte Themenbereiche ein: So gehe es bei (1) vor allem um die eigene „verbal performance" (ebd.: 13) der Rappenden, bei (2) um die Wirkung dessen auf die Rezipierenden. (vgl. ebd.: 13 f.)[17] (3) *Boasting* und (4) *Dissing* verdeutlichen den typischen Wettbewerbscharakter beim Rap[18]: Während *Boasting* die eigene Kompetenz (spielerisch ironisierend) anpreist, geht es beim *Dissing* darum, dem (fiktiven) Gegner diese abzusprechen, darüber hinaus aber auch generell seine Daseinsberechtigung im Hip-Hop infrage zu stellen oder noch weiter zu gehen. Die Autoren fassen dies als „verbal attack and symbolic humiliation" (ebd.: 15) zusammen. Wie in Kapitel 5.2 beschrieben ist in Raptexten die kulturelle Verankerung der Rappenden im Rahmen der Authentizität relevant. (5) *Place/Time-References* drücken dabei diese Verankerung aus, „[t]aken together, place and time references emphasizes rap's reality grounding, its anchoring in real space and time." (ebd.) (6) *Identification (Naming)* meint die Selbstbenennung der Rappenden (vgl. ebd.: 16). Die hierbei zumeist genutzten Künstlernamen können dabei als Mittel der Identitätsschaffung verstanden werden.[19] Bei der (7) *Representation* drücken die Rappenden ihre Repräsentativität aus und verorten sich im Rahmen des Hip-Hop-Diskurses. Das muss sich nicht nur auf lokale Verortung (Herkunft) beziehen, sondern kann auch verschiedene Strömungen, wie bspw. *Oldschool*, zum Thema haben. Die Autoren verweisen hinsichtlich lexikalischer Ressourcen hierzu auf „(i) various equivalents of the verb to represent, (ii) […] the English verb as a loan-word, (iii) […] other circumlocutions." (ebd.) Besonders (ii) ist für die vorliegende Untersuchung offensichtlich interessant.

5.1.2 Cultural References

Kulturelle Referenzen dienen zum einen ebenfalls der Verankerungssignalisierung sowohl in Pop- bzw. Hip-Hop-kultureller Hinsicht als auch in Herkunftshinsicht. Zum anderen kann damit aber auch die kulturelle Kompetenz als Daseinsberechtigung im Hip-Hop-Diskurs kommuniziert werden, indem relevante Diskurselemente angeführt werden. So schildern Berns und Schlobinski bspw. einen Fall, bei dem der Radio-Moderator einer Hip-Hop-Sendung, Mister Hawkeye, die Anmaßung des Titels *MC* (Master of Ceremony) bei der Selbstbenennung vieler Anrufer kritisiert und diese als nicht gerechtfertigt darstellt, solange die Subkultur-Kompetenz nicht durch eine Live-Performance bewiesen sei (vgl. 2003: 210 f.). Kulturelle Referenzen zeigen also Subkulturkompetenz und Verankerung und führen dabei zu Intertextualität. Androutsopoulos und Scholz sehen als Mittel hierzu Eigennamen, wobei sie auch

[17] Für Beispiele für alle Speech-Act-Patterns aus unterschiedlichen Sprachen siehe Androutsopoulos und Scholz 2002: 13–18.

[18] Siehe hierzu auch Margara 2018.

[19] Aus konversationsanalytischer Perspektive weisen Androutsopoulos und Scholz darauf hin: „While self-referentiality of naming utterances fits the overall profile of rap discourse, they can also have a structuring function, initiating a new turn." (2002: 16)

Markennamen einschließen (vgl. Androutsopoulos & Scholz 2002: 18). Dabei unterscheiden sie drei wesentliche Referenzbereiche: *People (Personalities)*, *Brand-Names* und *Fiction*, die sie noch weiter differenzieren (vgl. ebd.: 19), was für diese Untersuchung aber nicht weiter relevant ist.

5.1.3 Rap-Rhetorics

Androutsopoulos und Scholz verweisen hierzu vor allem auf eine Untersuchung Potters (1995) zu entsprechenden Charakteristika von Raptexten und führen an:

- tropes, in particular metonyms and metaphors
- comparisons
- acronyms (e. g. N.W.A. for Niggas with Attitude)
- spelling of words, in particular proper names, e. g. German female rapper Sister S. pronounces her name as S.I.S.T.E.R.S.
- "homonymic slippage", i. e. puns and other (quasi-)homophone lexical relations.

(Androutsopoulos & Scholz 2002: 21)

Während sich die Autoren auf *Tropes* und *Comparisons* beschränken (vgl. ebd.), wird in der vorliegenden Untersuchung *Homonymic Slippage* berücksichtigt und *Antithesis* als neue Kategorie ergänzt. Letzteres ist ein Ergebnis der Datensichtung, da sich hier, wie am Ende von Abschnitt 3 als fachsprachliche Besonderheit angeführt, Gegensatzpaare zeigen. Im Rahmen der *Rap-Rhetoric* zeigen sich sowohl Belege für die direkte Gegenüberstellung von OLD_SCHOOL und NEW_SCHOOL (1a), als auch Beispiele im übertragenden Sinn generell für *alt – neu* im Zusammenhang mit OLD_SCHOOL (1b). Um möglichst sensitiv vorzugehen, wurden hier auch solche Belege hier zugeordnet, bei denen die Antithese formal zwar vorkommt, inhaltlich eine Zuordnung aber diskutabel wäre (1c).

(1)

a. Ey, was Oldschool oder Newschool? [das_traurigste_lied_der_welt]
b. Ich bin zwar neu in den Charts, doch Oldschool wie Methusalem [freund_oder_feind]
c. Oldschool, Newschool, ich pass' in keine Schublade [streng_geheim]

5.1.4 Weitere Aspekte

Zusätzlich wurde der Versuch unternommen, die allgemeine Konnotation (positiv/negativ) i. S. d. geäußerten Haltung gegenüber OLD_SCHOOL und ALTE SCHULE zu bestimmen, um die Verhandlung von und mit diesem Diskurselement besser einordnen zu können. Hier wurde bei Belegen ohne erkennbare Wertung „neutral" eingetragen, bei Zweifelsfällen wurde „unklar" gewählt. Für Song- bzw. Raptexte sind für die Wortwahl natürlich auch Reimaspekte und ‚Klang' relevant. Da Reimaspekte eine intensive Auseinandersetzung mit größeren Anteilen jedes einzelnen Textes erfordert hätten (Reimschemata gehen über Paarreime natürlich

weit hinaus) und der Klang als subjektives Entscheidungskriterium nicht unmittelbar zugänglich ist, wurde beides von dieser Untersuchung ausgeschlossen.

5.2 Syntaktische/morphologische Verwendung

Abschnitt 3.3 hat verdeutlicht, welche Interpretationsprobleme es im vorliegenden Korpus hinsichtlich der grammatischen Einordnung vor allem für die anglizistische Nomination gibt, da die Schreibung häufig wenig aufschlussreich ist. Hinzu kommt, dass Rap- bzw. Songtexten zum Teil konzeptionelle Mündlichkeit zugesprochen wird und in der Tat für gesprochene Sprache typische syntaktisch reduzierte Strukturen auftreten, die weniger Interpretationskontext anbieten. Androutsopoulos diskutiert die häufig formulierte Hypothese, dass dies insbesondere auch für die Jugendsprache gelte (vgl. 2006: 292–301), die wiederum dem sprachlichen Stil von Raptexten zugesprochen wird (vgl. Derecka 2021: 94 f.). Diese Schwierigkeiten finden folgende Konsequenzen in der korpuslinguistischen Operationalisierung der syntaktischen/morphologischen Verwendung: Die ursprünglich angestrebte Berücksichtigung der Wortart wurde nach den ersten Versuchen aufgegeben. Für die native Nomination ist dies ohnehin von weniger Interesse, für OLD_SCHOOL waren die Interpretationsmöglichkeiten zu uneindeutig. Die Beispiele (2a–f) sollen nochmals die generelle Abweichung für die Groß- und Kleinschreibung sowie Getrennt- und Zusammenschreibung bzw. Schreibung mit Bindestrich illustrieren, die in manchen Kontexten für die Disambiguierung hinsichtlich der Wortart bei mangelnden syntaktischen Indizien notwendig, hier aber nicht ausreichend systematisch sind. Somit ist in zahlreichen Fällen eine Entscheidung zwischen adjektivischer und substantivischer Verwendung nicht möglich und daher ebenso wenig darüber, ob OLD_SCHOOL bspw. als pränominales Adjektivattribut im Rahmen eines Syntagmas oder als Bestandteil des Kompositums eingeordnet werden kann.

(2)

a. Da ist ne old-school party mit break-dancern	[boogie]
b. Jetzt rapp ich auf den Oldschool shit	[fick_rap]
c. Ich bin der perfekte Oldschool-Klischee Rapper	[kleines_stueck]
d. Wenn ihr mich mit Old School Rap verbindet, liegt ihr daneben	[warteschild]
e. Trag schwarze old-school Vans und Pants zur Pressekonferenz	[neptun]
f. Wir feiern heute eine Old School Party	[old_school_party]

Für die syntaktische Analyse folgt daraus, dass die attributive Verwendung nicht berücksichtigt wird. Hieraus folgt für die Erfassung der Komposita mit OLD_SCHOOL wiederum, dass eine Wertung als Kompositum vorgenommen wird, sobald diese vor ein Substantiv treten. Hinsichtlich der Verarbeitung erscheint dies wie zuvor erläutert plausibel, da durch das fehlende Flexionssuffix ohnehin keine Interpretation als pränominales Adjektiv im Rahmen eines Syntagmas angestoßen wird. Zur prädikativen Verwendung sei noch angemerkt, dass hieraus klassischerweise ebenfalls nicht zwischen Adjektiv und Substantiv differenziert werden kann, beide sind regulär möglich.

6 Untersuchungskorpus

Das Hip-Hop-Subkorpus mit 5.919.774 Tokens wurde mithilfe des Konkordanz-Programms AntConc mit einem Kontext von 20 Tokens nach allen Belegen für ALT + SCHULE (68 Belege) und OLD + SCHOOL (434 Belege) durchsucht. Die Suchanfragen wurden unterspezifiziert formuliert, um für eine explorative Sichtung möglichst wenig a priori auszuschließen. Daher mussten zunächst die Belege von insgesamt drei systematischen Problembereichen ausgeschlossen werden: Wie in Abschnitt 2.2 beschrieben, geht es bei dieser Untersuchung um den Vergleich der Verwendung der nativen und anglizistischen Nomination. Letztere kann als solche nur eingeordnet werden, wenn sie ‚im Deutschen' verwendet wird. Fälle von *Codeswitching*, also dem gänzlichen Wechsel der Sprache, wurden daher ausgeschlossen (3a). Ebenso wurde wörtlicher, also nicht-terminologischer Gebrauch als mögliches Ausschlusskriterium vor allem für die native Nomination vorgesehen, tatsächlich lag ein solcher Gebrauch interessanterweise nirgends vor (3b). Schließlich, und dies ist generell bei Frequenzbetrachtungen innerhalb von Rap- bzw. Songtexten zu beachten, wurden *Hooks*, also sich wiederholende Textabschnitte (auch ‚Refrains') nur für den Erstbeleg berücksichtigt (3c).

(3)

a. The cars be oldschool	[am_riden]
b. Ich bin nun Lehrerin an meiner alten Schule	[selbst konstruiertes Beispiel]
c. Sie sagt, ich sei so oldschool, oldschool (o- oldschool.)	[501; insgesamt 5 Wiederholungen dieser Hook] [20]

Nach dieser Bereinigung ergab sich für ALTE SCHULE eine Gesamtbelegzahl von 59, für OLD_SCHOOL eine Gesamtbelegzahl von 192. Anschließend wurden die in Abschnitt 5 angeführten Vergleichsdimensionen manuell annotiert. Hierbei hat sich gezeigt, dass der zwecks Einheitlichkeit ursprünglich zur Interpretation vorgesehene Kontext nur eines (graphematischen) Satzes für viele Belege nicht ausreichend gewesen ist, vor allem auf die pragmatische Betrachtung bezogen. Über diese Ausweitung des Interpretationskontextes hinaus wurde punktuell auch der gesamte Raptext betrachtet, um eine möglichst valide Einschätzung treffen zu können und die Fälle unklarer Kategorisierung zu reduzieren. In einigen Fällen – gerade hinsichtlich der Konnotation, des *Boastings* und *Dissings* – war weiteres Diskurswissen notwendig.[21] Da diese Studie hinsichtlich der Anwendbarkeit bspw. der speech act patterns auf das Vergleichspaar explorativ angelegt worden ist, wäre eine zusätzliche Validierung über Interrater-Reliabilität wünschenswert.

7 Analyse und Ergebnisse

[20] Klammerungen werden bei der Verschriftlichung der Raptexte meist genutzt, um sogenannte Adlibs auszuzeichnen, also kurze, wiederkehrende Einwürfe. Für eine Beschreibung und Entwicklungsdarstellung für Deutschland siehe: https://www.br.de/puls/musik/vorbild-us-rap-adlibs-100.html (Abruf 05.01.23).

[21] Für die sachkundige Unterstützung bei der Interpretation danke ich Leonard Stoll.

Nachfolgend werden die Ergebnisse des Vergleichs nacheinander vorgestellt. Dabei werden die Ergebnisse, wie zuvor beschrieben, nach pragmatischer und syntaktischer/morphologischer Verwendung verglichen. Hierzu werden zumeist die prozentualen Anteile an der jeweiligen Gesamtbelegzahl zur Veranschaulichung genutzt, in abweichenden Fällen wird darauf hingewiesen. Im Sinne der Validität sei nochmals darauf verwiesen, dass sich die Anzahl der Okkurrenzen beider Nominationen signifikant unterscheidet. Dies kann als ein erstes Ergebnis interpretiert werden: OLD_SCHOOL wird in den untersuchten Raptexten mehr als dreimal so häufig wie ALTE_SCHULE – also deutlich präferiert – genutzt, was mitunter womöglich auf die in Abschnitt 2 dargelegte kulturelle Genese des Hip-Hops bzw. Raps in Deutschland zurückzuführen ist, aber auch den ‚Nutzen' der anglizistischen Nomination gegenüber der nativen indizieren könnte.

7.1 Pragmatische Verwendung

Hinsichtlich der **(i)** ***Speech-Act-Patterns*** lassen sich keine wesentlichen Unterschiede identifizieren. Am auffälligsten ist, dass ALTE SCHULE häufiger im Kontext von *Boasting*, also der ironisierenden ‚Selbstverherrlichung', genutzt wird (61,0%) als OLD_SCHOOL (47,9%). Ebenso wird die native Nomination häufiger im Zusammenhang mit *Place/Time-References* genutzt (27,1%) als die anglizistische (15,6%). Interessant ist zudem, dass beide wesentlich häufiger für oder im Kontext von *Boasting* als *Dissing* genutzt werden, was insgesamt auf eine tendenziell positive Konnotation im Hip-Hop-Diskurs verweist, wenngleich die Anteile vom *Dissing* anzeigen, dass dies keineswegs von allen Diskursteilnehmenden so geäußert wird (vgl. Abbildung 1).

So finden sich Belege, in denen sich in der eigenen Kontextualisierung klar für OLD_SCHOOL ausgesprochen wird (4a) und Belege, in denen sich negativ wertend abgegrenzt wird (4b). Am Ende von Abschnitt 7.1 wird dies nochmals systematisierend eingeordnet. Zudem wird an den Beispielen deutlich, dass die Verwendung im *Boasting* oder *Dissing* kein hinreichendes Kriterium zur Konnotationsbestimmung ist. Positive wie negative Bewertungen von OLD_SCHOOL finden sich sowohl beim *Boasting* (4a–b) als auch beim *Dissing* (4c–d), was auch auf ALTE SCHULE (4g–h) zutrifft. Dennoch lässt sich die Tendenz bestätigen (vgl. Tabelle 1). Beispiel (4e) verdeutlicht, dass bspw. *Boasting* auch ohne eine Bewertung von OLD_SCHOOL erfolgen kann. An (a) wird ersichtlich, dass *Boasting* und *Dissing* auch gleichzeitig auftreten können. An den Anteilen von Belegen, in denen *Self-referential Speech* vorkommt, wird aber ebenfalls deutlich, dass die Autoren der Raptexte größtenteils irgendeine Form der Positionierung zur Diskurseinheit vornehmen, relativ unabhängig davon ob nativ oder anglizistisch.

(4)

a. Denn die Oldschool bleibt für ewig am rulen [microphone_checker]
b. Scheiß auf deinen Oldschool wir sind frisch [s_a_d_o_s]
c. Dein rap ist veraltet, Oldschool und fegt den Friedhof leer [am_riden]

d. Scheiß auf dich, ich feier nur die Oldschool-Hits [1988]
e. Von Oldschool bin ich Lehrer, von Newschool der Direktor [weiter_2]
f. Oldschool Style wie Breakdance auf Kartons [yes_yes_yo]
g. Ich bin so alte Schule mein Sound-System ist Deluxe [epos]
h. Fick die alte Schule, Junge, das hier ist der neue Scheiß [ganz_einfach]

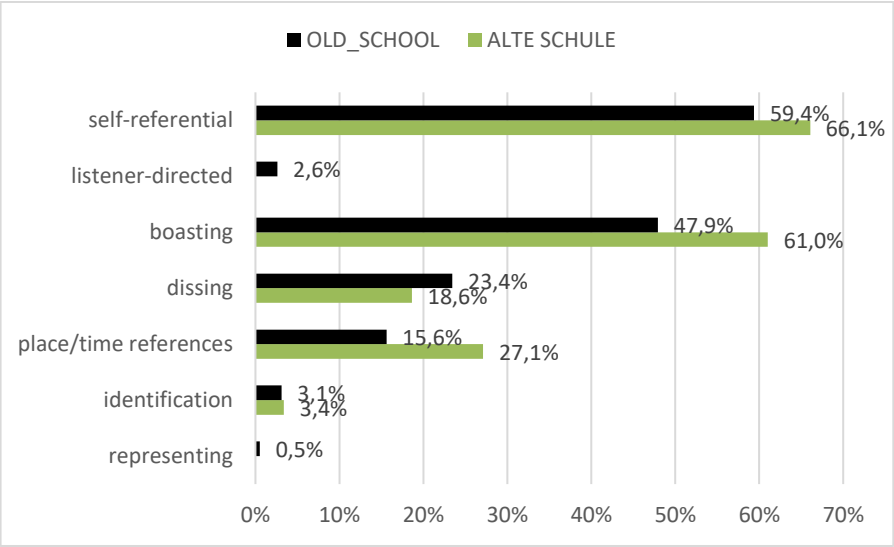


Abbildung 1: Verteilung der *Speech-Act-Patterns* bei OLD_SCHOOL und ALTE SCHULE.

	OLD_SCHOOL (n=123)		ALTE SCHULE (n=33)	
	positiv	negativ	positiv	negativ
Boasting	90,6%	9,4%	86,2%	13,4%
Dissing	34,2%	65,8%	37,5%	62,5%

Tabelle 2: Positive und negative Konnotation von OLD_SCHOOL und ALTE SCHULE bei *Boasting* und *Dissing*.

Die *Place/Time-References* können im Rahmen der Verankerung im Zusammenhang mit den **(ii)** ***Cultural References*** gesehen werden. Interessanterweise wird bei OLD_SCHOOL weitaus mehr Gebrauch von diesen Referenzen gemacht (38 Belege) als bei ALTE SCHULE (4 Belege). Beachtet werden muss hierbei aber natürlich auch der Unterschied in der Gesamtbelegzahl: Während damit insgesamt 19,8% der Belege für die anglizistische Nomi-

nation *Cultural References* beinhalten, liegt dieser Anteil bei der nativen Nomination bei lediglich 6,8%. Bei Betrachtung der Verteilung der *Cultural References* für OLD_SCHOOL wird deutlich, dass sich vor allem auf Personen bezogen wird (57,9%) (vgl. Abbildung 2), hier zumeist etablierte Persönlichkeiten vor allem der Hip-Hop-Kultur aus den USA (5a), wodurch vermutlich u. A. Subkulturkompetenz und -verortung ausgedrückt werden soll. Auf Fiktionales (5c) wird sich kaum bezogen (5,3%). Die wenigen Fälle stehen dann, wie die Markennamen (5b) (36,8%), zumeist im Zusammenhang mit Anhaltspunkten geteilten (Pop-)Kulturwissens ‚früherer Zeiten' – vermutlich um auch hier die Subkulturkompetenz und -verortung und vor allem die Authentizität im Zusammenhang mit OLD_SCHOOL zu unterstreichen.

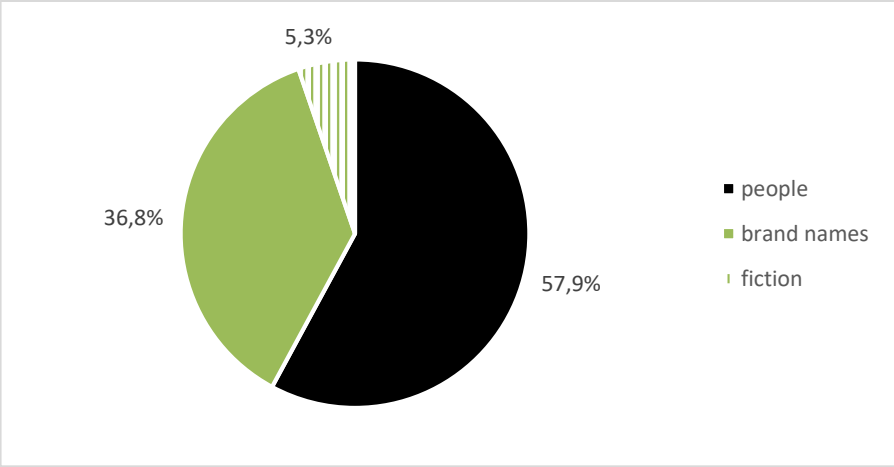


Abbildung 2: Verteilung der *Cultural References* bei OLD_SCHOOL.

Bei den Markennamen ist auffällig, dass sich vor allem auf Unterhaltungselektronik bezogen wird (bspw. 5e), die sich im Rahmen fortlaufenden technischen Fortschritts als Zeugnis einer ‚früheren Zeit' und zusätzlich im Rahmen von Freizeitgestaltung als Zeugnis einer geteilten Vergangenheit im Alltagsleben als Referenz anbietet, um dem zeitlichen Aspekt von OLD_SCHOOL zusätzlich Glaubwürdigkeit zuzutragen.

(5)

a. Ich bin Oldschool wie Afrika Bambaataa	[butterfly_effect]
b. Ich bin Oldschool wie MySpace oder Napster	[bandsalat]
c. GPC ist old-school wie Tom und Jerry	[cartoon]
d. Es ist Double Dragon, Oldschool, Nintendo	[double_dragon]
e. Spielen Playsi oldschool Tekken	[high]
f. Oldschool Trap-Phone, Moto Razr	[ice-1]
g. ihr seid oldschool so wie Nokia	[luzifer_1]
h. ich leb' oldschool wie Atari	[real_motherfucker]

Beispiele (5a–c) und (5g–h) veranschaulichen, wie solche Referenzen in Vergleichen, *Comparisons*, umgesetzt werden. Der Vergleich der (iii) *Rap-Rethorics* (Abbildung 3) weist für OLD_SCHOOL ein weitaus größeres Vorkommen von *Comparisons* aus (56,3%) als für ALTE_SCHULE (14,3%), was die Beobachtung der *Cultural References* hinsichtlich der Verteilung zwischen anglizistischer und nativer Nomination bestärkt – in mehr als der Hälfte aller Belege wird OLD_SCHOOL im Kontext von *Comparisons* genutzt. Der zuvor generell dargestellte Trend (vgl. Abbildung 2) zeigt sich hier in ähnlicher Form: 35,0% der *Comparisons* kommen mit *Cultural References* auf Personen vor, 17,5% mit Markennamen und 5,0% mit Fiktionalem. Damit treten also fast 60% aller *Comparisons* mit *Cultural References* nach der Definition von Adroutsopoulos und Scholz auf (vgl. Abschnitt 5.1.2)

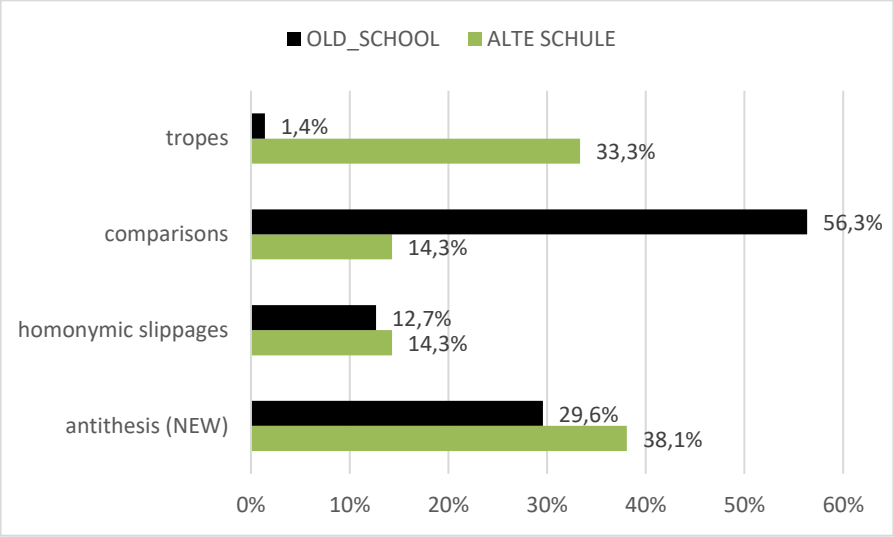


Abbildung 3: Verteilung der *Rap-Rhetorics* bei OLD_SCHOOL und ALTE SCHULE.

Insgesamt indiziert der Vergleich für beide Nominationen das umfangreiche ‚Sprachspiel', den für Rap konstitutiven *Skill*. Neben dem Unterschied bei den *Comparisons* ist hier vor allem auch die Differenz bei den *Tropes* auffällig. Der hier gefolgten Eingrenzung von Androutsopoulos und Scholz auf metaphorische Aspekte (vgl. 2002: 21 f.) beziehen sich die entsprechend annotierten Belege auf metaphorische ‚Spiele' mit der Domäne[22] Schule, ein erstes Beispiel hierzu ist bereits bei den *Speech-Act-Patterns* angeführt worden:

[22] In der Terminologie der von Androutsopoulos und Scholz angeführten konzeptuellen Metapherntheorie nach Lakoff/Johnson (2003) die sogenannte *Source Domain*.

(6)

a. Von Oldschool bin ich Lehrer, von Newschool der Direktor [weiter_2]

Dies ist der einzige Beleg für OLD_SCHOOL. Trotz geringerer Gesamtbelegzahl stellt sich bei ALTE SCHULE ein anderes Bild dar. Zwar muss einschränkend darauf hingewiesen werden, dass sich hinsichtlich der Dispersion die Belege auf zwei Quellen konzentrieren, dennoch kann dies, eingedenk der erwähnten geringeren Belegzahl, zumindest tendenziell als Hinweis angenommen werden.

(7)

a. Alte Schule, Neue Schule, wir schwänzen Schule [mmm_moerder_monster_muschel]
b. Die Truppen der alten Schule gelang‘ auf dem karierten Blatt ins Kreuzfeuer [papierkram]
c. Zu diesem Zeitpunkt ging‘ einige Kämpfer der alten Schule stiften. Papierflieger kreisten um das Einsatzgebiet [papierkram]

In Anbetracht des Titels *Papierkram* ist hierbei zu problematisieren, inwieweit die „alte Schule“ hier tatsächlich im Sinne des Vergleichspaars gemeint ist. Als mit den vorliegenden Daten nicht weiter nachweisbare Erklärung könnte dienen, dass für solche metaphorischen Vorgänge die native Nomination ggf. deshalb zugänglicher ist, weil das hiermit zusammenhänge Wortfeld, die Konzepte bzw. Frames im nativen Bereich für Muttersprachler abrufbarer sind. Eine ähnliche Begründung könnte bemüht werden, um den größeren Anteil antithetischer Belege, also solche mit der Gegenüberstellung von ‚alt‘ und ‚neu‘, zu erklären. Dennoch zeigt sich an den Beispielen, dass antithetische Konstruktionen bei der nativen wie der anglizistischen Nomination sehr ähnlich verwendet werden:

(8)

a. Ja old school aber kein Scheiss von gestern [afrob_kommt]
b. Alte Schule, trotzdem up-to-date wie Dr. Dre [alles_zersaegt]
c. Scheiß auf deinen Oldschool, wir sind frisch [don_t_like]
d. Fick die alte Schule, Junge, das hier ist der neue Scheiß [ganz_einfach]
e. Ich bin nicht Oldschool, nicht Newschool [leko_mio]
f. Alte Schule, neue Schule, wir schwänzen Schule [mmm_moerder_monster_muschel/44]
g. Ich bin zwar neu in den Charts, doch Oldschool wie Methusalem [freund_oder_feind]
h. Ich bin jung, fresh, jedoch Oldschool [power_1]
i. Ich bin die neue Ära und die alte Schule [55_interlude]
j. Alte Schule, die neue deutsche Welle [alte_schule]
k. Oldschool wird jetzt wieder wie neu sein [microphone_checker]
l. Ich mach Oldschool neu [weiter_2]

So lassen sich fast alle Belege in diese durch Beispiele dargestellten Gruppen unterteilen. Bei (8a–b) ordnen sich die Autoren selbst der OLD_SCHOOL / ALTEN SCHULE zu, betonen dabei aber, dass dies der Aktualität nicht abträglich sei. (8c–d) veranschaulicht eine klare Positionierung dagegen und hebt die eigene Aktualität hervor. Die Beispiele (8e–f) illustrieren

die Grenzfälle, da hier diskutiert werden könnte, ob von einer antithetischen Konstruktion ausgegangen werden kann oder aber nicht beide ‚Strömungen' im Diskurs gleichzeitig abgetan werden, ohne dabei die antonymische Relation zu thematisieren. Die Beispiele (8g–h) können gewissermaßen als Gegenentwurf zu (8a–b) gesehen werden, da nun die eigene Aktualität angeführt wird, aber hervorgehoben wird, dass man dennoch zu dieser Diskureinheit gehöre, also OLD_SCHOOL / ALTE_SCHULE sei. (8i–j) stellen in dieser Gruppe Grenzfälle dar. Lediglich für (8k–l) lassen sich keine vergleichbaren Belege für die native Nomination identifizieren, hier wird die Erneuerung der „Oldschool" proklamiert.
Insgesamt lassen sich, wie eingangs erwähnt, kaum wesentliche Verwendungsunterschiede zwischen nativer und anglizistischer Nomination feststellen. Vor allem bei den *Cultural References* zeigt sich jedoch die engere Einbettung in kulturelle Bezüge, die eine Verortung vornehmen und eine je spezifische Erfahrungsbasis kommunizieren. OLD_SCHOOL / ALTE SCHULE zeigt sich damit insgesamt als eine für den Hip-Hop-Diskurs in Raptexten relevante Diskurseinheit, die – nebenbei nochmals angemerkt – auch die in Abschnitt 3.3 angeführten orthografischen Besonderheiten zu plausibilisieren vermag. Dabei scheint OLD_SCHOOL / ALTE SCHULE weder gänzlich positiv oder negativ konnotiert und damit in der Selbst- und Fremdpositionierung sowohl für *Boasting* als auch für *Dissing* nutzbar und wird hier vor allem als eine Art ‚Gütesiegel' genutzt („In der Leute kein Recht dazu haben sich oldschool zu nennen" [in_einer_Welt]).

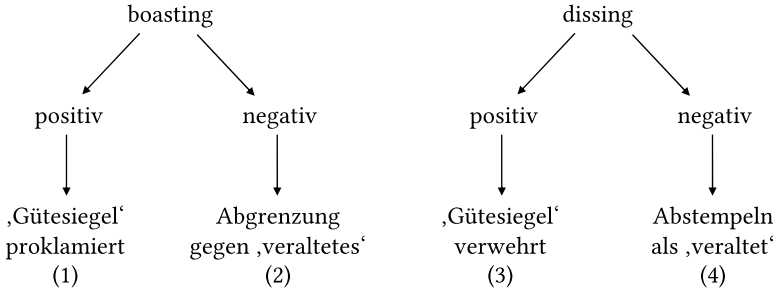


Abbildung 4: Selbst- und Fremdpositionierung zu OLD_SCHOOL / ALTE SCHULE.

Für alle diese Möglichkeiten (9; 1–4) lassen sich Beispiele für native wie anglizistische Nomination im Untersuchungskorpus finden. Hinsichtlich der Verteilung ist die Verwendung, wie bereits gezeigt, trotz unterschiedlicher Okkurrenzen sehr ähnlich.

(9)

(1) Denn Oldschool kann ich	[1000_bars/4]
(2) Das ist sicherlich kein Old School von gestern	[herz_und_seele/158]
(3) Du bist nicht Oldschool, sondern einfach nur Antik	[fresh_und_unbekannt/138]
(4) Scheiß auf deinen Oldschool, wir sind frisch	[s_a_d_o_s/308]

7.2 Syntaktische/morphologische Verwendung

Zunächst ist für die syntaktische Analyse der große Anteil von ALTE SCHULE bei der Kategorie „unklar“ (vgl. 4.1.4) hervorzuheben, ein Resultat mangelnden (Satz-)Kontexts. Daraus lässt sich also schließen, dass ALTE SCHULE im Untersuchungskorpus in weniger Belegen syntaktisch eingebunden ist. Die genauere Interpretation und Bezeichnung dieser Vorkommen ist m. E. n. in der aktuellen Forschung zu umstritten, als dass dies hier ausreichend gesichert und dem Umfang der Arbeit angemessen vorgenommen werden könnte. Die Beispiele (10a–e) illustrieren die unterschiedlichen Vorkommen, die bspw. als temporale (10a), modale (10b) oder lokale (10c) Adverbiale interpretiert werden können. Vor dem Hintergrund der Textsorte des Untersuchungskorpus muss zudem davon ausgegangen werden, dass Ambiguitätspotenziale – bspw. im Rahmen des *Signifyings* (vgl. 5.2) – womöglich bewusst genutzt werden. So kann (10c) insofern lokal verstanden werden, als dass der Interpret an der alten Schule Drogengeschäften nachgegangen ist, gleichzeitig könnte dies auch eine Art Fazit zur beschriebenen Tatsache sein, dass ihm nichts geschenkt worden sei – ein ‚hartes Leben‘ könnte ‚alte Schule‘ sein. Ebenfalls charakteristisch ist das Vorkommen in Aufzählungen (10e).

(10)

a. Zeit fliegt, alte Schule, damals Handtaschendiebe.	[30er_zone]
b. Ich bleib‘ grade, alte Schule, hustle jeden Tag für Patte.	[zwanni]
c. Keiner hat mir was geschenkt, alte Schule, Grasticker.	[intro_verlorene_tracks_ep]
d. Ich rap‘ auf G-Funk, Alte Schule, Classic.	[cripwalk]
e. Boppard, Handschlag, Deal, alte Schule.	[ueberall_zuhaus]

Bei diesen Belegen der nativen Nomination, aber vor allem auch bei der anglizistischen, lässt sich für die „unklar“ zugeordneten Fälle eine Tendenz zu Randpositionen von Sätzen feststellen. Funktional sind sie dabei häufig evaluierend. Dies soll nur einen groben Einblick vermitteln, theoretisch wird in dieser Arbeit nicht weiter auf die unklaren Fälle eingegangen, ihrer angemessenen Analyse und Zuordnung wäre hier zu viel Aufmerksamkeit zu widmen.[23]

Der hohe Anteil prädikativer Verwendung lässt sich gut mit den Ergebnissen des Vergleichs der pragmatischen Verwendung in Zusammenhang bringen: Die Positionierung von sich selbst oder anderen gegenüber OLD_SCHOOL / ALTE SCHULE, also die Zuschreibung oder Absage dieses ‚Gütesiegels‘, erfolgt meist über Kopulaverben mit OLD_SCHOOL / ALTE

[23] Eine umfangreichere Systematisierung zu Ellipsen und Fragmenten im Kontext von Jugendsprache nimmt Androutsopoulos vor (vgl. 1998: 292–301). Die Untersuchung Dereckas zu Texten des Rappers Haftbefehl thematisiert ebenfalls syntaktische Phänomene, zu denen sich zahlreiche der unklaren Fälle zuordnen ließen. (vgl. 2016: 116–123). Eine aktuelle Analyse und Interpretation *kopulaloser Nominalsätze*, die sich im Untersuchungskorpus auch häufiger finden lassen, bietet Behr (2016).

SCHULE als Prädikativa. Dass der Anteil prädikativer Verwendung bei der anglizistischen Nomination höher ist als bei der nativen (30,2% vs. 18,6%), kann als Resultat der morphologischen Unterdeterminiertheit interpretiert werden. OLD_SCHOOL kann sowohl als Substantiv als auch als Adjektiv prädikativ gebraucht werden und bedarf dabei keiner morphologischen Modifikation, lediglich einer graphematischen (Groß- und Kleinschreibung), die im Untersuchungskorpus aber aufgrund der erläuterten Aspekte nicht berücksichtigt wird (vgl. 3.3).

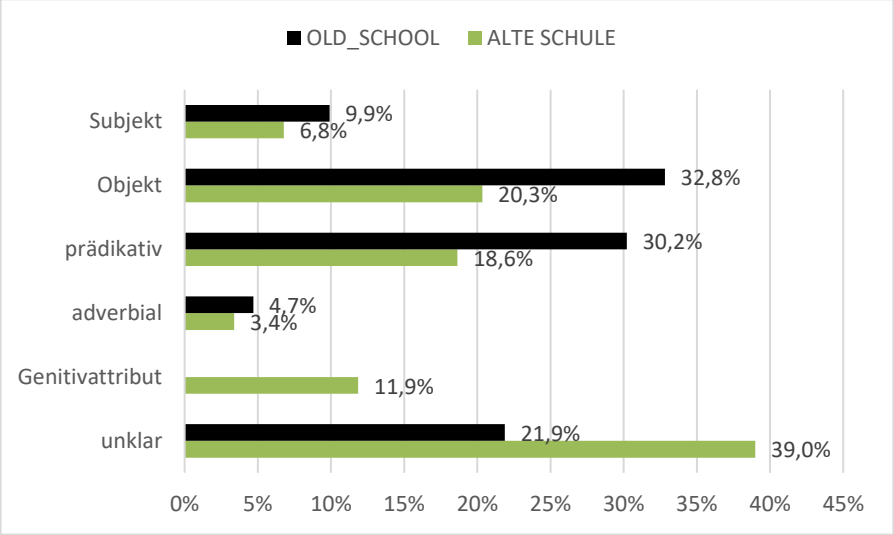


Abbildung 5: Syntaktische Verwendung von OLD_SCHOOL und ALTE SCHULE.

Zur adverbialen Verwendung ist zu erwähnen, dass es sich bei ALTE SCHULE ausschließlich um Belege in Form von Präpositionalphrasen handelt. Nur bei OLD_SCHOOL finden sich Belege wie in den Beispielen (a) und (b).

(11)

a. ich leb‘ oldschool wie Atari [real_motherfucker]
b. Ihr klingt Oldschool wie ein altes Tape [zwei_punkt_nuller]

Die morphologische Unterdeterminiertheit der anglizistischen Nomination zeigt sich am deutlichsten bei der Wortbildung. Lediglich einem belegten Kompositum bei ALTE SCHULE steht ein Anteil von 36% der Belege bei OLD_SCHOOL entgegen. Diese teilen sich wie folgt auf:

Dreigliedrige Komposita	Dreigliedrige Komposita (hybrid)	Viergliedrige Komposita (hybrid)	Fünfgliedrige Komposita (hybrid)	Derivation	Suffixoid *-mäßig*
60,9%	23,2%	1,4%	1,4%	10,1%	2,9%
Old-schoolstyle	*Oldschool-reimerei*	*Oldschool-Baggy-hosen*	*OldSchool-Trainbombing-Graffiti-Aufnahmen*	*Old-schooler*	*oldschool-eastcoast-mäßig*

Tabelle 3: Verteilung der Wortbildungen mit OLD_SCHOOL als Erstglied.

Am häufigsten finden sich Kompositionen mit *Rapper* (15,8%), gefolgt von *Beat* (7,0%) und dann *Flow*, *Style* und *Party* (jeweils 5,3%). Jeweils zwei Belege finden sich für Komposita mit *Aufnahme*, *Hit*, *Kassette*, *Legende*, *Mucke*, *Scheiß* und *Shit*. Für Komposita mit folgenden Substantiven gibt es jeweils nur einen Beleg: *Attitüde, Atze, Baggyhose, Bitch, Bounce, Boy, Chevy, Feeling, Girl, Groove, Klischee, Legende, Lied, Manier, Reimerei, Rhyme, Romantik, Sound, Tape, Track, Vibe* und *Trainbombing-Graffiti-Aufnahme*. Damit wird die Produktivität der Komposition mit OLD_SCHOOL im Vergleich zu ALTE SCHULE belegt.[24] Als Erklärung hierfür können die Signalwirkung englischen Wortmaterials im Zusammenhang mit dem Hip-Hop-Diskurs und die Zurschaustellung der eigenen Sprachkompetenz (*Skill*) durch hybride Komposita angeführt werden (vgl. Kapitel 5.2). Zusätzlich soll nochmal auf die mehrfach erwähnte morphologische Besonderheit des Englischen hingewiesen werden: Was hinsichtlich Schreibung und grammatischer Interpretation zu Schwierigkeiten führt, erweist sich für die Komposition als äußerst hilfreich: Die nicht vorhandene Flexion des Adjektivs im Englischen erlaubt es, die anglizistische Nomination formal 1:1 in ein Kompositum zu überführen. Dies ist nicht nur hinsichtlich des Wortmaterials ökonomischer, sondern damit korrespondierend auch für die Verarbeitung, so zumindest die Annahme. Während beim einzigen Kompositum für die native Nomination (*Alte Schule-Boxer*; schatzmeister] das Flexionssuffix eine Interpretation als attributives Adjektiv und damit als Syntagma anstößt, wird dieser Prozess bspw. bei *Oldschoolreimerei* von der Wortform vermutlich nicht induziert. Zudem müsste bei Kompositionen mit ALTE SCHULE wie in Kapitel 3.3 erläutert eine Durchkopplung mit Bindestrichen erfolgen, was die Komplexität zusätzlich erhöht. Zudem geht mit dem Flexionssuffix bei der nativen Nomination auch immer eine zusätzliche Silbe einher, die bei der anglizistischen Nomination eingespart wird.

Diese morphologische Flexibilität aufgrund der Unterdeterminiertheit zeigt sich ebenfalls beim *-er*-Derivat *Oldschooler*. Ein Versuch, diese Bildung analog für die native Nomination vorzunehmen, scheint schwierig. Ob die *-er*-Derivation hier tatsächlich auf die morphologische Integration ins Deutsche verweist oder aber bereits im Englischen so gebildet worden

[24] Onysko deutet diese „productive nests“ (2007: 41) mit Anglizismen in Hybridkomposita als Integrationsmarker.

ist, wo dies auch als produktives Muster angenommen wird (vgl. Barz 2008: 47; Duden 2021: 78), bleibt unklar. In einschlägigen Wörterbüchern des Deutschen und Englischen findet sich kein Eintrag hierzu, wohl aber für *preschooler* und *Grundschüler*, was die Möglichkeit zur -*er*-Derivation in diesem Kontext für beide Sprachen unterstreicht. Die Schwierigkeiten der Unterscheidung, ob es sich um eine ‚fertige' Entlehnung aus einer fremden Sprache oder um Fremdwortbildung innerhalb des Deutschen handelt, ist ein bekanntes Problem (vgl. bspw. Seiffert 2005).

8 Fazit

Diese exemplarische Untersuchung konnte für eine spezifische Domäne verschiedene Anreize für die Verwendung von Anglizismen perspektivieren: Neben der sozialindexikalischen Leistung des Englischen generell, die für den Hip-Hop-Diskurs von besonderer Bedeutung ist, scheint es auch stichhaltige wortformale Eigenschaften zu geben, die Anglizismen hier für Nominationen gerade im fachsprachlichen Kontext qualifizieren. Die Eignung von Fremdwörtern für Fachsprachen wird zwar häufig angeführt (vgl. bspw. Chang 2005), die Einsichten dieser Untersuchung weisen aber auf weitere Aspekte als die üblichen wie Kürze oder Internationalität hin. Zwar wird auch hier im Ergebnis Kürze thematisiert, dies aber vor allem im Hinblick auf morphologische Reduziertheit bzw. Unterdeterminiertheit. Durch die fehlende Adjektivflexion im Englischen können mit solchen prototypischen Nominationen aus Adjektiv und Subjektiv problemlos(er) komplexere Komposita gebildet werden, was gerade für entsprechende Diskurseinheiten aus terminologischer Sicht besonders produktiv scheint – und sich in zahlreichen solcher anglizistischen Nominationen im Deutschen und Komposita mit diesen ausdrückt. Diese morphologische Eigenschaft schlägt sich in syntaktischer Flexibilität nieder, die auch adjektivische Verwendung ohne formale morphologische Modifikation trifft für viele andere dieser Nominationen zu. Somit können sich, bei der terminologisch relevanten Schonung der Wortform, Wortbildungsnester ausbilden. Daneben sichert die Eigenschaft ‚fremdsprachig' als ‚Fremdsprachenklammer' im Kontrast zum nativen Kontext ebenfalls die begriffliche Interpretation, sie zeichnet die Begriffsgrenzen zusätzlich aus.

Für OLD_SCHOOL / ALTE SCHULE selbst hat sich der Status als Hip-Hop-spezifischer Fachbegriff im Untersuchungskorpus bestätigt, da komplexe Verwendungsmuster und diskursive Relevanz herausgearbeitet werden konnten (was übrigens auch die Schreibung <Alte Schule> rechtfertigte). Hierbei war das adaptierte Analyseframework von Androutsopoulos und Scholz (2002) grundlegend und konnte für die Analyse eines Vergleichspaars innerhalb nur einer Sprache angepasst nutzbar gemacht werden. Methodisches Potenzial liegt in der Validierung durch Interrater-Reliabilität und weiteren Auseinandersetzung mit möglichen Verzerrungen durch Frequenz- und Verteilungseffekte. Damit wurde auch das Hip-Hip-Subkorpus vom Songtextekorpus für verschiedene linguistische Forschungsrichtungen angeschnitten. Es böte sich an, in einem analogen Vorgehen weitere Anglizismen in Deutschraptexten zu untersuchen, die womöglich zusätzliche Nutzungsaspekte – sowohl für Anglizismen generell als auch für bestimmte Lexeme in der Hip-Hop-Domäne – offenlegen

könnten. Hier zeigt sich im Zusammenspiel zugespitzt, was sich an Möglichkeiten durch die spezifischen Eigenschaften des Englischen für eine wortbildungsaffine Sprache wie das Deutsche eröffnet.

9 Literatur

Altleitner, Margret. 2007. Der Wellness-Effekt: die Bedeutung von Anglizismen aus der Perspektive der kognitiven Linguistik. Frankfurt am Main: P. Lang.

Androutsopoulos, Jannis. 2003. „‚jetzt speak something about italiano.' Sprachliche Kreuzungen im Alltagsleben". Osnabrücker Beiträge zur Sprachtheorie (65): 79–109.

———. 2007. „Style online: Doing hip-hop on the German-speaking Web". In Style and social identities: alternative approaches to linguistic heterogeneity, hrsg. Peter Auer. Berlin; New York: de Gruyter, 279–317.

Androutsopoulos, Jannis K. 1998. Deutsche Jugendsprache: Untersuchungen zu ihren Strukturen und Funktionen. Frankfurt am Main; New York: P. Lang.

——— (Hg.). 2003. HipHop: Globale Kultur - lokale Praktiken. Bielefeld: Transcript-Verlag.

Androutsopoulos, Jannis K. / Alexandra Georgakopoulou (Hg.). 2003. Discourse Constructions of Youth Identities. Amsterdam Philadelphia: Benjamins.

Androutsopoulos, Jannis K. / Arno Scholz. 2003. „Spaghetti Funk: Appropriations of Hip-Hop Culture and Rap Music in Europe". Popular Music and Society 26(4): 489–505.

Androutsopoulos, Jannis / Arno Scholz. 2002. „On the recontextualization of hip-hop in European speech communities: A contrastive analysis of rap lyrics". Philologie im Netz (19): 1–42.

Auer, Peter (Hg.). 1998. Code-Switching in Conversation: Language, Interaction and Identity. Transferred to digital pr. London: Routledge.

Bachtin, Michail Michajlovič / Adelheid Schramm / Michail Michajlovič Bachtin / Fëdor Michajlovič Dostoevskij. 1971. Probleme der Poetik Dostoevskijs. München: Hanser.

Behr, Irmtraud. 2016. „Kopulalose Nominalsätze". In Fragmentarische Äußerungen, Eurogermanistik, hrsg. Jean-François Marillier und Élodie Vargas. Tübingen: Stauffenburg Verlag, 137–56.

Bergmann, Rolf / Dieter Nerius (Hg.). 1997. Die Entwicklung der Großschreibung im Deutschen von 1500 bis 1700. Heidelberg: Winter.

Berns, Jan / Peter Schlobinski. 2003. „Constructions of identity in German hip-hop culture". In Discourse constructions of youth identities, Pragmatics and beyond, Amsterdam Philadelphia: Benjamins, 197–219.

Braselmann, Petra. 2002. „Englisch in der Romania". In Deutsch – Englisch – Europäisch: Impulse für eine neue Sprachpolitik, Thema Deutsch, Mannheim: Dudenverlag, 298–332.

Bredel, Ursula / Tilo Reißig (Hg.). 2022. Weiterführender Orthographieerwerb. 3., durchgesehene und aktualisierte Auflage. Baltmannsweiler: Schneider Verlag Hohengehren.

Buchmann, Franziska. 2015. Die Wortzeichen im Deutschen. Heidelberg: Universitätsverlag Winter.

Burger, Harald. 2015. Phraseologie: eine Einführung am Beispiel des Deutschen. 5., neu bearbeitete Aufl. Berlin: Erich Schmidt.

Busse, Ulrich. 1993. Anglizismen im Duden: eine Untersuchung zur Darstellung englischen Wortguts in den Ausgaben des Rechtschreibdudens von 1880-1986. Tübingen: Niemeyer.

———. 2008. „Anglizismen im Deutschen: Entwicklung, Zahlen, Einstellungen“. In Sprachkontakt und Mehrsprachigkeit: Zur Anglizismendiskussion in Deutschland, Österreich, der Schweiz und Italien, Sprache, Literatur und Geschichte, hrsg. Sandro M. Moraldo. Heidelberg: Winter, 37–68.

Carstensen, Broder / Regina Schmude / Ulrich Busse. 2001. Anglizismen-Wörterbuch: der Einfluss des Englischen auf den deutschen Wortschatz nach 1945. Berlin: W. de Gruyter.

Chang, Youngick. 2005. Anglizismen in der deutschen Fachsprache der Computertechnik: eine korpuslinguistische Untersuchung zu Wortbildung und Bedeutungskonstitution fachsprachlicher Komposita. Frankfurt am Main; New York: P. Lang.

Derecka, Małgorzata. 2021. Patchworkdeutsch – Sprachlich-kulturelle Interferenz in den Songtexten von Haftbefehl. Berlin: Peter Lang.

Eichinger, Ludwig M. 2008. „Anglizismen im Deutschen meiden – warum das nicht so leicht ist“. In Sprachkontakt und Mehrsprachigkeit: zur Anglizismendiskussion in Deutschland, Österreich, der Schweiz und Italien, Sprache, Literatur und Geschichte, hrsg. Sandro M. Moraldo. Heidelberg: Winter, 69–93.

Eisenberg, Peter. 2013. Grundriss der deutschen Grammatik. Bd. 1: Das Wort. 4., aktualis. u. überarb. Aufl. Stuttgart Weimar: Metzler.

———. 2018. Das Fremdwort im Deutschen. 3., überarbeitete und erweiterte Auflage. Boston: Walter de Gruyter.

Fiedler, Sabine. 2014. Gläserne Decke und Elefant im Raum: phraseologische Anglizismen im Deutschen. Berlin: Logos Verlag Berlin GmbH.

Fleischer, Wolfgang. 1996. „Phraseologische, terminologische und onymische Wortgruppen als Nominationseinheiten“. In Nomination - fachsprachlich und gemeinsprachlich, hrsg. Clemens Knobloch und Burkhard Schaeder. Opladen: Westdeutscher Verlag, 147–70.

Fleischer, Wolfgang / Irmhild Barz / Marianne Schröder. 2012. Wortbildung der deutschen Gegenwartssprache. 4. Auflage, völlig neu bearbeitet. Berlin Boston: De Gruyter.

Fuhrhop, Nanna. 2007. Zwischen Wort und Syntagma: zur grammatischen Fundierung der Getrennt- und Zusammenschreibung. Tübingen: M. Niemeyer.

———. 2015. Orthografie. Vierte, aktualisierte Auflage. Heidelberg: Universitätsverlag Winter.

———. 2022. „System der Getrennt- und Zusammenschreibung“. In Weiterführender Orthographieerwerb, Deutschunterricht in Theorie und Praxis, Baltmannsweiler: Schneider Verlag Hohengehren, 107–28.

Glück, Helmut / Michael Rödel (Hg.). 2016. Metzler Lexikon Sprache. 5., aktualisierte und überarbeitete Auflage. Stuttgart: J.B. Metzler Verlag.

Görlach, Manfred (Hg.). 2001. A dictionary of European anglicisms: a usage dictionary of anglicisms in sixteen European languages. Oxford [England] ; New York: Oxford University Press.

Güler Saied, Ayla. 2012. Rap in Deutschland: Musik als Interaktionsmedium zwischen Partykultur und urbanen Anerkennungskämpfen. 1. Auflage. Bielefeld: transcript.

Hennig, Mathilde / Jan Georg Schneider / Ralf Osterwinter / Anja Steinhauer. 2021. Duden - Sprachliche Zweifelsfälle: das Wörterbuch für richtiges und gutes Deutsch. 9., überarbeitete und erweiterte Auflage. Berlin: Dudenverlag.

Hoberg, Rudolf. 2002. Deutsch - Englisch - Europäisch: Impulse für eine neue Sprachpolitik. Mannheim: Dudenverlag.

Jacobs, Joachim. 2005. Spatien: zum System der Getrennt- und Zusammenschreibung im heutigen Deutsch. Berlin: De Gruyter.

Keim, Inken / Wilfried Schütte (Hg.). 2002. Soziale Welten und kommunikative Stile: Festschrift für Werner Kallmeyer zum 60. Geburtstag. Tübingen: Narr.

Kluge, Friedrich / Elmar Seebold. 2011. Etymologisches Wörterbuch der deutschen Sprache. 25., durchgesehene und erw. Aufl. Berlin; Boston: De Gruyter.

Knobloch, Clemens / Burkhard Schaeder (Hg.). 1996. Nomination - fachsprachlich und gemeinsprachlich. Opladen: Westdeutscher Verlag.

Kunkel-Razum, Kathrin u. a. (Hg.). 2020. Duden - die deutsche Rechtschreibung: auf der Grundlage der aktuellen amtlichen Rechtschreibregeln. 28., völlig neu bearbeitete und erweiterte Auflage. Berlin: Dudenverlag.

Lakoff, George / Mark Johnson. 2003. Metaphors we live by. Chicago: University of Chicago Press.

Langner, Heidemarie C. 1995. Die Schreibung englischer Entlehnung im Deutschen: eine Untersuchung zur Orthographie von Anglizismen in den letzten hundert Jahren, dargestellt an Hand des Dudens. Frankfurt am Main; New York: P. Lang.

Margara, Andreas. 2018. „‚Ich zerstöre meinen Feind' – Die Evolution von Battle-Rap in Deutschland". IDS SPRACHREPORT (4): 2–9.

Marillier, Jean-François, und Élodie Vargas, hrsg. 2016. Fragmentarische Äußerungen. Tübingen: Stauffenburg Verlag.

Moraldo, Sandro M. (Hg.). 2008. Sprachkontakt und Mehrsprachigkeit: zur Anglizismendiskussion in Deutschland, Österreich, der Schweiz und Italien. Heidelberg: Winter.

Müller, Hans-Georg. 2016. Der Majuskelgebrauch im Deutschen: Gross- und Kleinschreibung theoretisch, empirisch, ontogenetisch. Berlin; Boston: De Gruyter.

Müller, Peter O. (Hg.). 2005. Fremdwortbildung: Theorie und Praxis in Geschichte und Gegenwart. Frankfurt am Main; New York: P. Lang.

Munske, Horst Haider, hrsg. 2004a. Deutsch im Kontakt mit germanischen Sprachen. Tübingen: Niemeyer.

———. 2004b. „Englisches im Deutschen. Analysen zum Anglizismenwörterbuch". In Deutsch im Kontakt mit germanischen Sprachen, Germanistische Linguistik, hrsg. Horst Haider Munske. Tübingen: Niemeyer, 155–74.

Noll, Volker / Sylvia Thiele. 2004. Sprachkontakte in der Romania: zum 75. Geburtstag von Gustav Ineichen. Tübingen: M. Niemeyer.

Onysko, Alexander. 2007. Anglicisms in German: borrowing, lexical productivity, and written codeswitching. Berlin; New York: Walter de Gruyter.

Pejčeva, Neli Christova. 2014. Akzeptanz englischen Wortgutes in Lifestyle-Magazinen: eine Untersuchung der Motivierbarkeit der Übernahme von Anglizismen vor dem Hintergrund des gesellschaftlichen Wertewandels; am Beispiel des österreichischen „Wiener" und des bulgarischen „Egoist". Hamburg: Kovač.

Potter, Russell A. 1995. Spectacular vernaculars: hip-hop and the politics of postmodernism. Albany: State University of New York Press.

Rampton, Ben. 1998. „Language crossing and the redefinition of reality“. In Code-switching in conversation: language, interaction and identity, hrsg. Peter Auer. London: Routledge, 290–320.

Rhys, Larysa F. / Olena Y. Bondarchuk / Larysa A. Pasyk. 2021. „Angleichung der Neuanglizismen an das System der deutschen Gegenwartssprache“. Філологія 1(50): 107–11.

Roelcke, Thorsten. 2020. Fachsprachen. 4., neu bearbeitete und wesentlich erweiterte Auflage. Berlin: Erich Schmidt Verlag.

Scherer, Carmen / Anke Holler (Hg.). 2010. Strategien der Integration und Isolation nicht-nativer Einheiten und Strukturen. Berlin: De Gruyter.

Schlobinski, Peter / Gaby Kohl / Irmgard Ludewigt. 1993. Jugendsprache: Fiktion und Wirklichkeit. Opladen: Westdeutscher Verlag.

Schneider, Roman. 2022. „Zwischen Schriftlichkeit und Mündlichkeit: Songtexte in der deskriptiven Sprachforschung”. In Sprachreport 1/2022. 38–50.

Scholz, Arno. 2004. „Die Nutzung von Angloamerikanismen zwischen Bedürfnis und Luxus. Spanische, französische, italienische und deutsche Beispiele aus Hip-Hop-Zeitschriften“. In Sprachkontakte in der Romania: zum 75. Geburtstag von Gustav Ineichen, Tübingen: M. Niemeyer, 259–72.

Seiffert, Anja. 2005. „Probleme synchroner Fremdwortbildungsforschung“. In Fremdwortbildung: Theorie und Praxis in Geschichte und Gegenwart, Dokumentation germanistischer Forschung, Frankfurt am Main; New York: P. Lang, 219–39.

Siekmeyer, Anne. 2007. Form und Gebrauch komplexer englischer Lehnverben im Deutschen: eine empirische Untersuchung. Bochum: Universitätsverlag Brockmeyer.

Spitzmüller, Jürgen. 2005. Metasprachdiskurse: Einstellungen zu Anglizismen und ihre wissenschaftliche Rezeption. Berlin; New York: W. De Gruyter.

Streeck, Jürgen. 2002. „Hip-Hop-Identität“. In Soziale Welten und kommunikative Stile: Festschrift für Werner Kallmeyer zum 60. Geburtstag, Studien zur deutschen Sprache, hrsg. Inken Keim und Wilfried Schütte. Tübingen: Narr, 537–57.

Verlan, Sascha. 2003. „HipHop als schäne Kunst betrachtet - oder: die kulturellen Wurzeln des Rap“. In HipHop: Globale Kultur - lokale Praktiken, Cultural studies, hrsg. Jannis K. Androutsopoulos. Bielefeld: Transcript-Verlag, 138–46.

Wetzler, Dagmar. 2006. Mit Hyperspeed ins Internet: zur Funktion und zum Verständnis von Anglizismen in der Sprache der Werbung der Deutschen Telekom. Frankfurt am Main; New York: P. Lang.

Yang, Wenliang. 1990. Anglizismen im Deutschen: am Beispiel des Nachrichtenmagazins Der Spiegel. Tübingen: Niemeyer.

Zifonun, Gisela. 2010. „Von ‚Bush administration‘ zu ‚Kohl-Regierung‘: Englische Einflüsse auf die deutsche Nominalkonstruktionen?“ In Strategien der Integration und Isolation nicht-nativer Einheiten und Strukturen, Linguistische Arbeiten, hrsg. Carmen Scherer und Anke Holler. Berlin: De Gryuter, 165–82.

Zwischenstaatliche Kommission für deutsche Rechtschreibung, hrsg. 2005. Deutsche Rechtschreibung: Regeln und Wörterverzeichnis: amtliche Regelung. Tübingen: G. Narr.

Internetquellen

Verein Deutsche Sprache: Anglizismenindex. https://vds-ev.de/arbeitsgruppen/deutsch-in-der-oeffentlichkeit/ag-anglizismenindex/ (Abruf: 12.01.23).

Patrick, Vanessa (2016): Das sind Adlibs – und so kamen sie nach Deutschland. https://www.br.de/puls/musik/vorbild-us-rap-adlibs-100.html (Abruf: 05.01.23).

Korpora

IDS (2022): *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache* 2022-I (Release vom 08.03.2022), Mannheim: Leibniz-Institut für Deutsche Sprache. PID: 00-04B6-B898-AD1A-8101-4.

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, S. 835-841. URI: https://songkorpus.de

Korrespondenzanschrift

Marco Gierke
Leibniz-Institut für Deutsche Sprache
gierke@ids-mannheim.de

Katrin Hein

Beinahe-ums-Leben-kommen-in-Regenpfützen und *Chauvi-Macho-Macker-Stuss* – kreative Wortbildungen in Songtexten

Abstract

Im Zentrum dieses Beitrags steht die Analyse kreativer Wortbildungsprodukte in Songtexten. Der Fokus liegt somit bewusst auf solchen Wortbildungen, die nicht den Weg ins Lexikon finden, sondern gerade aufgrund ihres okkasionellen Charakters einen erhöhten Grad an Expressivität aufweisen, der dann gezielt für die spezifische kreative Qualität von Songtexten genutzt wird.
Solche okkasionellen komplexen Wörter, die sich in theoretischer Hinsicht innerhalb der Domäne der ‚Extravagant Morphology' verorten lassen, werden über das Kriterium der Wortlänge aus dem Songkorpus herausgefiltert und im Anschluss hinsichtlich ihrer formalen sowie semantisch-pragmatischen Besonderheiten analysiert. Im Vordergrund steht dabei die Frage, wodurch die Kreativität der insgesamt 183 Bildungen des Untersuchungskorpus getriggert wird. Die Analyse zeigt, dass expressive Effekte in Songtexten offenbar sowohl durch die Verwendung markierter Wortbildungsmuster als auch durch den Rückgriff auf ‚auffällige' Lexik erzeugt werden. Zum einen ist der Anteil markierter Wortbildungsmuster wie der Phrasenkomposition und anderer phrasaler Wortbildungen gegenüber klassischen Textsorten wie Zeitungstexten deutlich erhöht. Zum anderen wird durch die Verwendung einer umgangssprachlichen, vulgären, brutalen oder poetischen Lexik, aber auch mit unmarkierten Wortbildungsmustern wie der prototypischen Determinativkomposition, Aufmerksamkeit erregt. Insgesamt erweist sich das Songkorpus dabei als wahre Fundgrube für kreative Wortbildungsprodukte.

Keywords: Wortbildung, Komposita, Okkasionalismen, Extravagant Morphology

1 Einleitung

„Unverkennbar bedienen […] sich [Songtexte] sprachlicher Kniffe und Innovationen, um Aufmerksamkeit zu wecken. Künstler weichen teilweise bewusst von Konventionen ab und überdehnen Regeln, formulieren überraschend und nutzen dabei Freiheiten bei der Wortkomposition oder -stellung" (SCHNEIDER 2022, 40). Eben dieses Charakteristikum von Songtexten wird im folgenden Beitrag für den Bereich der Wortbildung diskutiert und durch eine ausführliche empirische Analyse von kreativen Wortbildungen im Songkorpus (*www.songkorpus.de*; vgl. SCHNEIDER 2022) illustriert. Der Fokus liegt dabei auf Okkasionalismen wie *Klickelkrackelzickzackland* oder *Beinahe-ums-Leben-kommen-in-Regenpfützen*, „die nur für eine bestimmte kommunikative Gelegenheit geschaffen werden" und „mitunter nur aus einem

speziellen Kontext heraus verständlich [sind]“ (GRAMMIS 2022, „Okkasionalismus“). Es handelt sich dabei um Wortbildungsprodukte, die im Sinne von Ortner et al. (1984, 167) als „nichtusuell“ zu bezeichnen sind, im weitesten Sinne „einer Erwartungsnorm zuwider[laufen]“ und eben keinen Eingang ins Lexikon finden (vgl. Kapitel 2). Als Indikator für Kreativität, bzw. als empirische Operationalisierungsmöglichkeit, wird dabei das gut objektivierbare Kriterium der Wortlänge herangezogen: Indem die längsten Wortbildungen des Songkorpus morphologisch analysiert werden (vgl. Kapitel 3.1), sollten – so die Annahme – insbesondere kreative Wortbildungsprodukte herausgefiltert werden.

Das Herzstück des empirischen Teils (s. Kapitel 3) bildet die linguistische Analyse des Untersuchungskorpus, das insgesamt 183 komplexe Wörter umfasst. Diese werden zunächst nach Wortbildungstypen klassifiziert und gegebenenfalls in Bezug auf weitere formale und/oder semantische Auffälligkeiten kommentiert. Das Ziel dieses Beitrags besteht jedoch nicht nur darin, kreative Wortbildungen im Songkorpus zu identifizieren und zu klassifizieren. Vielmehr zieht sich die Frage, wodurch solche Bildungen Aufmerksamkeit wecken, wie ein roter Faden durch den gesamten Analyseteil. Resultiert die spezifische Qualität der Songkorpus-Wortbildungen v.a. aus dem Rückgriff auf markierte Wortbildungsmuster wie die Phrasenkomposition (vgl. HEIN 2015)? Oder ist es v.a. die Wahl der Lexik – z.B. vulgäre Lexik wie in *Schwanz-in-den-Mund-Style* – innerhalb der Wortbildungsmuster, durch die kreative Wortbildungen Aufmerksamkeit auf sich ziehen? Oder eine Kombination aus beidem?

Darüber hinaus wird versucht, zumindest am Rande auch Bezüge zu Vorkommen und Verteilung von Wortbildungstypen in klassischen schriftsprachlichen Textsorten wie Zeitungstexten zu ziehen – diese sind jedoch eher kursorischer Natur und basieren nicht auf expliziten (empirischen) Vergleichen mit anderen Korpora wie z.B. dem Deutschen Referenzkorpus (DeReKo) (LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE 2011). Dennoch sind solche In-Bezug-Setzungen wichtig, um einerseits die spezifische Rolle von Wortbildung im Songkorpus herauszuarbeiten und andererseits die Wortbildungsforschung selbst durch die Berücksichtigung nicht-klassischer Textsorten zu bereichern. Schließlich stellt das Songkorpus aufgrund seines „weitgehend mündlich konzipierte[n] Charakters“ eine „komplementäre Datenquelle für empirische Untersuchungen“ (SCHNEIDER 2022, 49) dar, die es ermöglicht, z.B. die noch wenig erforschte Rolle von Wortbildung in der gesprochenen Sprache zu untersuchen (vgl. aber HEIN/ANTONIOLI in Vorb.; HELMER 2022; STUMPF 2021). Nicht zuletzt ist auch die bewusste Fokussierung okkasioneller Bildungen ein wichtiger Aspekt der Wortbildungsforschung, da Wortbildung zugleich von Regelhaftigkeit und Kreativität geprägt ist und auch okkasionelle Instantiierungen von Wortbildungsmustern etwas über die dahinterstehenden, regelhaften Strukturen aussagen.

Der Schwerpunkt des Beitrags liegt auf der Analyse „auffälliger“ komplexer Wörter im Songkorpus (s. Kapitel 3). Vorab erfolgt in Kapitel 2 eine knappe theoretische Verortung okkasioneller Wortbildungsprodukte innerhalb des Gesamtbereichs der deutschen Wortbildung. Kapitel 4 fasst abschließend die zentralen Ergebnisse der Analyse zusammen.

2 Zur Funktion von kreativer Wortbildung (in Songtexten)

Durch Wortbildung entstehen „neue lexikalische Einheiten, Lexeme“ (FLEISCHER/BARZ 2012, 10). Allerdings wird nicht jede Wortneubildung zwangsläufig auch ins Lexikon übernommen, auch wenn diese Möglichkeit grundsätzlich immer besteht (FLEISCHER/BARZ 2012, 18; 24). Der vorliegende Beitrag beschäftigt sich bewusst mit solchen Bildungen, die nicht den Weg ins Lexikon finden, sondern gerade aufgrund ihres okkasionellen Charakters einen erhöhten Grad an Expressivität aufweisen, der dann gezielt für die spezifische kreative Qualität von Songtexten genutzt wird. Abgesehen davon sind okkasionelle Wortbildungsprodukte aber auch von Interesse für die Wortbildungsforschung selbst, weil „ihre Bildungsweise und Frequenz Auskunft über die Produktivität der Modelle geben können“ (FLEISCHER/BARZ 2012, 25).

Die Funktion von Wortbildung liegt einerseits in der „Versprachlichung von Begriffen“ (DONALIES 2007, 3), d.h. in der Bezeichnung von (neuen) Konzepten. Andererseits wird aber auch wegen „bestimmter Erfordernisse der Satz- und Textbildung“ (z.B. Wortartwechsel) auf Wortbildungsprodukte zurückgegriffen (FLEISCHER/BARZ 2012, 3). Allgemein wird zwischen Erstbenennungen und Zweitbenennungen unterschieden. Während erstere auf ein objektives Ausdrucksbedürfnis zurückgehen, bei dem es darum geht, neue Konzepte zu benennen (z.B. *simsen*), sind Zweitbenennungen wie z.B. die Ersetzung von *Altenheim* durch *Seniorenresidenz* eher subjektiv begründet: „Neue Wörter werden in Äußerungssituationen gebildet, in denen Sprechern ein geeignetes Wort für einen Begriff fehlt oder in denen sie die verfügbaren Wörter für unpassend oder nicht treffend halten; sei es, dass ein ganz neuer Gegenstand oder Sachverhalt genannt (auch: bezeichnet) werden muss, sei es, dass eine neue Sicht auf eine Sache zum Ausdruck gebracht werden soll“ (WÖLLSTEIN/DUDENREDAKTION 2016, 4:652).

All diese Funktionen von Wortbildung sind grundsätzlich auch für das Songkorpus relevant. Die im empirischen Teil (s. Kapitel 3) fokussierten kreativen Wortbildungen sind aber v.a. auf subjektive Ausdrucksbedürfnisse zurückzuführen, genauer gesagt auf den Wunsch des Sprechers, beim Hörer eine bestimmte Wirkung zu erzielen (ERBEN 2006, 21f.). Im Vordergrund scheint in dieser Textsorte das Erzielen expressiver, poetischer Effekte sowie das Ringen um Aufmerksamkeit zu stehen. ‚Auffälligkeit‘ in diesem Sinne wird wohl gerade auch dadurch erzielt, dass – durch den Rückgriff auf ungewöhnliche Lexik und/oder unkonventionelle Wortbildungsmuster – alternative Benennungen (Zweitbenennungen) geschaffen werden. Es geht also nicht bzw. nicht ausschließlich darum, dass in Songtexten ein Bedarf besteht, völlig neue Konzepte durch die Bildung neuer Wörter zu bezeichnen: „Vielleicht werden dabei nicht immer echte Benennungslücken geschlossen, auf jeden Fall aber geht es um die regelhafte neuartige Kombination bekannter Elemente im Sinne von Wittgensteins Sprachspielen.“ (SCHNEIDER 2022, 42).

Der Status als ‚Okkasionalismus‘ wird üblicherweise mit dem Status als Hapax Legomenon, d.h. mit dem nur einmaligen (oder mit wenigen) Vorkommen in einem bestimmten Korpus verknüpft. Gerade dieses nicht rekurrente Vorkommen macht die Bildungen wohl auch in

gewissem Sinne attraktiv: „Solche Wortschönheiten gefallen oft auch, weil sie selten vorkommen und von ausgeprägtem Sprachgefühl oder kreativer Sprachlust zeugen“ (ZIFONUN 2021, 273, zitiert nach SCHNEIDER 2022, 42).

Durch den bewussten Fokus auf okkasionelle, d.h. nicht-lexikalisierte Wortbildungsprodukte in Songtexten wird hier zugleich die textdistinktive Funktion von Wortbildung fokussiert, welche diese – neben einer textkonstitutiven Funktion – auszeichnet (FLEISCHER/BARZ 2012, 26). Die textdistinktive Komponente der Wortbildung impliziert, dass komplexe Wörter nicht nur zur Individualität von Einzeltexten führen können, sondern auch „dazu beitragen, eine Textsorte zu charakterisieren und sie von anderen Textsorten zu unterscheiden“ (FLEISCHER/BARZ 2012, 30). Es wird davon ausgegangen, dass die Textsorte „Songtext“ sich durch spezifische Wortbildungsprodukte auszeichnet, die in anderen Texten nicht vorkommen bzw. nicht funktionieren und bei den Rezipienten auch nicht auf Akzeptabilität stoßen würden. Okkasionelle Wortbildungen mit textdistinktiver Funktion werden von Fleischer/Barz (2012, 31) unter dem Schlagwort der „stilbildende[n] Potenz der Wortbildung“ diskutiert und wie folgt charakterisiert: „Okkasionelle Wortbildungen […] tragen zur Steigerung der Expressivität bei, regen an und unterhalten“ (FLEISCHER/BARZ 2012, 32). In Kapitel 3 wird noch deutlich werden, inwiefern die Bildungen aus dem Songkorpus durch eben dieses Charakteristikum einen Beitrag zur Spezifik von Songtexten leisten.

Die hier im Fokus stehenden kreativen Wortbildungen lassen sich in theoretischer Hinsicht innerhalb der Domäne der „Extravagant Morphology“ (EITELMANN/HAUMANN 2022) verorten. Der Begriff „extravagant“ wird dabei im Sinne Haspelmaths (1999) verwendet[1] und ist an die so genannte ‘Maxime der Extravaganz’ geknüpft, welche von Haspelmath (1999, 1055) im Kontext von Grammatikalisierungsprozessen angeführt und wie folgt definiert wird: “Talk in such a way that you are noticed”. Die Maxime der Extravaganz ist eine von insgesamt fünf Maximen, die für den Sprachwandel eine Rolle spielen, und spiegelt die Beobachtung wider, dass es bei einer erfolgreichen Kommunikation nicht nur darum geht, verstanden zu werden (vgl. EITELMANN/HAUMANN 2022, 2): „[…] the maxim of extravagance seeks to grasp the observation that communicative behaviour does not boil down to achieving the goal of being understood – a goal that would be sufficiently achieved by adhering to the maxims of clarity and conformity. In order to be noticed, language users would also overstep boundaries and ‘choose new ways of saying old things’ (HASPELMATH 1999, 1057)” (EITELMANN/HAUMANN 2022, 2).

In ihrem kürzlich erschienenen Sammelband “Extravagant Morphology“ wenden Eitelmann und Haumann (2022) den bisher auf die Ebene der Grammatikalisierung bezogenen Begriff der ‚extravagance‘ erstmals auf den Bereich der Wortbildung an. Dabei bestehen enge Bezüge zum Begriff der „linguistic creativity“, der auch im Titel dieses Beitrags („kreative Wortbildungen“) erscheint: „Creativity in a more narrow, linguistic sense captures innovative language use that goes beyond the productive application of rules, often associated with a ludic quality” (EITELMANN/HAUMANN 2022, 4). Genau darum – dies haben die bisherigen

Ausführungen zur Natur von Songtexten und der Funktion von okkasionellen Wortbildungsprodukten bereits deutlich gemacht – scheint es bei Songtexten u.a. zu gehen.

Wie sich Kreativität bzw. Extravaganz in der Wortbildung konkret manifestiert, wird im folgenden Kapitel durch eine detaillierte linguistische Analyse von Wortbildungsprodukten im Songkorpus exemplarisch demonstriert.

3 Empirische Untersuchung

Der Fokus der nachstehenden empirischen Untersuchung liegt auf der Analyse kreativer Wortbildungsprodukte im Songkorpus. Dabei wird nicht nur herausgearbeitet und illustriert, welche Typen kreativer Wortbildungsprodukte in Songtexten vorkommen, sondern auch der Frage nachgegangen, wodurch genau solche Bildungen Aufmerksamkeit wecken: Resultiert die spezifische Qualität der Songkorpus-Wortbildungen v.a. aus dem Rückgriff auf markierte Wortbildungsmuster, und/oder aus einer auffälligen Lexik innerhalb der Wortbildungsmuster? Diskutiert werden soll dabei außerdem, ob und inwiefern sich die komplexen Wörter im Songkorpus – bzw. die beobachtbaren Verteilungsverhältnisse der Wortbildungstypen – grundlegend von anderen klassischeren Textsorten wie Zeitungstexten unterscheiden.[2]

3.1 Vorgehen

Um bei der Analyse von Anfang an mit potentiellen Kandidaten für kreative Wortbildungsprodukte zu arbeiten, wird die Untersuchungsbasis zunächst auf die Liste der eintausend längsten Wörter im Songkorpus (‚longwords') beschränkt.[3] Wortlänge wird somit als Indikator für Kreativität betrachtet. Der Vorteil daran ist, dass es sich dabei um ein verhältnismäßig objektives Kriterium handelt, welches sich empirisch gut operationalisieren lässt. Natürlich ist die Heranziehung der ‚longwords' aber keineswegs ein Garant für das Herausfiltern kreativer Wortbildungen – wie noch ersichtlich werden wird, enthält das Untersuchungskorpus auch lexikalisierte Wörter – die okkasionellen Bildungen sind aber eindeutig in der Überzahl. Das gängige Kriterium zur Identifizierung okkasioneller Bildungen – der sogenannte Status als Hapax Legomenon – kann innerhalb des Songkorpus nicht ohne Weiteres angewendet werden. Wenn überhaupt, müsste für die komplexen Wörter des Songkorpus anhand eines Referenzkorpus überprüft werden, ob es sich tatsächlich um Einmalbildungen handelt. Dieser Aufwand steht in der vorliegenden Untersuchung jedoch in keinem Verhältnis zum Zweck.[4]

Die zugrunde gelegte ‚longwords'-Liste enthält neben den Einzelwörtern auch Angaben zu ihrer Frequenz (im Songkorpus), sowie dem Archiv, in dem sie auftreten. Außerdem ist für jedes Belegwort auch der Satzkontext angegeben. Letzterer ist, insbesondere für ein adäquates Verständnis von Ad-hoc-Wortbildungen, nicht zu unterschätzen und wird daher in den Beispielübersichten jeweils mitgeliefert – ebenso wie der konkrete Künstlername bzw. das Archiv, aus dem die jeweilige Bildung stammt. Da diese Liste der eintausend längsten Wörter jedoch zu lang für eine detaillierte Wortbildungsanalyse ist, wird die Untersuchungsbasis noch weiter eingegrenzt, und zwar auf alle Wörter mit einer Wortlänge von mindestens 24 Buchstaben. Dieses Kriterium wurde nach der groben Sichtung aller 'longwords' festgelegt,

da dies eine Grenze zu sein scheint, ab der die komplexen Wörter „uninteressanter“, das heißt weniger kreativ werden. Durch die Anwendung dieses ‚Mindestens-24-Buchstaben-Kriteriums‘ ergibt sich ein Untersuchungskorpus mit insgesamt 183 Wörtern. Diese werden im folgenden Unterkapitel (s. Kapitel 3.2) zunächst nach Wortbildungstypen klassifiziert und gegebenenfalls in Bezug auf weitere formale und/oder semantische Auffälligkeiten kommentiert. Zu diesem Zweck wird zum einen auf die gängigen Wortbildungstypen aus Fleischer/Barz (2012) zurückgegriffen, zum anderen aber auch auf markiertere Typen der „phrasalen Wortbildung“ (LAWRENZ 2006; vgl. HEIN 2015) deren Ausgangseinheiten nicht lexemischer, sondern eben phrasaler Natur sind.

3.2 Analyse: Wortbildungstypen im Songkorpus

Klassifiziert man die 183 Bildungen des Untersuchungskorpus hinsichtlich ihres Wortbildungstyps, dominiert mit einem Anteil von ca. 56 % auch im Songkorpus der Prototyp der Komposition, nämlich die nominale Determinativkomposition (z.B. *Lebensänderungsschneiderei*, vgl. Tabelle 1); adjektivische Determinativkomposita wie *erdstrahlungsabweisend* sind erwartungsgemäß deutlich weniger frequent (4 Bildungen). Andere zentrale Wortbildungsarten kommen – im Vergleich zu klassischen Textsorten – auffallend selten vor: Bei nur 4 Bildungen des Untersuchungskorpus handelt es sich um Derivate (z.B. *unaufhaltbar*), Konvertate sind nicht enthalten. Drei Bildungen sind als adjektivische Kopulativkomposita einzuordnen, z.B. *schuppenschultrig-selbstgerecht*.

WB-Typ	# Types	Beispiel
Determinativkompositum (N)	102	*Lebensänderungsschneiderei*
Phrasenkompositum (N)	51	*Beckham-im-Gedächtnis-Grätsche*
Phrasenkonvertat	8	*Saufen-bis-der-Arzt-kommt*
Phrasenderivat (N; A)	9	*Schlachtvieh-auf-Tiertransporter-Zerrer; Candle-in-the-Wind-mäßig*
Derivat (N; A)	4	*Beliebigkeit ; u.n.a.u.f.h.a.l.t.b.a.r*
Determinativkompositum (A)	4	*erdstrahlungsabweisendem*
Kopulativkompositum (A)	3	*schuppenschultrig-selbstgerechte*
Phrasenkompositum (A)	2	*Oma-Schlüpfer-riechendes*
	183	

Tabelle 1: Wortbildungstypen im Untersuchungskorpus.

Auffallend häufig, und im Vergleich zu anderen Textsorten wie z.B. Zeitungstexten deutlich erhöht, ist der Anteil, den sogenannte „phrasale Wortbildungen“ innerhalb der analysierten Songkorpus-Bildungen ausmachen. Es handelt sich dabei um „neuere Wortbildungstypen“, zu denen gemäß Lawrenz (2006, 1) „unterschiedliche Typen von Phrasenkomposition, Phrasenderivation und Phrasenkonversion“ gehören. Mit einem Anteil von ca. 28 % sind nominale Phrasenkomposita wie *Beckham-im-Gedächtnis-Grätsche* sogar der zweithäufigste Wortbildungstyp im Untersuchungskorpus, häufiger sind nur prototypische Determinativkomposita. In zwei Fällen handelt es sich um adjektivische Phrasenkomposita, z.B. *Oma-Schlüpfer-riechendes (Miststück)*, ein Phänomen, das für das Deutsche bisher nicht (empirisch) untersucht wurde (vgl. GÜNTHER/KOTOWSKI/PLAG 2018 für das Englische). Darüber lassen sich im Songkorpus auch die beiden anderen Wortbildungstypen nachweisen, die Lawrenz zusammen mit der Phrasenkomposition unter dem Dach der „phrasalen Wortbildung“ beschreibt, nämlich

nominale und adjektivische Phrasenderivate (9 Belege, z.B. *Candle-in-the-Wind-mäßig*) und Phrasenkonvertate (8 Belege, z.B. *Saufen-bis-der-Arzt-kommt*).

Hervorzuheben ist, dass der Anteil von Phrasenkomposita – insbesondere gegenüber prototypischen Komposita aus nicht-phrasalen Konstituenten – im Untersuchungskorpus ungewöhnlich hoch ist. Während der Anteil von „Komposita mit Wortgruppen oder Sätzen bzw. Satzfragmenten (als 1. Konstituente)“ gemäß des Standardwerks von Ortner et al. (1991, 4:6) innerhalb der Gruppe der nominalen Komposita bei 3,2 % liegt, sind Phrasenkomposita im Songkorpus halb so häufig wie prototypische Komposita. Insgesamt kann man davon ausgehen, dass phrasale Einheiten hier bewusst zur Bildung komplexer Wörter herangezogen werden, um Expressivität zu triggern – eben weil solche Bildungen vom Erwarteten abweichen.

Wie die nachstehende (s. Kapitel 3.2.1 und 3.2.2) feinkörnige formale und semantische Analyse der frequenten Wortbildungstypen aus Tabelle 1 zeigt, wird im Songkorpus aber auch durch die Verwendung klassischer Wortbildungstypen Aufmerksamkeit erregt. Markierte Wortbildungstypen sind also offenbar nur eine mögliche Quelle für die Schöpfung kreativer komplexer Wörter.

3.2.1 Klassische Wortbildungstypen: Komposita & Derivate

Wie bereits erwähnt, bilden nominale Determinativkomposita mit einem Anteil von ca. 56 % im Songkorpus den gängigsten Wortbildungstyp. Adjektivische Determinativkomposita (s. Übersicht 1) hingegen spielen mit einem Anteil von nur ca. 2%, wie in klassischen Textsorten auch, eine deutlich untergeordnete Rolle.

1. Da steht " Reinigt die Aura, schützt die Haut vor Verfall, mit **erd-strah-lungs-ab-wei-sendem** Metall.“ Das Konzept, es geht auf, sie verkaufen wie blöd auf Messen und munter Events rauf und runter, bei Psychogetanze, mit ei'm Wort, das ganze scheinesoterische Sülz-Alphabet. (Stoppok)[5]
2. Vielleicht war es der Messias, der nach zweitausend Jahr'n Noch mal gekommen ist, und du hast ihn nicht gefahr'n, Mit deinem chromblitzenden, **air-condition-daunenweichen** Thron. (Mey/Misc)

Übersicht 1: Adjektivische Determinativkomposita im Songkorpus.

Das adjektivische Kompositum in (1) fällt durch seine unkonventionelle Schreibung, genauer gesagt durch die Abtrennung der einzelnen Silben durch Bindestriche, auf und wird im Kontext einer ironischen Charakterisierung des Runs auf esoterische Produkte und Events verwendet. Das Erstglied sättigt dabei eine durch das deverbale Zweitglied eröffnete Leerstelle (was wird abgewiesen?). In Beispiel (2) ist nicht ganz eindeutig, ob die unmittelbaren Konstituenten in einem subordinierten Verhältnis stehen, also ob „daunenweich“ durch „aircondition“ näher bestimmt wird (z.B. ‚weich durch aircondition‘), oder ob die beiden Teile in einem koordinativen Verhältnis stehen (z.B. ‚aircondition und daunenweich‘).

Der Großteil der Determinativkomposita im Songkorpus ist nominal. Von den 102 nominalen Determinativkomposita können 19 Bildungen als lexikalisiert gelten, z.B. *Sehnenscheidenentzündung*, *Minderwertigkeitskomplexe*, *Führerscheinzulassungsstelle* usw. Solche Bildungen werden an dieser Stelle ausgeklammert, weil sie in Verbindung mit dem Thema ‚Kreativität' nicht unmittelbar relevant sind. Die Tatsache, dass es sich bei den restlichen 83 Gesamtkomplexen um Ad-Hoc-Bildungen[6] handelt, ist ein klares Indiz für den kreativen Charakter der komplexen Wörter im Songkorpus zu werten. Es ist davon auszugehen, dass der vermehrte Gebrauch solcher Ad-hoc-Bildungen kein Zufall ist, sondern eben dazu dient, Aufmerksamkeit zu erzielen. Ad-hoc-Bildungen können somit auch als sprachliches Mittel zur Umsetzung künstlerischer Intentionen gewertet werden.

Nachstehend (s. Übersicht 2) werden die nominalen Determinativkomposita des Untersuchungskorpus genauer betrachtet. Dabei steht die Frage im Vordergrund, wie es in Songtexten gelingt, mit einem recht konventionellen Bildungsmuster wie der nominalen Determinativkomposition dennoch expressive oder witzige Effekte zu erzielen.

Vorweg ist anzumerken, dass Okkasionalität bereits als eine Ursache für expressive Effekte zu betrachten ist – schließlich erzeugt Noch-nicht-Gehörtes ohne Frage mehr Aufmerksamkeit als der Rückgriff auf solche Bildungen, die bereits fest im Sprachgebrauch etabliert sind. Welche Ursachen für Expressivität lassen sich darüber hinaus feststellen?

Einige der nicht-lexikalisierten Bildungen erregen allein durch ihre Komplexität bzw. Länge Aufmerksamkeit, so z.B. *Autobahnraststättenspielplatzklettergerüst* in (3). Diese Bildung ist mit 42 Buchstaben nicht nur die Längste unter den okkasionellen determinativen Komposita, sondern auch in morphologischer Hinsicht komplex, da es sich um ein rekursives Kompositum handelt, dessen unmittelbare Konstituenten selbst wiederum Komposita sind. Andere Bildungen hingegen, exemplarisch illustriert in (4) und (5), fallen durch Reduplikationen im Erstglied auf. Man kann sich gut vorstellen, inwieweit Komposita mit Reimdopplungen wie *Krickelkrackelzickzackland* gezielt für Rhythmik und Klang von Songs eingesetzt werden.

3. Du traust dich nicht nach Haus und sitzt hier schon seit Stunden oben auf dem **Autobahnraststättenspielplatzklettergerüst**. (Fettes Brot)
4. **Krickelkrackelzickzackland** und keine geraden Geraden. (Fettes Brot)
5. Teddi, du wirbelst los mit deinen **Schlotter-Schlacker-Gummibeinen**. (Udo Lindenberg)
6. Die Absage könnte auch gut von einem deutschen Unternehmen sein: "Leider stellen wir zu keinem Zeitpunkt, weder jetzt noch später, **Selbstmordattentatsbewerber** über fünfundfünfzig ein." (Hannes Wader)
7. Hey Baby, ich sag Goodbye zur **Lebensänderungsschneiderei**, ich bin doch kein Schnarcho. (Udo Lindenberg)
8. Jede **Beautybloggerhurentocher** plus alle Rapper. (Frauenarzt/HipHop)
9. Und wenn, nur ganz kleines bisschen wie 'ne **Bremsstreifen-Boxershorts**. (257er/HipHop)
10. Es ist das **Rapdeutschlandkettensägenmassaker**. (K.I.Z./HipHop)

11. Laptop, Rapgott, **Lederjacken-Prollschiene**, Ersguterjunge, yeah mein Label eine Goldmine. (Bushido/Charts)
12. Du zeigtest Dich betroffen von der **Zeitverfluggeschwindigkeit**. (Tocotronic/Misc)
13. Aber stolpernd folge ich dir durch deine **Kindheitserinnerungswälder** bis es dunkel wird und dann folgst du mir. (Element of Crime)
14. **Kosmetikartikelüberschuss** und Demokratiedefizit. (Maeckes/HipHop)

Übersicht 2: Nominale (okkasionelle) Determinativkomposita im Songkorpus.

Eine weitere Besonderheit, die im Gegensatz zu den in (3) bis (5) skizzierten Typen nicht strukturell fassbar ist, liegt in der (grotesken) Kombination von semantisch-thematischen Konzepten, die gemäß Weltwissen nichts miteinander zu tun haben bzw. nicht zusammenpassen. Es handelt sich dabei um eine der Hauptursachen von Expressivität innerhalb der Gruppe der nominalen Determinativkomposita. Die Bildung *Selbstmordattentatsbewerber* in (6) ist insofern grotesk, als man sich für die Durchführung eines Selbstmordattentats wohl üblicherweise nicht auf klassischem Wege bewirbt; ein Blick auf den Gesamtbeleg zeigt, dass die Bildung insgesamt ironisch eingebettet ist. Ebenso *Lebensänderungsschneiderei* in (7): Üblicherweise ist das im Erstglied genannte Konzept ‚Leben' nichts, das man analog zu Kleidungstücken nach Wunsch in einer Änderungsschneiderei (Zweitglied) abändern kann.

Ein Aufmerksamkeitstrigger, der sich durch alle Wortbildungstypen zieht, ist außerdem der Rückgriff auf vulgäre Lexik (z.B. *Beautybloggerhurentochter* in (8) oder *Bremsstreifen-Boxershorts* in (9)) sowie die Verwendung brutaler (z.B. *Rapdeutschlandkettensägenmassaker* in (10)) bzw. umgangssprachlicher Lexik (z.B. *Lederjacken-Prollschiene* in (11)). Der expressive Charakter solcher Belege hat weniger damit zu tun, dass es sich um morphologisch komplexe Wörter handelt, sondern ist vielmehr auf die Wortwahl zurückzuführen. Beleg (11) zeigt anschaulich und exemplarisch, dass natürlich auch die Einbettung der Wortbildungsprodukte wiederum Expressivität erzeugen bzw. die in den komplexen Wörtern bereits angelegte Expressivität noch erhöhen kann: *Lederjacken-Prollschiene* ist hier nicht nur Teil einer rhythmisch wirksamen Aneinanderreihung von Komposita, sondern reimt sich zusätzlich mit dem Ende (*Goldmine*) der Songzeile.

Neben diesen v.a. durch negativ belastete Lexik auffallenden Bildungen finden sich im Songkorpus mit *Zeitverfluggeschwindigkeit* (12) und *Kindheitserinnerungswälder* (13) aber auch Komposita, die schöne Bilder zeichnen und von fast poetischer Natur sind. Das Vorkommen solch poetisch anmutender Komposita könnte durchaus an bestimmte Archive gekoppelt sein, schließlich stammen die Bildungen in (12) und (13) erwartungsgemäß nicht aus dem HipHop-Genre, sondern aus Songtexten von ‚Element of Crime' und ‚Tocotronic'. Insgesamt sind solche positiv-poetischen Komposita in dem analysierten Songkorpus-Ausschnitt jedoch gegenüber den Bildungen mit vulgärer oder brutaler Lexik in der Minderheit, was sicherlich eher der Zusammensetzung des Korpus als der grundsätzlichen Natur von Songtexten geschuldet ist. Beleg 14 zeigt schließlich, wie das an sich bereits negativ konnotierte Kompositum *Kosmetikartikelüberschuss* durch die Kontrastierung mit dem entgegengesetzten Muster X-*Defizit* (*Demokratiedefizit*) in seiner abwertenden Wirkung noch verstärkt wird. Durch

die direkte Gegenüberstellung der beiden Komposita wird hier eine Kritik zum Ausdruck gebracht: Während es eher unwichtige Dinge wie Kosmetik im Überfluss gibt, ist ein so wichtiges (geistiges) Gut wie die Demokratie hingegen Mangelware.

Neben den zahlreichen Determinativkomposita, die exemplarisch in den Übersichten 1 und 2 illustriert sind, befinden sich unter den ‚longwords' des Songkorpus nur 4 Derivate (s. Übersicht 3). Dies ist aus der Perspektive der Wortbildung in jedem Fall als überraschender Befund zu werten, da die Derivation zu den zentralen Wortbildungsarten des Deutschen zählt und in anderen klassischen Textsorten quantitativ nicht weniger bedeutsam als die prototypische Determinativkomposition ist. Wie dieser Befund zu deuten ist, bleibt unklar. Entweder besteht in einer speziellen Textsorte wie den Songtexten einfach generell ein geringeres Bedürfnis für den Rückgriff auf diesen Wortbildungstyp, oder Derivate fallen aufgrund des hier angesetzten Kriteriums der Wortlänge von mindestens 24 Buchstaben durch das Raster.

Die Anzahl der Derivate innerhalb der ‚longwords' ist nicht nur gering, darüber hinaus sind auch nur zwei der insgesamt vier Derivate für den Fokus dieses Beitrags auf kreative Wortbildungsprodukte relevant. Die lexikalisierten Bildungen *betriebswirtschaftlich* und *Beliebigkeit* werden daher an dieser Stelle nicht näher betrachtet. Übersicht 3 zeigt die beiden okkasionellen Derivate im Songkorpus; beide sind adjektivisch.

15. Alle wollen die Box mit dem **hörspielentertainmenthaften**, nicen Stuff auf dem Hoodtape. (Kollegah/HipHop)
16. Denn wir sind **u.n.a.u.f.h.a.l.t.b.a.r.** (Azad/HipHop)

Übersicht 3: Adjektivische (okkasionelle) Derivate im Songkorpus.

Die Ableitung *hörspielentertainmenthaft* in (15) wirkt auffällig, obwohl es sich in grammatischer Hinsicht um eine regelgemäße Verwendung des Ableitungsmusters X-*haft* handelt, welches laut Fleischer/Barz (2012, 336) hauptsächlich mit substantivischen Basen auftritt und sowohl zur Ableitung komplexer als auch simplizischer Basen dient. Eventuell lässt sich die Markiertheit der „Vergleichsbildung" *hörspielentertainmenthaft* darauf zurückführen, dass das komplexe Basiswort, das mit *-haft* abgeleitet wird, aus einem englischsprachigen Grundwort und einem deutschsprachigen Bestimmungswort besteht.

Die Ableitung *unaufhaltbar* in (16) fällt in der verschriftlichen Version des Songtextes zunächst durch die Verwendung von Punkten zwischen den Einzelbuchstaben auf – vermutlich wird damit auf die akustische Buchstabe-für-Buchstabe-Ausbuchstabierung im Song hingewiesen. Die Bildung wirkt deswegen auffällig, weil mit *unaufhaltsam* eine lexikalisierte Parallelform/Konkurrenzbildung zu *unaufhaltbar* existiert (vgl. FLEISCHER/BARZ 2012, 348); die Ableitung mit dem Suffix *–sam* ist in diesem Fall daher als markiert zu betrachten.

3.2.2 Phrasenkomposita und andere phrasale Wortbildungen

Die longwords des Songkorpus enthalten insgesamt 53 Phrasenkomposita - 96 % davon sind nominal, so dass der Schwerpunkt der nachstehenden Analyse auf der nominalen Phrasenkomposition liegt. Es ist davon auszugehen, dass dies, analog zur prototypischen Determinativkomposition, das produktivste Submuster der Phrasenkomposition darstellt.

Zugleich untermauern die beiden adjektivischen Bildungen in Übersicht 4 aber, dass das Bildungsmuster im Deutschen grundsätzlich auch zur Bildung von Adjektiven herangezogen werden kann (vgl. LAWRENZ 2006, 7). Auffallend ist hier, dass es sich in beiden Fällen nicht um ‚echte Adjektive', sondern um Partizipialattribute handelt. Außerdem dienen beide Bildungen in Übersicht 4 dazu, eine Person genauer zu charakterisieren bzw. – wie der Belegkontext zeigt – abzuwerten.

17. **Dennoch-für-die-Zukunft-sparende** Deutsche. (Favorite/HipHop)
18. Du **nach Oma-Schlüpfer-riechendes** Miststück. (Die Ärzte/Misc)

Übersicht 4: Adjektivische Phrasenkomposita im Songkorpus.

Unterzieht man die nominalen Phrasenkomposita des Songkorpus einer genaueren syntaktischen Analyse, fällt ein grundlegender Unterschied zur Phrasenkomposition in der geschriebenen Zeitungssprache (DeReKo) (vgl. HEIN 2015) und in der gesprochenen Sprache (FOLK-Korpus) (vgl. HEIN/ANTONIOLI in Vorb.) auf: Sowohl in DeReKo (LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE 2011) als auch im FOLK-Korpus (vgl. LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE 2022) bilden Nominalphrasen als Erstglieder mit einem Anteil von ca. 80 % den dominierenden Formtypen. Satzähnliche Gebilde hingegen, deren Integration in Komposita gemäß Hein (2015) zu expressiven Gesamtkomplexen führt, sind das zweithäufigste Erstglied, treten mit einem Gesamtanteil von ca. 18 % (DeReKo) bzw. 11 % (FOLK) aber deutlich seltener auf. Im Songkorpus hat die Mehrheit der nominalen Phrasenkomposita zwar ebenfalls eine Nominalphrase als Erstglied, allerdings sind Letztere mit einem Anteil von ca. 61 % hier längst nicht so dominierend wie in DeReKo oder FOLK. Satzähnliche Einheiten bilden auch im Songkorpus den zweithäufigsten Erstgliedtyp, sind aber mit einem Anteil von 35 % deutlich präsenter als in den vorhergehenden Untersuchungen (s. Übersichten 5 und 6).

Wenn mit Hein (2015) davon ausgegangen wird, dass das Bildungsmuster der Phrasenkomposition aus zwei Submustern besteht, von denen eines eher unmarkiert ist (Phrase ohne den Status einer kommunikativen Minimaleinheit als Erstglied, z.B. *Fünf-Tage-Woche*) und eines expressiv bzw. markiert ist (Phrase mit dem Status einer kommunikativen Minimaleinheit oder satzähnliches Gebilde als Erstglied, z.B. *Zu-mir-oder-zu-dir-Gequatsche*)[7], lässt sich allein aus den quantitativen Verhältnissen im Songkorpus bereits Folgendes ableiten: Nicht nur, dass der Anteil phrasaler Wortbildungen bzw. Phrasenkomposita im Besonderen hier gegenüber klassischeren Textsorten deutlich erhöht ist und als Indikator für die Kreativität der Wortbildungsprodukte im Songkorpus verstanden werden kann. Auch ein genauerer Blick auf die enthaltenen Phrasenkomposita deutet auf einen erhöhten Anteil solcher Formtypen hin, die

als besonders expressiv gelten. Als Zwischenfazit kann daher festgehalten werden, dass der Rückgriff auf markierte Wortbildungstypen in jedem Fall als eine Ursache für den expressiven Charakter der komplexen Wörter im Songkorpus zu betrachten ist. Eine detailliertere Analyse der nominalen Phrasenkomposita zeigt jedoch u.a., dass auch an sich bereits markierte Formtypen durch andere Faktoren in ihrer Expressivität noch verstärkt werden, bzw. dass relativ unmarkierte Formtypen dank anderer Faktoren dennoch als expressiv wahrgenommen werden.

19. Hält mich fest in ihrer Hand, im **Vater-Mutter-Kinder-Land**. (Tocotronic/Misc)
20. Jeden Morgen ein **Drei-Minuten-Frühstücksei** und eine Runde mit dem Hund; Pünktlich bei der Arbeit sein, pünktlich wieder Schluss; Jeden Tag in die gleiche Richtung, ohne zu fragen, wieso; Jede Nacht dieselben Gesichter in denselben Fernsehshows. (Die Toten Hosen/Misc)
21. Ich bin der Vater von mehreren **Sapiens-Haflinger-Kreuzungen**. (Pimpulsiv/HipHop)
22. Kralle Kralle Kralle, und jetzt machen sie dich alle, mit deinem **Chauvi-Macho-Macker-Stuss**. (Udo Lindenberg)
23. Rudi läuft in die erste **Beckham-im-Gedächtnis-Grätsche**. (Sido/HipHop)
24. Ersguterjunge **Schwanz-in-den-Mund-Style**. (Bushido/HipHop)

Übersicht 5: Nominale Phrasenkomposita im Songkorpus (Nominalphrase als Erstglied).

Übersicht 5 illustriert exemplarisch den im Songkorpus dominierenden Formtypen der Phrasenkomposition, nämlich Bildungen mit einer Nominalphrase in Erstgliedposition. Die Bildungen in (19) und (20) erregen kaum Aufmerksamkeit und wären auch außerhalb des spezifischen Kontextes von Songtexten denkbar. Bei *Drei-Minuten-Frühstücksei* in (20) liegt im Erstglied eine durch ein Numerale erweiterte Nominalphrase vor – eine Spielart der Nominalphrase, die in DeReKo (vgl. HEIN 2015, 446) das häufigste Erstglied bildet (über 20 % des Untersuchungskorpus). Im Songkorpus hingegen ist dieser Typus nur 5-mal vertreten – dies könnte darauf zurückzuführen sein, dass er sich aufgrund seiner „Unauffälligkeit" nur wenig für die Ziele und Ausdrucksbedürfnisse von Songtexten eignet.

Vater-Mutter-Kinder-Land in (19) weist, genau wie die komplexen Wörter in (21) und (22), in Erstgliedposition implizit koordinierte Nominalphrasen auf (vgl. HEIN 2015, 194 f.). Während die Bildung in (19) aufgrund der ‚Kompatibilität' der im Erstglied koordinierten Konzepte (Vater, Mutter und Kind) kaum auffällig ist, sind die formal vergleichbaren Bildungen in (21) und (22) deutlich expressiver. Deren Markiertheit ist weniger auf das Bildungsmuster selbst zurückzuführen, sondern eher auf die groteske Koordination von divergierenden Konzepten und Sprachstilen (*Sapiens-Haflinger-Kreuzungen*) bzw. auf die klangvolle Kopplung von Konstituenten (*Chauvi-Macho-Macker-Stuss)*. Eine expressive Wirkung kann aber abgesehen von der Koordination selbst auch durch die einzelnen Lexeme hervorgerufen werden (vgl. Kapitel 3.2.1), z.B. durch deren umgangssprachlichen Charakter. Dies gilt z.B. für *Chauvi-Macho-Macker-Stuss*. Bei den Bildungen in (23) und (24) handelt es sich syntaktisch gesehen um präpositional erweiterte Nominalphrasen. Aufmerksamkeit wird hier durch die Aussage selbst (*Beckham-im-Gedächtnis-Grätsche*) bzw. durch den vulgären Charakter (*Schwanz-in-den-Mund-Style*) erzielt.

Übersicht 6 zeigt exemplarisch Phrasenkomposita mit einem Satz oder einer satzähnlichen Einheit als Erstglied.

25.Sind stets auf der Suche nach dem nächsten Kick Mit diesem verwegnen **Geht-nicht-gibt's-nicht-Blick** Bohrn sie furchtlos und behende Löcher in Tische und Wände, Überschwemmen, legen Brände Und bringen nie etwas zu Ende – Sind so kleine Hände! (Mey/Misc)
26. Aber das hier ist **Ich-finde-du-bist-weak-Rap**. (Juse Ju/HipHop)
27.Und ich werd ' nicht müd ', den Reichtum und die Launen Und den Aberwitz der Schöpfung zu bestaunen: Kugelfisch, Rohrdommel, Steinlaus, Milbe, Maibock, doch indes, Die schönste , bunte Vielfalt hat das menschliche Gesäß : Es gibt dicke Pöter, Und todschicke Pöter, Es gibt selbstbewusste **„ich-fang'-alle-Blicke"-Pöter**. (Mey/Misc)
28. Ich fühl mich wie ein **Plakatier'n-Verboten-Plakat**. (Kunze/Misc)
29.**Talking-Böser-Traum-Blues** (Hannes Wader)
30.Du machst nicht mit beim **Liter-Brennendes-Öl-Trinken-Wettbewerb**? (Trailerpark/HipHop)
31.In feindlichen **Schützengräben-explodier-Granaten-Style**. (Azad/HipHop)

Übersicht 6: Nominale Phrasenkomposita im Songkorpus (Satz/satzähnliche Einheit als Erstglied).

Zunächst zur syntaktischen Analyse: In (25) bis (27) ist jeweils ein Aussagesatz enthalten – in zwei Fällen ist dieser jeweils in der ersten Person Singular, d.h. aus der Ich-Perspektive, formuliert (*Ich-finde-du-bist-weak-Rap* und *„ich-fang'-alle-Blicke"-Pöter*). Das Erstglied von *Plakatier'n-Verboten-Plakat* in (28) ist als Satzellipse zu klassifizieren, die trotz des fehlenden Verbs als kommunikative Minimaleinheit fungiert (vgl. HEIN 2015, 203f.).
In syntaktischer Hinsicht schwieriger zu fassen sind hingegen die Phrasenkomposita in (29) bis (31). Im Fall von *Talking-Böser-Traum-Blues* wird die Analyse durch die Kombination englischer und deutscher Lexik bzw. der Satzstellungsregeln der beiden Sprachen erschwert. Eine einsprachige, syntaktisch vollständige Auflösung des elliptischen Erstglieds könnte „Hier-spricht-dein-böser-Traum-Blues" sein. Bei den Bildungen in (30) und (31) ist in Erstgliedposition jeweils eine Verbgruppe realisiert, die aus einem Verbalkomplex und den entsprechenden Komplementen besteht. Formale Besonderheiten liegen hier insofern vor, als die Verbgruppen jeweils nicht vollständig sind: Während in *Liter-Brennendes-Öl-Trinken-Wettbewerb* in (30) das flektierte Numeral fehlt („Einen-Liter-brennendes-Öl-Trinken"), enthält das Erstglied in (31) als verbales Element lediglich einen Verbstamm. Zumal ist anhand der nicht durchgängigen Markierung mit Bindestrichen in *In feindlichen Schützengräben-explodier-Granaten-Style* unklar, welche Elemente tatsächlich Teil des Erstglieds sind. Das Phänomen syntaktisch unvollständiger bzw. syntaktisch nur schwer kategorisierbarer Erstglieder ist jedoch keine exklusive Eigenschaft der Songkorpus-Bildungen, sondern wurde auch für Phrasenkomposita in DeReKo (vgl. HEIN 2015, 199–201) bereits nachgewiesen. Eventuell könnte

man hier von einem bewussten Einsatz ungrammatischer bzw. grammatisch nicht vollständiger Strukturen sprechen, da auch diese wiederum Aufmerksamkeit auf Seiten des Hörenden wecken.

Auch jenseits ihrer formalen Eigenschaften sind einige der Phrasenkomposita in Übersicht 6 besonders interessant bzw. kommentierungswürdig. Der Beleg in (25) ist einer der wenigen Belege im Songkorpus, in denen das Erstglied verfestigt und somit im weitesten Sinne bereits „vorgefertigt" ist – ein Phänomen, das für die Phrasenkomposition insgesamt eine wichtige Rolle spielt (vgl. STEYER/HEIN 2018). Konkret geht „*Geht's nicht, gibt's nicht*" in (25) auf einen Werbeslogan des Praktiker-Baumarkts zurück. Die Tatsache, dass in den Phrasenkomposita des Songkorpus offenbar kaum auf Vorgefertigtes zurückgegriffen wird, kann als Indiz für eine erhöhte Kreativität der Bildungen gedeutet werden – schließlich beläuft sich der Anteil verfestigter Erstglieder in den Phrasenkomposita aus DeReKo auf ca. 65 % (vgl. HEIN 2015, 450f.).[8]

Auch bei den Zweitgliedern gibt es, wie die rückläufige Sortierung der insgesamt 183 Phrasenkomposita des Songkorpus zeigt, einen hohen Grad an Kreativität, bzw. wenig Wiederholung. Lediglich die Zweitglieder *Rap*, *Style*[9] und *Pöter* treten mehrfach auf. *Ich-finde-du-bist-weak-Rap* in (26) steht exemplarisch für das Muster [XP-*Rap*]. Zugleich handelt es sich auch hier wieder um ein Beispiel, das zeigt, inwiefern die Vermischung mehrerer Sprachen im Erstglied expressive Effekte triggern kann. Die Bildung in (27) – *„ich-fang'-alle-Blicke"-Pöter* – ist eines von diversen Phasenkomposita mit dem Zweitglied *Pöter*, das, wie auch aus dem Volltextbeleg ersichtlich wird, umgangssprachlich im Sinne von „Gesäß" verwendet wird. Hier zeigt sich einmal mehr, dass Kreativität sich nicht nur durch die Wahl markierter Wortbildungstypen oder formal/semantisch auffälliger Erstglieder, sondern auch durch die Lexemwahl in Zweitgliedposition manifestiert.

Die expressive bzw. witzige Wirkung von *Plakatier'n-Verboten-Plakat* in (28) ist auf den enthaltenen Widerspruch zurückzuführen: An einer Stelle, an der das Plakatieren explizit verboten ist, sollten erst gar keine Plakate zu finden sein. Zudem weckt auch der im Volltextbeleg gezogene Vergleich zwischen dem Gefühlszustand des Singenden und einem Plakat, welches das Plakatieren verbietet, Aufmerksamkeit (*„Ich fühl mich wie ein Plakatier'n-Verboten-Plakat"*). Die beiden Phrasenkomposita in (30) und (31) aus Übersicht 6 fallen insbesondere wegen der Brutalität der damit transportierten Bilder auf. *Schützengräben-explodier-Granaten-Style* enthält eine Anspielung auf Kriegsmetaphorik; *Liter-Brennendes-Öl-Trinken-Wettbewerb* ist zusätzlich zu dem brutalen Bild in Erstgliedposition durch die Absurdität der Gesamtbedeutung gekennzeichnet (‚Wettbewerb, in dem es darum geht, brennendes Öl zu trinken').

Zum Abschluss der Analyse der im Songkorpus enthaltenen nominalen Phrasenkomposita stellt sich die Frage, ob bestimmte Genres und/oder Künstler in besonderem Maße Gebrauch von diesem markierten Wortbildungsmuster machen. Insgesamt zeigen die 51 Phrasenkomposita eine breite Streuung über die Archive des Songkorpus, auch wenn nicht alle Archive

bzw. Künstler in der Liste vertreten sind. Auffallend ist, dass sich die meisten dieser Bildungen, und zwar mit einem Anteil von ca. 43 %, in HipHop-Songs finden. Dass sich ein markiertes Wortbildungsmuster wie die Phrasenkomposition insbesondere in Verbindung mit Sprechgesang manifestiert, in dem ein besonderer Fokus auf Sprache gerichtet wird, ist sicherlich wenig überraschend.

Im HipHop-Archiv ist der Anteil von Phrasenkomposita nicht nur am höchsten, sondern hier finden sich auch am ehesten Phrasenkomposita-Muster mit lexikalischem Anker. Solche musterhaften Verwendungen desselben Zweitglied-Lexems sind im Songkorpus insgesamt aber sehr selten, wenn man bedenkt, dass die 51 nominalen Phrasenkomposita nur drei Zweitglied-Lexeme aufweisen, die mehrfach vorkommen.[10] Zudem sind die einzelnen lexikalischen Muster selbst mit drei ([XP-*Style*]*;* [XP-*Pöter*]) bzw. vier Instantiierungen ([XP-*Rap*]) nicht besonders produktiv. Es kann daher festgehalten werden, dass sich die Phrasenkomposita im Songkorpus durch einen hohen Grad an Kreativität auszeichnen. Aufgrund der Tatsache, dass sich zweitgliedzentrierte lexikalische Muster am ehesten im HipHop-Archiv finden, könnte man außerdem darüber mutmaßen, ob insbesondere die Ausdrucksweise im HipHop relativ stark vorgegebenen Konventionen folgt und bestimmte Formulierungsmuster hier eine besondere Rolle spielen.

Neben der Phrasenkomposition finden sich im Songkorpus noch weitere Typen der „phrasalen Wortbildung", nämlich Phrasenkonvertate (s. Übersicht 7) und Phrasenderivate (s. Übersicht 8). Die Ableitungs- bzw. Konversionsbasis wird hier, anders als in prototypischen Konvertaten und Derivaten (vgl. Kapitel 3.2.1), nicht von einem Lexem, sondern von einer phrasalen Einheit gebildet (vgl. LAWRENZ 2006, 8–10). Beide Phänomene sind mit 8 bzw. 9 Belegen aber deutlich weniger prominent vertreten als die Phrasenkomposition.

32. Um kam er einmal im Jahr nach Hause nach Gronau, in unsere kleine Heimatstadt , erzählte er von den sieben Meeren und von wüsten Puffs und vom **Saufen-bis-der-Arzt-kommt**. (Udo Lindenberg)
33. Ein **Beinahe-ums-Leben-kommen-in-Regenpfützen** (ah). (Prezident/HipHop)

Übersicht 7: Phrasenkonvertate im Songkorpus.

Die Bildungen in (32) und (33) illustrieren exemplarisch den im Songkorpus beobachtbaren Rückgriff auf Phrasenkonvertate. In *Saufen-bis-der-Arzt-kommt* wird eine feste Wortverbindung nominalisiert. Diese Nominalisierung wird im Volltextbeleg durch „vom" angezeigt, d.h. durch die Verschmelzung aus Artikel und Präposition. *Beinahe-ums-Leben-kommen-in-Regenpfützen* mutet auf den ersten Blick fast philosophisch oder poetisch an. Dieser Eindruck wird durch die ansonsten eher vulgäre Wortwahl im Song von „Prezident" aber eher zunichte gemacht.[11] Die Nominalisierung einer Phrase, aus der dieses Wortbildungsprodukt hervorgegangen ist, wird durch den unbestimmten Artikel „ein" am Beginn des Volltextbelegs in (33) evident.

Übersicht 8 zeigt eine Auswahl von Phrasenderivaten im Songkorpus. Während Phrasenkonvertate immer nominal sind, sind im Songkorpus sowohl nominale (34) als auch adjektivische Phrasenderivate (35) vertreten.

34. Ihr **Unschuld'ge-zu-lebenslanger-Haft-im-Zoo-Einsperrer**, Ihr Gänsestopfer, Ihr **Schlachtvieh-auf-Tiertransporter-Zerrer**! (Mey/Misc)
35. So leg ich vorsorglich fest, was eines Tags in meiner steht , Dass mein letztes Inserat nicht auch noch in die Hose geht , […] Ich will nicht noch ' nen Verriss, ich will keine Lubhudelei'n, Nicht , dass noch Mike Krüger **Candle-in-the-Wind-mäßig** zum Schluss " Mein Gott Walter " für den traurigen Anlass umdichten muss! (Mey/Misc)

Übersicht 8: Phrasenderivate im Songkorpus.

Die Songtext-Zeile in (34) enthält mit *Unschuld'ge-zu-lebenslanger-Haft-im-Zoo-Einsperrer* und *Schlachtvieh-auf-Tiertransporter-Zerrer* gleich zwei nominale Phrasenderivate, die sich zudem aufgrund ihres Zweitglieds reimen. In beiden Fällen handelt es sich um Ableitungen mit dem Suffix *-er*, welches hier zur Bildung von Nomina Agentis herangezogen wird. Auffallend ist, dass die Verbalgruppen, die als Ableitungsbasis fungieren, in beiden Bildungen jeweils eine negative Tiermetaphorik enthalten. Dementsprechend werden die Gesamtkomplexe in (34) dazu verwendet, eine Kritik zum Ausdruck zu bringen und die Adressaten jeweils negativ zu charakterisieren. Die Spezifik der phrasalen Wortbildung erlaubt es dabei, die Kritik wesentlich expliziter zu formulieren, als dies bei prototypischen Derivaten der Fall wäre.

Candle-in-the-Wind-mäßig ist als relationales Adjektiv zu klassifizieren, das aus der Ableitung einer Nominalphrase mit dem Suffix *-mäßig* hervorgegangen ist. Während *-mäßig* von Fleischer/Barz (vgl. 2012, 346f.) zu den regulären Suffixen gezählt wird, behandelt Lawrenz (2006, 9) analoge Bildungen wie *kleineleutemäßig* als Typ, bei dem „Halbsuffixe […] an Nominalphrasen gehängt [werden], um Adjektive abzuleiten." Zudem handelt es sich bei dem Phrasenderivat in (35) um ein komplexes Wort, dessen Bedeutung sich nur mit Weltwissen erschließen lässt. Auf der Bedeutungsseite wird hier auf den legendären Auftritt Elton Johns mit dem Song „Candle in the wind" im Rahmen von Lady Dianas Beerdigung angespielt. Die Bildung *Candle-in-the-wind-mäßig* ist nur verständlich, wenn dieser Bezug hergestellt wird.

4 Fazit

Die Analyse der ‚longwords' in Kapitel 3 hat gezeigt, dass das Songkorpus eine wahre „Fundgrube" (SCHNEIDER 2022, 40) für kreative Wortbildungsprodukte darstellt. Dies ist keineswegs überraschend, wenn man bedenkt, dass „Lyrics […] eigenwillig [sind] und […] vom Üblichen ab[weichen]" (SCHNEIDER 2022, 143). Das In-den-Blick-Nehmen ‚unüblicher', d.h. okkasioneller Wortbildungen wiederum ist in zweierlei Hinsicht gewinnbringend: Erstens sind „gerade okkasionelle Wortbildungen insofern von besonderem Interesse für die synchrone Wortbildungslehre, als ihre Bildungsweise und Frequenz Auskunft über die Produktivität der Modelle geben können" (FLEISCHER/BARZ 2012, 25). Zweitens wird der Blick auf

Wortbildungsphänomene durch die Hinzuziehung eines „komplementär faszinierenden Forschungsgegenstand[s]“ (SCHNEIDER 2022, 143) wie dem Songkorpus erweitert. Songtexte als Fundgrube für Okkasionalismen stellen eine wichtige Ergänzung zur Beschreibung von Wortbildungsphänomenen auf der Basis von klassischen Textsorten wie Zeitungstexten dar.

In diesem Beitrag wurde versucht, okkasionelle Wortbildungsprodukte über das Kriterium der Wortlänge herauszufiltern – ein Kriterium, das gegenüber alternativen Herangehensweisen wie der Fokussierung von Neologismen, Hapax legomena oder Bindestrich-Schreibungen zumindest objektiv und auch empirisch gut operationalisierbar ist. Auch wenn nicht klar ist, wie viele okkasionelle Wortbildungsprodukte durch die Fokussierung auf die längsten Wörter verpasst wurden (‚Recall‘), so ist die erreichte ‚Precision‘ insofern sehr hoch, als es sich bei fast allen ‚longwords‘ zugleich auch um kreative Wortbildungsprodukte handelt.

Einer der interessantesten Befunde der empirischen Untersuchung ist der hohe Anteil (ca. 28 %) von Phrasenkomposita wie *Ich-finde-du-bist-weak-Rap* gegenüber prototypischen Determinativkomposita (ca. 56 %) aus lexematischen Konstituenten wie *Lederjacken-Prollschiene* – eine Verteilung, die von der in klassischen Textsorten wie Zeitungstexten deutlich abweicht (vgl. Kapitel 3.2) (vgl. ORTNER U. A. 1991, 4:6). Eine weitere deutliche Abweichung von anderen schriftsprachlichen Textsorten ist der extrem niedrige Anteil (ca. 2 %) von Derivaten. Dies könnte als Ergebnis der spezifischen Ausdrucksbedürfnisse von Songtexten interpretiert werden, allerdings ist das Ergebnis insofern mit Vorsicht zu genießen, als das extrem geringe Vorkommen von Derivaten auch dem 24-Buchstaben-Kriterium geschuldet sein könnte. Möglicherweise sind viele Derivate nicht ins Untersuchungskorpus eingegangen, weil sie aus weniger als 24 Buchstaben bestehen.

Der gegenüber der Standardsprache deutlich erhöhte Anteil phrasaler Wortbildungsprodukte zeigt, dass Kreativität bzw. Aufmerksamkeit in Songtexten offenbar (auch) durch den Rückgriff auf markierte Wortbildungsmuster erzeugt wird. Dies ist jedoch nicht das einzige zur Verfügung stehende Mittel: Als ‚auffällig‘ ist in jedem Fall auch die Lexik zu etikettieren, die bei der Realisierung der Wortbildungsmuster verwendet wird. Das Spektrum reicht dabei von vulgär (z.B. *Bremsstreifen-Boxershorts*) und brutal (z.B. *Rapdeutschlandkettensägenmassaker*) über umgangssprachlich (z.B. *Chauvi-Macho-Macker-Stuss*) bis poetisch (z.B. *Kindheitserinnerungswälder*). Dies führt dazu, dass auch völlig unmarkierte Wortbildungstypen wie die Determinativkomposition oder die Derivation in Songtexten expressive Effekte erzeugen.

Als Fazit kann daher festgehalten werden, dass sich kreative Wortbildungsprodukte in Songtexten einerseits durch den Rückgriff auf auffällige, von der Norm abweichende Wortbildungstypen auszeichnen, andererseits aber auch durch die Wahl auffälliger Lexik innerhalb der Wortbildungsmuster charakterisiert sind. Es ist also ein Zusammenspiel struktureller und lexikalischer Faktoren, das im Songkorpus für eine erhöhte Kreativität sorgt.

Songtexte sind künstlerische Akte des Sprachgebrauchs. Ihr linguistischer Mehrwert, der im vorliegenden Beitrag für den Bereich der Wortbildung illustriert wurde, zeigt einmal mehr,

wie wichtig es ist, Wortbildung nicht nur strukturell, sondern auch im konkreten Sprachgebrauch – und dies innerhalb unterschiedlicher medialer Realisierungen – „im Sinne einer pragmatischen Wortbildung“ (STUMPF 2021, 76) zu untersuchen und zu beschreiben.

[1] ‘Extravaganz’ wird zum Teil auch im Sinne von ‘Expressivität’ verstanden bzw. damit gleichgesetzt. HASPELMATH (1999, 1057) argumentiert jedoch dafür, dass ‚extravagant‘ die treffendere der beiden Bezeichnungen sei: “According to a dictionary definition, expressive means ‘showing very clearly what someone thinks or feels’, so in this sense ‘‘expressivity’’ would not be different from clarity […] and it would not explain why speakers should use an innovated word for a sense that for a long time has successfully been expressed by different means.”

[2] Dies erfolgt jedoch weder systematisch noch durch explizite empirische Kontraststudien, sondern ist aus Machbarkeitsgründen von exkurshaftem Charakter.

[3] Ich danke Roman Schneider für die automatische Extraktion der längsten Wörter des Songkorpus.

[4] Auch Überlegungen, auf der Basis einer Liste aller Wörter mit Bindestrichschreibungen zu einer möglichst umfangreichen Liste kreativer Wortbildungen zu gelangen, wurden letztlich verworfen. Aus anderen empirischen Untersuchungen (vgl. z.B. HEIN 2015) ist bereits bekannt, dass die Bindestrichschreibung kein stabiles Merkmal markierter Wortbildungsmuster darstellt.

[5] Für jeden Beleg wird die Quelle jeweils in der Form ‚(KÜNSTLER/ARCHIV)‘ angegeben. In Fällen, in denen Künstler und Archiv deckungsgleich sind, enthält die Quellenangabe folglich nur ein Element (z.B. Udo Lindenberg).

[6] Natürlich enthalten auch die Ad-hoc-Bildungen teils lexikalisierte Konstituenten, als Ganzes gesehen handelt es sich aber um Okkasionalismen.

[7] Zum Konzept der ‚kommunikativen Minimaleinheit‘ vgl. Zifonun et al. (1997, 86).

[8] Der Verfestigungsgrad des Erstglieds wurde jedoch nicht in einer mit Hein (2015) vergleichbaren systematischen Weise untersucht und annotiert.

[9] Drei der vier Belege mit dem Zweitglied Style stammen jedoch aus ein und demselben Song (Azad: „unaufhaltbar“): *Totengräber-Sparten-Style*; *Keine-Begrenzung-des-Schaden-Style*; *Schützengräben-explodier-Granaten-Style*.

[10] Es gibt nur ein Erstglied, das mehr als einmal im untersuchten Songkorpus-Ausschnitt vorkommt, nämlich [*Menschenrechts*-X] (2 Vorkommen). Zudem liegt nur ein Gesamtkomplex vor (*Minderwertigkeitskomplex*), zu dem es mehrere Tokens gibt.

[11] *Beinahe-ums-Leben-kommen-in-Regenpfützen* ist im Song von ‚Prezident‘ wie folgt eingebettet: „Mit mir selbst, immer so 'n Gefühl von fünf vor zwölf. Von noch nicht, schon zu spät, jetzt erst recht, fick' die Welt, warum nicht, hin und weg, Prezi-spring-ins-Trümmerfeld. Permanent den Arsch auf Grundeis, aber trägt das Himmelszelt Als Kopfbedeckung (yeah), wer nicht 'n Bisschen übertreibt. Kommt auf keinen grünen Zweig, ich weiß bescheid. Ist ein Leben auf den Zehenspitzen (yeah). Ein Rodeoritt mit schweren Geschützen (yeah). Ein Beinahe-ums-Leben-kommen-in-Regenpfützen (ah) (https://genius.com/Prezident-bis-unterskinn-lyrics).

Literatur

Donalies, E. (2007). Basiswissen Deutsche Wortbildung. Tübingen; Basel: Francke.

Eitelmann, M./ Haumann, D. (eds.) (2022). Extravagant morphology. Amsterdam/Philadelphia: John Benjamins Pub.

Erben, J. (2006). Einführung in die deutsche Wortbildungslehre. 5., durchges. und erg. Aufl. Berlin: Schmidt.

Fleischer, W./ Barz, I. (2012). Wortbildung der deutschen Gegenwartssprache. 4. Auflage, völlig neu bearb. von Irmhild Barz; unter Mitarb. von Marianne Schröder. Berlin; Boston: de Gruyter.

grammis (2022). Grammatisches Informationssystem „grammis". Mannheim: Institut für Deutsche Sprache. http://grammis.ids-mannheim.de. doi: 10.14618/grammis.

Günther, C./ Kotowski, S./ Plag, I. (2018). „Phrasal compounds can have adjectival heads: Evidence from English". *English Language & Linguistics* 24(1), 75-95. doi:10.1017/S1360674318000229.

Haspelmath, M. (1999). „Why is grammaticalization irreversible?" *Linguistics. An interdisciplinary journal of the language sciences* 37(6), 1043-1068.

Hein, Katrin. (2015). Phrasenkomposita im Deutschen: empirische Untersuchung und konstruktionsgrammatische Modellierung. Tübingen: Narr.

Hein, Katrin/ Antonioli, G. (in Vorb.). „Phrasenkomposita im gesprochenen Deutsch an der Schnittstelle von Wortbildungs- und Gesprächsforschung". In: Murelli, A./ Gaeta, L. (eds.) (2024). Sammelband zur Tagung „Das heutige gesprochene Deutsch zwischen Sprachkontakt und Sprachwandel" (Universität Turin, Sept. 2022). Reihe Germanistische Linguistik. Berlin: De Gruyter.

Helmer, H. (2022). „Okkasionalismen im gesprochenen Deutsch. Bedeutungserklärungen zwischen Notwendigkeit und interaktiver Ressource". *Deutsche Sprache* 2/2022, 97-123.

Leibniz-Institut für Deutsche Sprache (2011.). Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2011-I (Release vom 29.03.2011). http://www.ids-mannheim.de/DeReKo.

Leibniz-Institut für Deutsche Sprache (2022). Datenbank für gesprochenes Deutsch (DGD). Version 2.18 (25.7.2022). https://dgd.ids-mannheim.de.

Lawrenz, B. (2006). Moderne deutsche Wortbildung. Phrasale Wortbildung im Deutschen: linguistische Untersuchung und sprachdidaktische Behandlung. Hamburg: Dr. Kovač.

Ortner, H./ Ortner, L. (1984). Zur Theorie und Praxis der Kompositaforschung. Tübingen: Narr.

Ortner, L./ Müller-Bollhagen, E./ Ortner, H./ Wellmann, H./ Pümpel-Mader,M./ Gärtner, H. (1991). Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache. Band 4: Substantivkomposita. Düsseldorf: Schwann.

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

Steyer, K./ Hein, K. (2018). „Satzwertige usuelle Wortverbindungen und gebrauchsbasierte Muster". In: Engelberg,S./ Lobin, H./ Steyer, K./ Wolfer, S. (eds.) (2018). Wortschätze: Dynamik, Muster, Komplexität. Jahrbuch / Institut für Deutsche Sprache 2017. Berlin: De Gruyter.

Stumpf, S. (2021). „Passe-partout-Komposita im gesprochenen Deutsch. Konstruktionsgrammatische und interaktionslinguistische Zugänge im Rahmen einer pragmatischen Wortbildung“. *Zeitschrift für Germanistische Linguistik* 49 (1): 33–83.

Wöllstein, A./ Dudenredaktion (eds.) (2016). Duden - Die Grammatik. Unentbehrlich für richtiges Deutsch. 9., Vollständig überarbeitete und aktualisierte Auflage. Der Duden in zwölf Bänden. Band 4. Mannheim: Dudenverlag.

Zifonun, G./ Hoffmann, L./ Strecker, B. (1997). Grammatik der deutschen Sprache 1. Berlin: de Gruyter.

Zifonun, G. (2021). Das Deutsche als europäische Sprache. Ein Portrait. Berlin; Boston: De Gruyter.

Korrespondenzanschrift

Katrin Hein
Leibniz-Institut für Deutsche Sprache
hein@ids-mannheim.de

Elke Donalies

Phraseme im Songkorpus: Etabliertes in Anti-Establishment-Texten

1 Definition Phrasem

Phraseme (griech. *phrasis* 'rednerischer Ausdruck') sind polylexikal, das heißt, sie bestehen aus mindestens zwei Wörtern, zum Beispiel *Gespenster sehen* oder *gähnende Leere*[1]. Phraseme sind Wiederholungen; sie sind als Ganzes etabliert. Phraseme können idiomatisch sein, demotiviert, metaphorisch, zum Beispiel *mit dem Feuer spielen*, müssen das aber nicht, zum Beispiel *zwischen Hoffen und Bangen* oder *kreuz und quer*. Phraseme versprachlichen genauso wie ein Wort einen einzigen Begriff, eine Einheit des Denkens, eine Idee, ein Konzept, zum Beispiel *von der Hand in den Mund leben*.

2 Zugrundeliegende Datensammlung

Für diesen Beitrag habe ich je 19 zufällige Seiten Songtexte von Element of Crime, Fettes Brot, Udo Lindenberg, Stefan Stoppok, Konstantin Wecker und Marius Müller-Westernhagen ausgewertet. Pro Seite waren das jeweils rund 650 Wörter, insgesamt also rund 74.100 Wörter. Daraus habe ich 140 Phraseme extrahiert (siehe Abschnitt 10 (Anhang)). Gesammelt habe ich so lange, bis sich die verschiedensten Aspekte (Struktur, Stil, Inhalte etc.) mehrfach wiederholten; insofern ist mein – ja eher kleines – Teilkorpus durchaus gesättigt. Überdurchschnittlich phrasemreich sind im untersuchten Datensample mit 41 Phrasemen die Texte von Stoppok (29%). Für weitere statistische Feststellungen solcherart ist mein Teilkorpus zu klein.

Die Autoren habe ich unter dem Aspekt *Themenvergleichbarkeit* ausgewählt. Die Themen sind: Alkohol, Drogen, Leben, Lieben, Scheitern. Bei Konstantin Wecker kommt explizit Politisches hinzu.

3 Zentraler Aspekt dieses Beitrags

Spätestens seit der Verleihung des Literaturnobelpreises an Bob Dylan 2016 können meines Erachtens Songtexte getrost als Literatur gelten. Sie haben Lyrikstatus. Leser zeitgenössischer Lyrik erwarten nach Engelberg & Rapp (2018, S. 31) „ungewöhnliche Wortkombinationen", also gerade nicht die etablierten Kombinationen der Phraseme. Außerdem singen alle ausgewählten Autoren gegen das Establishment an, gegen vorgefertigtes Leben, 08/15 und Denkschablonen; das könnte eine Antihaltung gegen sprachlich

[1] Die in diesem Beitrag angeführten Phraseme sind – so nicht anders markiert – authentische Belege aus dem hier ausgewerteten Teilkorpus des Songkorpus (Schneider, 2022 (Stand April 2022)).

Etabliertes, gegen Phraseme inkludieren. Phraseme konservieren – und konservativ sind die untersuchten Autorentexte ganz sicher nicht.

Wie genau sieht also der Phrasemgebrauch in meinem Teilkorpus des Songkorpus aus?

4 Strukturen

Die Nutzung bestimmter Phrasemstrukturen entspricht der allgemeinen Nutzung. Die Autoren brauchen wie alle Autoren vor allem verbale Kombinationen und adjektivbegleitete Substantive. Sie greifen häufig aber auch auf ganze Sätze zurück.

Am meisten genutzt werden mehr oder weniger komplexe Verbphraseme mit Präposition (Beleg 1 und 2) oder ohne Präposition (Beleg 3).

(1) **unter Strom stehen**
Ich liebe hohe Spannung und stehe meistens unter Strom.
Lindenberg, Bis ans Ende der Welt

(2) **sich zum Narren machen**
Habe ich die Wahl. Soll ich mich zum Narren machen. Wenn man mich bezahlt.
Westernhagen, Ich will es wissen

(3) **jemandem den Marsch blasen**
Er kriecht der Wirtschaft in den Arsch und bläst dem Rest der Welt den Marsch.
Wecker, Amerika 2001

Häufig sind es sogenannte Somatismen. Das sind anthropozentrische Phraseme, die physische oder psychische, meist negative Befindlichkeiten (Belege 4-6) oder Mimisches und Gestisches versprachlichen (Beleg 7). Siehe auch Mieder (2020).

(4) **die Nerven verlieren**
und weißt du noch, wie du geweint hast und wie ich die Nerven verlor?
Element of Crime, Die Party am Schlesischen Tor

(5) **an die Nieren gehen**
Es wurmt dich und es geht dir an die Nieren, überall geht's nur ums Profitieren.
Stoppok, Mir stinkt's auch

(6) **in die Knie gehen**
Schlimmer als Silvester am Brandenburger Tor eingeklemmt und blau zwischen hunderttausend Fremden zu steh'n und noch vor Mitternacht in die Knie zu geh'n ist das Gefühl, das ich habe, wenn ich dich wiederseh und du tust, als ob nichts gewesen wär.
Element of Crime, Karin Karin

(7) **ohne mit der Wimper zu zucken**
Hast du den Trottel gesehen, der eben, ohne auch nur mit einer Wimper zu zucken, verkehrtherum in die Einbahnstraße fuhr?
Element of Crime, Bevor ich dich traf

Typische Verbphraseme sind daneben Funktionsverbgefüge. Funktionsverbgefüge werden definiert als „die Verbindung von einem verbalen Bestandteil (einem Funktionsverb, das arm ist an lexikalischer Bedeutung) und einem nominalen Bestandteil (zumeist einer Nominalgrupe im Akkusativ oder einer Präpositionalgruppe), abgeleitet von korrespondierenden Verben“(Helbig, 2006, S. 166). Solche in anderen Korpora üblichen, immer ein bisschen gestelzt wirkenden und deshalb allenthalben sprachkritisch belächelten oder sogar bekämpften Gefüge, etwa *in Erfahrung bringen, in Anspruch nehmen, zu Ende gehen, in Erscheinung treten, in Abrede stellen, Dank sagen, einen Entschluss fassen*, habe ich in meinem Teilkorpus nicht gefunden. Das hat wohl vor allem stilistische Gründe. Funktionsverbgefüge passen nicht zum Stil der untersuchten Songtexte. Siehe Abschnitt 5 (Stilistisches).

Bei den Substantivphrasemen dominieren solche aus Substantiv und Adjektivattribut (Beleg 8 und 9). Bei den selteneren mit zwei Substantiven (Beleg 10 und 11) dominieren Paarformeln, die durch eine Präposition verortet werden und rhythmisch oder alliterierend gleichklingen.

(8) **heiße Spur**
Der Detektiv, der niemals schlief, rund um die Uhr auf heißer Spur.
Lindenberg, Auf heißer Spur

(9) **gähnende Leere**
denn in meinem Kühlschrank herrscht wieder einmal gähnende Leere.
Fettes Brot, Das Lied vom Ende

(10) **zwischen Hoffen und Bangen**
Du hast mir Treue geschworen. Zwischen Hoffen und Bangen. Wurde ich geboren.
Westernhagen, Clown

(11) **in Nacht und Nebel**
Hab Dich gefunden in Nacht und Nebel. Hast mich gefesselt, hast mich geknebelt.
Westernhagen, Hey Honey

Die dritte größere Gruppe bilden Satz-, überwiegend Einfachsatzphraseme. Es sind alltagsnahe Sprichwörter unbekannter Autorschaft, die Untersuchungsgegenstände der Parömiologie (Beleg 12), oder Zitate berühmter Autoren (Beleg 13), auch Geflügelte Worte genannt (griech. *épea pteróenta*). Beleg 13 stammt aus Goethes Faust I. Teil (1808).

(12) **Wie man in den Wald hineinruft, so schallt es heraus**
Wie man's in den Wald hineinruft, so schallt's heraus. Wir stellen uns mit

Flüstertüten vor's Alsterhaus.
Fettes Brot, Bundeskanzler

(13) **Da steh ich nun, ich armer Tor.**
Da steh ich nun, ich armer Tor, und würd mich gern an Weisheitslehren berauschen.
Wecker, Dass alles so vergänglich ist

Wie allgemein sind auch bei den Autoren meines Teilkorpus Adverbphraseme rar. Es sind immer rhythmisch oder alliterierend gleichklingende Paarformeln.

(14) **hin und her**

(15) **kreuz und quer**
Die ganze Kohle überweist er - hin und her - kreuz und quer und ganz viel Geld in die Dritte, Vierte oder Fünfte Welt.
Lindenberg, Commander Superfinger

(16) **landauf, landab**
hörts ihr uns alle mitanander ned schrein, landauf, landab, alle, die noch am Lebn sind.
Wecker, Der Baum

Etwas häufiger wiederum sind satzgrammatisch unterschiedlich strukturierte Phraseme mit der Vergleichspartikel *wie*. Beleg 17 ist ein Beispiel für die ebenfalls raren Adjektivphraseme.

(17) **so dumm wie Brot**
Bist du schlau oder bist du so dumm wie Brot?
Element of Crime, Die Party am Schlesischen Tor

(18) **wie aus dem Ei gepellt**
Wieder alles renoviert hier. Wie aus dem Ei gepellt. Kann man sich nicht vorstelln, dass das einem nicht gefällt.
Stoppok, Frisch renoviert

(19) **bluten wie ein Schwein**
Jetzt ist der Unsinn bald vorbei, das war auch allerhöchste Zeit. Schau her, ich blute wie ein Schwein. Ist mir egal, du tust mir leid.
Element of Crime, Immer nur geliebt

5 Stilistisches

Der Sprachstil der sechs Autoren ist dezidiert leger bis derb alltagssprachlich. Daher gibt es zahlreiche Belege für derbe Phraseme, die sich in den meisten anderen Sprachkorpora mutmaßlich weniger dicht auffinden lassen.

(20) **auf die Kacke hauen**
Ausgerechnet jetzt lieg' ich im Krankenhaus und da komm ich wohl lebend nicht mehr raus. Wo ich gerade jetzt auf die Kacke hauen wollte und im nächsten Monat meine Rente krieg'n sollte.
Stoppok, Ausgerechnet jetzt

(21) **die Arschkarte ziehen**
ja auch das große Aktienglück ist längst verflogen und nun hat er die goldene Arschkarte gezogen.
Lindenberg, Der Millionär hat keine Kohle mehr

(22) **Arsch auf Grundeis gehen**
Es gibt Tage, da geht mir der Arsch auf Grundeis, da könnt ich heulen, obwohl ich keinen Grund weiß, da scheint keine Sonne, da geht kein Licht an.
Stoppok, Gute-Laune-Blues

(23) **bluten wie ein Schwein**
Jetzt ist der Unsinn bald vorbei, das war auch allerhöchste Zeit. Schau her, ich blute wie ein Schwein. Ist mir egal, du tust mir leid.
Element of Crime, Immer nur geliebt

(24) **dummes Huhn**
Ob ich neidisch bin? Ha! Ich doch nicht, dummes Huhn.
Element of Crime, Geh doch hin

(25) **blöde Kuh**
Am liebsten habe ich meine Ruh. Das gilt auch für dich, du blöde Kuh.
Westernhagen, Herr D.

(26) **armes Schwein**
Sie lassen dich draußen stehen, so böse und gemein, und wieder bist du das arme Schwein.
Lindenberg, Club der Millionäre

Die Schnauze voll haben sogar gleich drei Autoren. Genaugenommen spricht aus den Texten aller sechs Autoren eine solche Grundhaltung.

(27) **die Schnauze voll haben**
Ausgerechnet jetzt, wo's richtig losgehn soll, ausgerechnet jetzt hast du die Schnauze voll und haust mir einfach ab, total unverschämt!
Stoppok, Ausgerechnet jetzt
Er hatte die Schnauze von diesem Leben voll, er wär so gern ausgeflippt.
Lindenberg, Der Malocher
Ruf mich an, wenn du die Schnauze voll hast. Komm vorbei, wenn du nicht weißt, wohin.
Element of Crime, Dieselben Sterne

Gelegentlich werden aber auch alltagssprachferne, elitär bildungssprachliche Phraseme herangezogen. Sie stammen aus dem gemeinsamen Bildungshintergrund der gebildeten Autoren und ihrer gebildeten Zuhörer. Phraseme dieser Art bewahren und transportieren kollektives Wissen, kollektive Kultur.

(28) **Doch alle Lust will Ewigkeit.** (Nietzsche, Zarathustra)
das muß doch jetzt die Liebe sein. Und feuchte Haut und plötzlich Mut. Und alle Lust will Ewigkeit.
Wecker, Bleib nicht liegen

(29) **Die Milch der frommen Denkungsart** (Schiller, Wilhelm Tell)
Draußen hinterm Fenster sitzt ein Kind und rührt in der Milch der frommen Denkungsart herum.
Element of Crime, Draußen hinterm Fenster

6 Varianten

Im Prinzip sind Phraseme variabel. So müssen Verb-, Substantiv- und Adjektivphraseme der Morphosyntax angepasst werden.

(30) **um den heißen Brei reden**
Ich rede nicht gern um den heißen Brei: Ich wollte euch nie erziehen.
Wecker, An meine Kinder

In Satzphraseme werden nach Bedarf andere Satzelemente eingeschoben.

(31) **Der Lack ist ab.**
Der Lack ist bei uns beiden zwar schon ab, doch alten Resten eine Chance, mal sehen, ob es noch klappt.
Element of Crime, Alten Resten eine Chance

Daneben gibt es Variation etwa als „Austausch von Wörtern“(Ortner, 1982, S. 283). Ortner nennt aus ihrem Pop- und Rockkorpus *die ewigen Rockgründe* zu *die ewigen Jagdgründe* und *der große Blonde mit den flinken Fingern* zum Filmtitel *Der große Blonde mit dem schwarzen Schuh* (1972).

In ungereimter Sprache ist das nicht notwendig; dagegen geschieht der Austausch von Wörtern in meinem durchgehend gereimten Teilkorpus naturgemäß häufig des Reimes wegen. In den Belegen 32 und 33 reimt sich *befallen/ knallen, schlagen/ sagen.* Der Austausch der Wörter bleibt dabei semantisch folgenlos.

(32) **auf die Kacke hauen**
Das Pferd blieb zu Haus vom Selbstmitleid befallen, der Bär ging immer aus, um richtig auf die Kacke zu knallen.
Stoppok, Gelbes Pferd, grüner Bär

(33) **Regentropfen, die an mein Fenster klopfen, die sagen dir ...**
Regentropfen, die an mein Fenster schlagen, werden mir immer wieder nur das eine sagen: Du kommst nie mehr zurück.
Stoppok, Nie mehr zurück

Auch der reimbedingte Vertausch von Wörtern – in Beleg 34 ist es Schweiß/ heiß – hat keinen semantischen Einfluss auf das Phrasem.

(34) **Blut, Schweiß und Tränen**
Er wühlt in Sex und Tränen, Blut und Schweiß und fährt jetzt Cabrio. Sein Preis ist heiß.
Wecker, Der dumme Bub 3

Dagegen hat in Beleg 35 der Wortaustausch wesentliche Auswirkungen auf die Phrasembedeutung. Die variierten Wörter endreimen sich auf die ursprünglichen und schaffen so eine besondere Nähe zum eigentlichen Phrasem. Es handelt sich um eine systematische Anspielungsvariante, für die typisch die „Übernahme der syntaktischen Struktur“ist (Lange, 1998, S. 189). Ausgesagt wird, dass *Dichter und Denker* aufs gegenwärtige Vaterland nicht zutrifft, sondern *Richter und Lenker.*

(35) **Land der Dichter und Denker**
Ach, du mein schauriges Vaterland, du Land der Richter und Lenker!
Wecker, Ach, du mein schauriges Vaterland

Auch Beleg 36 stellt durch Austausch eines Wortes und durch Vertausch der Bezugssubstantive die Aussage des Phrasems elementar infrage.

(36) **Reden ist Silber, Schweigen ist Gold**
Schweigen ist feige. Reden ist Gold.
Westernhagen, Schweigen ist feige

Tatsächlich werden also in meinem Teilkorpus mit hohem Anteil an Anti-Establishment-Texten die etablierten Aussagen der Phraseme mitunter durch knallharte Gegenentwürfe torpediert. Die Autoren entziehen damit den Phrasemen, die sonst argumentativ als unwiderlegbare Wahrheiten funktionieren, ihre Unwiderlegbarkeit.

Außer durch Aus- und Vertausch von Wörtern werden Phraseme durch Zusätze variiert. Die Zusätze nuancieren semantisch und stellen mitunter das Phrasem in verblüffende Kontexte.

(37) **alles auf eine Karte setzen**
Alte Männer setzen alles auf die letzte große Creditkarte.
Lindenberg, Der Greis ist heiß

Schließlich wird ein Phrasem durch Weglassen einzelner Wörter variiert. Das Weglassen ist notwendig, damit eine zusätzliche Aussage Platz hat. In Beleg 38 sind es die ölgleichen Tränen, die wütend machen; in Beleg 39 ist es die Dummheit der Allzu-Lauten.

(38) **Öl ins Feuer gießen**
Deine Tränen sind noch einmal richtig Öl im Feuer meiner Wut.
Element of Crime, Finger weg von meiner Paranoia

(39) **Hunde, die bellen, beißen nicht.**
Die Hunde, die so laut bellen, das sind nicht unbedingt die ganz hellen.
Stoppok, Ich sach ma so

7 Verbindung zweier Phraseme

Einerseits stehen zwei Phraseme in einem Satz zusammen, ohne dass dadurch ein spezieller Effekt erzielt wird. Ihr Zusammentreffen ist eher zufällig.

(40)(41) **heiliger Strohsack + starker Tobak**
Heiliger Strohsack, das ist für manchen starker Tobak.
Fettes Brot, Da draußen

(42)(43) **jemandem einen Strich durch die Rechnung machen + auf dem Schlauch stehen**
Zieh den Strich nicht durch die Rechnung, steh nicht dauernd auf dem Schlauch.
Stoppok, Jackpot

Andererseits werden Phraseme offenbar bewusst miteinander kontaminiert. Ihre Aussagen beziehen sich aufeinander, bedingen einander. So wird in Beleg 44 + 45 zum lesbaren Buch, wer wie gedruckt lügt.

(44)(45) **lügen wie gedruckt + in jemandem wie in einem Buch lesen**
Ich lüg wie gedruckt, du liest in mir wie im Buch, sagst „Dank dir für den Besuch!“, du hast von mir jetzt echt genug.
Fettes Brot, Bring mich nach Haus

(46)(47) **über Leichen gehen + vor die Hunde gehen**
Das ist halt das Schöne am deutschen Verbund: wir gehen über Leichen und die Andern vor die Hund.
Wecker, Clevermänner, Eastlandrunner

(48)(49) **Die Wände haben Ohren + Mauern des Schweigens**
Haben die Wände hier Ohren oder sind das Mauern des Schweigens?
Fettes Brot, Echo

8 Witzige Verwendung

Bewusst eingesetzt werden metaphorische Phraseme, die wörtlich weitergedacht werden. Das Phrasem ist dabei präsent; der Witz entsteht aus der Überraschung, aus dem Unerwarteten der (Be)Deutung.

(50) **am Arsch sein**
Hier ist einer, der dich braucht. Wenn du am Arsch bist, stinkt's mir auch.
Stoppok, Mir stinkt's auch

(51) **die Hand für etwas ins Feuer legen**
Ich leg meine Hand in das Feuer vom Würstchengrill unten am Fluss dafür, dass nicht alles umsonst war und jeder nur tut, was er muss.
Element of Crime, Kaffee und Karin

Wörtlich nimmt auch Fettes Brot das Phrasem *Wie man in den Wald hineinruft, so schallt es heraus.* Ausprobiert wird, ob das Phrasem tatsächlich zutrifft – offenbar nicht.

(52) **Wie man in den Wald hineinruft, so schallt es heraus.**
Ich steh jeden Morgen früh auf und ruf laut in den Wald hinein. Irgendwann hat mir mal jemand erzählt, so schallts auch wieder raus. Kam lange keiner mehr zu Besuch. Ich hab genug vom Alleinesein und irgendwie ist alles so irgendwo zwischen oh yeah und okay. [...] *Ohne Echo, Echo, kein Echo. Ich brauch ein Echo, Echo, ein Echo. Gib mir ein Echo, Echo, ein Echo!*
Fettes Brot, Echo

Manchmal spielen die Autoren mit dem Doppelsinn eines Phrasems. Das Phrasem *in die Röhre kucken* ist deutlich vor Erfindung des TVs nachgewiesen und hat folglich etymologisch nichts mit dem TV zu tun – vermutlich kommt es aus der Jägersprache, wo enge Dachsbauten *Röhre* heißen – wird aber wahrscheinlich gar nicht so selten mit dem TV neumotiviert. Der Malocher in Beleg 53 jedenfalls kuckt TV und dabei in die Röhre, geht also im Leben leer aus. In Beleg 54 sind die erotisch aufgeladenen feuchten Träume zu verregneten Träumen umgedeutet.

(53) **in die Röhre kucken**
Und dann schmiss er's mit Karacho voll ins TV und schrie: Ihr glaubt wohl ich bin nicht ganz dicht! Jeden Abend Fusel schlucken und dann in die Röhre kucken und dann pennen und dann wieder zur Schicht.
Lindenberg, Der Malocher

(54) **feuchte Träume**
Es regnet, begossen wird die Welt. Wer jetzt nicht schläft, verfällt der feuchten Träumerei.
Element of Crime, Es regnet

Niederdeutsch *Na, denn man tau* 'Na, dann mal zu! Auf geht's! Tschüss!' bezieht Stoppok doppelsinnig auf seinen kaputten Kühlschrank.

55) **Na, denn man tau**
Jetzt nach all den Jahren taut er auf, ich kann's noch immer nicht verstehen. [...] *Mein Freund der Kühlschrank, ja dann man tau.*
Stoppok, Der Kühlschrank

Gerne werden Phraseme, die etablierten Verbindungen, von den unetablierten Autoren karikiert, lächerlich gemacht, jedenfalls hinterfragt. Beleg 59 spielt mit den Analogien von *nicht ganz bei Trost sein* 'verrückt sein' und *untröstlich* 'ohne Trost'. Wobei einer, der untröstlich ist und Spaß dabei hat, wirklich nicht ganz bei Trost sein kann.

56) **Was kostet die Welt?**
Freunde, was kostet die Welt? Eins fünfzig! Mmh, bezahl ich, ist ja wahrlich günstig.
Fettes Brot, Dionysos

(57) **den Hals nicht vollkriegen können**
Viele Flaschen können den Hals nicht vollkriegen. Hier kommt endlich der Korken.
Fettes Brot, Da draußen

(58) **Die Gedanken sind frei.**
Denn die Gedanken sind frei, keiner kann sie googlen.
Fettes Brot, Crazy World

(59) **bei Trost sein**
Frag' mich nicht, ob ich noch bei Trost bin. Den ganzen Tag untröstlich und Spaß dabei.
Element of Crime, Im Himmel ist kein Platz mehr für uns zwei

9 Resümee

Das Songkorpus erlaubt Einblicke in bestimmte gesellschaftliche Diskurse, die in anderen Sprachkorpora weniger zur Geltung kommen. Das zeigt sich auch bei der Analyse von Phrasemen im Songkorpus.

Phraseme sind etablierte Wortkombinationen; sie konservieren kollektives Wissen, kollektive Kultur. Element of Crime, Fettes Brot, Udo Lindenberg, Stefan Stoppok, Konstantin Wecker, Marius Müller-Westernhagen, die Autoren meines kleinen Teilkorpus, sind Anti-Establishment und alles andere als konservativ. Zwar verwenden sie häufig Phraseme verschiedenster Struktur und Art, karikieren sie aber auch häufig, spielen lässig mit ihnen, hinterfragen ihre Bedeutung, verändern ihre Bedeutung. Ihre spezielle Haltung bedingt spezielle Phraseme und spezielle Phrasemvarianten.

References

Engelberg, S., & Rapp, I. (2018). Die Gräten einer Harfe. Metaphorische Transformation und ihre morphosyntaktische Grundlage. In E. Winter-Froemel (Ed.), *Sprach-Spiel-Kunde. ein Dialog zwischen Wissenschaft und Praxis* (p. 31-34). Berlin etc.: de Gruyter.

Helbig, G. (2006). Funktionsverbgefüge – Kollokationen – Phraseologismen. Anmerkungen zu ihrer Abgrenzung im Lichte der gegenwärtigen Forschung. In U. Breuer

& I. Hyvärinen (Eds.), *Wörter – Verbindungen. Festschrift für Jarmo Korhonen zum 60. Geburtstag* (p. 165-174). Frankfurt etc.: Peter Lang.

Lange, M. (1998). Die Verwendung sprachlicher Vorlagen in Texten der Anzeigenwerbung. In D. Hartmann (Ed.), *"Das geht auf keine Kuhhaut" – Arbeitsfelder der Phraseologie. Akten des westfälischen Arbeitskreises Phraseologie/Parömiologie 1996. (= Studien zur Phraseologie und Parömiologie)* (p. 169-198). Bochum:Brockmeyer.

Mieder, W. (Ed.). (2020). *"mit dem Kopf durch die Wand": Sprichwörtliche Somatismen in der modernen Lyrik.* Burlington/ Vermont: The University of Vermont.

Ortner, L. (1982). Wortschatz der Pop-/Rockmusik. In *Sprache der Gegenwart (53).* Düsseldorf:Schwann.

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

Korrespondenzanschrift

Elke Donalies
Leibniz-Institut für Deutsche Sprache
donalies@ids-mannheim.de

10 Anhang

Ach wie gut, dass keiner weiß, dass ich Rumpelstielzchen heiß. (Märchen der Brüder Grimm) **Variante** *Ach, wie gut, dass keiner weiß: Der Greis ist: Häh? Der Greis ist heiß.* Lindenberg, Der Greis ist heiß
Alle Wege führen nach Rom. Variante *Viele Wege führen nach Rom, doch nur einer führt zu dir. Du sagst ich kenn' ihn schon, bitte zeig ihn mir. Was soll ich denn in Rom?* Fettes Brot, 6 Million Ways To Rome Choose One
alles auf eine Karte setzen Variante *Alte Männer setzen alles auf die letzte grosse Creditkarte.* Lindenberg, Der Greis ist heiß
alter Schwede *Knack den Jackpot, alter Schwede, und dann hol dir das Geld ab.* Stoppok, Jackpot
am Arsch sein *Hier ist einer, der dich braucht. Wenn du am Arsch bist, stinkt's mir auch.* Stoppok, Mir stinkt's auch
am Ende des Regenbogens (irische Sage/ Filmtitel) *Am Ende des Regenbogens legt der Regen noch einen Zahn zu und überflutet im Überschwang, gleich hinter dem S-Bahn-Übergang, den Weg mit leeren Flaschen, Steinen und Schlamm.* Element of Crime, Am ersten Sontag nach dem Weltuntergang
an die Nieren gehen *Es wurmt dich und es geht dir an die Nieren, überall geht's nur ums Profitieren.* Stoppok, Mir stinkt's auch
armes Schwein *Sie lassen dich draußen stehen, so böse und gemein und wieder bist du das arme Schwein.* Lindenberg, Club der Millionäre
Arsch auf Grundeis gehen *Es gibt Tage, da geht mir der Arsch auf Grundeis, da könnt ich heulen obwohl ich keinen Grund weiß, da scheint keine Sonne, da geht kein Licht an.* Stoppok, Gute-Laune-Blues
auf dem Schlauch stehen *Zieh den Strich nicht durch die Rechnung, steh nicht dauernd auf dem Schlauch.* Stoppok, Jackpot **auf den Knien rutschen** *Rutsch mir den Buckel runter, aber rutsch nicht auf deinen Knien. Jetzt hast Du ja alles gebeichtet und ich hab dir alles verziehn.* Stoppok, Gesellschaftsspiele
auf den Senkel gehen *Was mir dann total auf den Senkel geht, für mich das allerallerletzte: Aufgesetzte gute Laune, aufgesetzte. Kalter Kaffee, harte Brötchen, alte Witze, unbequeme Stühle* Stoppok, Gute-Laune-Blues
auf die Kacke hauen *Ausgerechnet jetzt lieg' ich im Krankenhaus und da komm ich wohl lebend nicht mehr raus. Wo ich gerade jetzt auf die Kacke hauen wollte und im nächsten Monat meine Rente krieg'n sollte.* Stoppok, Ausgerechnet jetzt **Variante** *Das Pferd blieb zu Haus vom Selbstmitleid befallen, der Bär ging immer aus, um richtig auf die Kacke zu knallen.* Stoppok, Gelbes Pferd, grüner Bär
auf die Schnauze fallen Variante *Wenn man, so wie ich, schon mal richtig auf die Fresse gefallen ist, weil ein Mädchen gesagt hat: "Alles klar", und die Sache dann doch ganz anders gelaufen ist, hat man wenig Mut, es noch mal zu riskieren.* Lindenberg, Bitte keine Love-Story
auf Zack sein *Weil der Papst is 'n Hammer, ist enorm auf Zack. Er erkennt 50 Länder an ihrem Geschmack, oh yeah. Weil er doch ständig den Boden abküsst.* Lindenberg,

Benedictum-Benedactum
auferstanden aus Ruinen und der Zukunft zugewandt (DDR-Nationalhymne) **Variante** *Auferstanden aus Ruinen und den Reimen zugewandt* Fettes Brot, Friedhof der Nuscheltiere
aufgefahren in den Himmel (Apostolisches Glaubensbekenntnis) *Keine Namen. Kein Versprechen. Keine Lügen. Wahre Liebe. Aufgefahren. In den Himmel. Für Momente. Sind wir eins.* Westernhagen, Dreh dich nicht um
bei Trost sein *Frag' mich nicht, ob ich noch bei Trost bin. Den ganzen Tag untröstlich und Spaß dabei.* Element of Crime, Im Himmel ist kein Platz mehr für uns zwei
blanker Neid *Dreh meine Joints nur vom Feinsten. Da spricht bei Dir nur blanker Neid.* Fettes Brot, Da draußen
blauer Planet *Auf dem blauen Planet ist ja wohl alles zu spät.* Lindenberg, Der blaue Planet
blöde Kuh *Am liebsten habe ich meine Ruh. Das gilt auch für dich, du blöde Kuh.* Westernhagen, Herr D.
Blut, Schweiß und Tränen (Churchill 1940) **Variante** *Er wühlt in Sex und Tränen, Blut und Schweiß und fährt jetzt Cabrio. Sein Preis ist heiß.*Wecker, Der dumme Bub 3
bluten wie ein Schwein *Jetzt ist der Unsinn bald vorbei, das war auch allerhöchste Zeit. Schau her, ich blute wie ein Schwein. Ist mir egal, du tust mir leid.* Element of Crime, Immer nur geliebt
Da steh ich nun, ich armer Tor. (Goethe, Faust) *Da steh ich nun, ich armer Tor, und würd mich gern an Weisheitslehren berauschen.* Wecker, Dass alles so vergänglich ist
den Hals nicht vollkriegen können *Viele Flaschen können den Hals nicht vollkriegen. Hier kommt endlich der Korken.* Fettes Brot, Da draußen **der große Wurf** *Der große Wurf ist meist ein Bumerang. Knapp vorbei ist auch daneben. Warum geht so viel im Leben schief?* Stoppok, Flügel
Der Haussegen hängt schief *Früher war alles anders, ich hatte mein geregelten Stress. Konnt' mich stundenlang drüber auslassen wie schlecht es mir ging, und wie schief, wie schief der Haussegen hing.* Stoppok, Lotto gewonn'n
Der Lack ist ab. *Der Lack ist bei uns beiden zwar schon ab, doch alten Resten eine Chance, mal sehen, ob es noch klappt.* Element of Crime, Alten Resten eine Chance
die Arschkarte ziehen *ja auch das große Aktienglück ist längst verflogen und nun hat er die goldene Arschkarte gezogen.* Lindenberg, Der Millionär hat keine Kohle mehr
Die Fahne hoch. (Horst-Wessel-Lied 1929) *Ein Ölfeld brennt. Es ist schon dunkel. Die Fahne hoch. Und die Hosen müssen runter.* Westernhagen, Neger
Die Gedanken sind frei. *Denn die Gedanken sind frei, keiner kann sie googlen.* Fettes Brot, Crazy World
die Hand für etwas ins Feuer legen *Ich leg meine Hand in das Feuer vom Würstchengrill unten am Fluss dafür, dass nicht alles umsonst war und jeder nur tut, was er muss.* Element of Crime, Kaffee und Karin
die Kirche im Dorf lassen **Variante** (Leo Kirch, Medienunternehmer) *Man hätt' den Kirch im Dorfe lassen sollen.* Wecker, Der dumme Bub 3

die Köpfe rollen lassen *Damals ließ er die Köpfe rollen, ja er war schon immer überaus tüchtig.* Stoppok, Ehrenmann
die Mauern des Schweigens *Haben die Wände hier Ohren oder sind das Mauern des Schweigens?* Fettes Brot, Echo
die Milch der frommen Denkungsart (Schiller, Wilhelm Tell 1804) *Draußen hinterm Fenster sitzt ein Kind und rührt in der Milch der frommen Denkungsart herum.* Element of Crime, Draußen hinterm Fenster
die Nerven verlieren *und weißt du noch, wie du geweint hast und wie ich die Nerven verlor?* Element of Crime, Die Party am Schlesischen Tor
die Schnauze voll haben *Ausgerechnet jetzt, wo's richtig losgehn soll, ausgerechnet jetzt hast du die Schnauze voll und haust mir einfach ab, total unverschämt!* Stoppok, Ausgerechnet jetzt *Er hatte die Schnauze von diesem Leben voll, er wär so gern ausgeflippt.* Lindenberg, Der Malocher *Ruf mich an, wenn du die Schnauze voll hast. Komm vorbei, wenn du nicht weißt, wohin.* Element of Crime, Dieselben Sterne
Die Wände haben Ohren. *Haben die Wände hier Ohren oder sind das Mauern des Schweigens?* Fettes Brot, Echo
Doch alle Lust will Ewigkeit (Nietzsche, Zarathustra 1885) *das muß doch jetzt die Liebe sein. Und feuchte Haut und plötzlich Mut. Und alle Lust will Ewigkeit.* Wecker, Bleib nicht liegen
Dritte Welt **Variante** *Die ganze Kohle überweist er - hin und her - kreuz und quer und ganz viel Geld in die Dritte, Vierte oder Fünfte Welt.* Lindenberg, Commander Superfinger
Dukaten kacken *Ich krieg's nicht mehr geregelt mit den laufenden Kosten, wie wär es mit einem Ölscheich aus dem mittleren Osten? Auch wenn ich ein Esel bin, ich kack keine Dukaten.* Stoppok, Du brauchst Personal
dumm wie Brot *Bist du dabei oder bist du so gut wie tot? Bist du schlau oder bist du so dumm wie Brot?* Element of Crime, Die Party am Schlesischen Tor
dummes Huhn *Ob ich neidisch bin? Ha! Ich doch nicht, dummes Huhn.* Element of Crime, Geh doch hin
ein Mann wie ein Baum *Ich war ein Cowboy, ein Mann wie ein Baum.* Lindenberg, Cowboy
ein Schiff mit acht Segeln (Brecht, Seeräuber-Jenny 1928) *Und kein Schiff mit acht Segeln lag drunten am Kai. Für diese Herren war die Party vorbei.* Wecker, Der Virus
eine Leiche im Keller haben *Leichen im Keller, Beton im Gemüt und viel zu lang schon allein.* Element of Crime, Ganz leicht
einem nackten Mann in die Tasche greifen *Mein lieber Herr Gerichtsvollzieher, wie wär's denn mit nem Gläschen Bier? Ich hab noch eine Flasche Pils, die teil'n wir uns, wenn Du willst. Wir sind ja quasi schon Bekannte, wenn auch nicht gerade Blutsverwandte.* [...] *Aber auch Du müsstest doch wissen: Dass man einem nackten Mann nicht in die Tasche greifen kann.* Stoppok, Der nackte Mann
einen Knick in der Optik haben *Ein Knick in der Optik, ein Kratzen im Hals und viel zu ängstlich mit dir.* Element of Crime, Ganz leicht

einen Strich durch die Rechnung machen Variante *Zieh den Strich nicht durch die Rechnung, steh nicht dauernd auf dem Schlauch.* Stoppok, Jackpot
Erlöse uns von dem Übel! (Apostolisches Glaubensbekenntnis) *Oh Django, erlöse uns von dem Übel und füll Bier in unsere Kübel.* Stoppok, Django
feuchte Träume *Es regnet, begossen wird die Welt. Wer jetzt nicht schläft, verfällt der feuchten Träumerei.* Element of Crime, Es regnet
Flink wie Windhunde, zäh wie Leder und hart wie Kruppstahl (Hitler 1935) *Ich lebe. Rock and roll steht wieder mal. Wir sind wieder hart wie Stahl. Die Familie ist gesund. Was soll's.* Westernhagen, Lass uns leben
Froh zu sein bedarf es wenig und wer froh ist ist ein König. Variante *Fett zu sein bedarf es wenig und wer fett ist, hört den König.* Fettes Brot, Definition von fett
gähnende Leere *denn in meinem Kühlschrank herrscht wieder einmal gähnende Leere.* Fettes Brot, Das Lied vom Ende
Gespenster sehen *Dabei hat es gar nix mit dem Streit zu tun und mit dem kaputten Fenster. Ich hab das Gefühl, du siehst allmählich überall Gespenster.* Stoppok, Alles so schwarz
Getretner Quark wird breit, nicht stark. (Goethe, Westöstlicher Divan 1819) *Ich fürchte zu wissen, warum du anrufst und mir erzählst, dass es im Kern für diese Jahreszeit zu kalt ist. Und dass du dieses Jahr so gern mal wirklich richtig in den Süden oder auch nach Dänemark gefahren wärst, wenn man dich ließe. Getretener Quark wird breit, nicht stark.* Element of Crime, Getretener Quark
goldener Käfig *Sitz in nem goldenen Käfig. Von allem, was ich brauch oder nicht, hab ich mehr als genug, denn mach ich mal piep, krieg ich genau was ich will.* Stoppok, Goldener Käfig
graue Maus *Alle Mauerblümchen, alle grauen Mäuschen, alle Streberleichen, alle Pickelfressen, alle hässlichen Entlein.* Fettes Brot, Falsche Entscheidung
grün und blau schlagen Variante *Ich hab jetzt Sachen an, die du nicht magst und die sind immer grün und blau. Ob ich wirklich Sport betreibe, interessiert hier keine Sau.* Element of Crime, Delmenhorst
hässliches Entlein *Alle Mauerblümchen, alle grauen Mäuschen, alle Streberleichen, alle Pickelfressen, alle hässlichen Entlein.* Fettes Brot, Falsche Entscheidung
heiliger Strohsack *Heiliger Strohsack, das ist für manchen starker Tobak.* Fettes Brot, Da draußen
heiße Spur *Der Detektiv, der niemals schlief, rund um die Uhr auf heißer Spur.* Lindenberg, Auf heißer Spur
hin und her *Die ganze Kohle überweist er - hin und her - kreuz und quer und ganz viel Geld in die Dritte, Vierte oder Fünfte Welt.* Lindenberg, Commander Superfinger
Hinter den Kulissen von Paris ist das Leben noch einmal so süß. (Mireille Mathieu 1969) *Ich hau' jetzt ab nach Paris, da ist das Leben so süß.* Lindenberg, Der Malocher
hinter den sieben Bergen bei den sieben Zwergen *Ich geh' über sieben Berge und über sieben Brücken und hüpf' noch kurz durch's Minenfeld und dann bin ich auch schon da - in der jungen Welt.* Lindenberg, Der Generalsekretär

hoch hinaus wollen *Doch ich wollte hoch hinaus und dann bin ich da weg. Und so zog ich nach Hollywood, Mel Brooks war mein Regisseur.* Lindenberg, Cowboy
hoch und heilig versprechen *Du versprachst mir hoch und auch ziemlich heilig, ich wär ab jetzt dein Macker. Und ich hatte noch nicht mal meinen Deckel bezahlt, da warst Du schon vom Acker.* Stoppok, Herzlos
hohle Sprüche *Ich muss zugeben, ich bin immer wieder platt. Du bist die Königen der hohlen Sprüche.* Stoppok, Die Königin
Hunde, die bellen, beißen nicht. **Variante** *Die Hunde, die so laut bellen, das sind nicht unbedingt die ganz hellen.* Stoppok, Ich sach ma so
Ich glaub, es hackt. *Doch jetzt wird's arg, jetzt zieht es an, du erzählst mir alles alles über deinen Mann. Wie und was und wo, ich glaub es hackt.* Stoppok, Die Königin
im Arsch sein *Du bist nicht mehr, wie du warst. Unsre Liebe ist im Arsch.* Fettes Brot, Du driftest nach rechts
im Bilde sein *Dass die Welt verrückt ist, das weiß ich genau. Das seh' ich jeden Abend in der Tagesschau. Ich bin im Bilde, mir kann man nichts erzähl'n. Ich hab so ziemlich alles schon im Fernsehen geseh'n.* Stoppok, Alles nur'n Film
Im Dunkeln ist gut munkeln. **Variante** *Liebling, lass uns tanzen. Denn tanzen darf ein jeder Jud. Neger, die sind dunkel. Im Dunkeln läßt sich's munkeln.* Westernhagen, Mit Pfefferminz bin ich dein Prinz
Im Westen nichts Neues (Buchtitel Remarque 1928) *Im Westen nichts Neues. Hannawald ist gestürzt. Die Börse im Keller. Deine Sorgen, deine Sorgen.* Westernhagen, Es ist an der Zeit
in die Gänge kommen *Dann kann's passier'n, dass ich erst nachmittags so richtig in die Gänge komme.* Stoppok, Alles so schwarz **in die Knie gehen** *Schlimmer als Silvester am Brandenburger Tor eingeklemmt und blau zwischen hunderttausend Fremden zu steh'n und noch vor Mitternacht in die Knie zu geh'n ist das Gefühl, das ich habe, wenn ich dich wiederseh und du tust, als ob nichts gewesen wär.* Element of Crime, Karin Karin
in die Röhre kucken *Und dann schmiss er's mit Karacho voll ins TV und schrie: Ihr glaubt wohl ich bin nicht ganz dicht! Jeden Abend Fusel schlucken und dann in die Röhre kucken und dann pennen und dann wieder zur Schicht.* Lindenberg, Der Malocher
in jemandem wie in einem Buch lesen *Ich lüg wie gedruckt, du liest in mir, wie im Buch, sagst "Dank dir für den Besuch!", du hast von mir jetzt echt genug.* Fettes Brot, Bring mich nach Haus
in Nacht und Nebel *Hab Dich gefunden in Nacht und Nebel. Hast mich gefesselt, hast mich geknebelt.* Westernhagen, Hey Honey
ins Schleudern kommen *Wenn ich mal ins Schleudern komme, bist du da und hältst mich fest.* Lindenberg, Baby, wenn ich down bin
Jeder soll nach seiner Façon glücklich werden. (Friedrich der Große 1740) *Räucherstäbchen und Wildreis und Abende auf dem Balkon. In Eppendorf ist morgen Flohmarkt und jeder nach seiner Façon.* Element of Crime, Ein Hotdog unten am Hafen

jemandem den Buckel runterrutschen *Rutsch mir den Buckel runter, aber rutsch nicht auf deinen Knien. Jetzt hast Du ja alles gebeichtet und ich hab dir alles verziehn.* Stoppok, Gesellschaftsspiele
jemandem den Marsch blasen *Er kriecht der Wirtschaft in den Arsch und bläst dem Rest der Welt den Marsch.* Wecker, Amerika 2001
jemanden auf dem Kieker haben *Alles so Typen bei denen ich nicht grade beliebt war. Ganz besonders der eine hatte mich auf dem Kieker.* Fettes Brot, Die meisten meiner Feinde
jemanden hängen lassen *Du bist die eine, die mich niemals, die mich niemals hängen läßt.* Lindenberg, Baby, wenn ich down bin
jemandem in den Arsch kriechen *Er kriecht der Wirtschaft in den Arsch und bläst dem Rest der Welt den Marsch.* Wecker, Amerika 2001
jemanden sticht der Hafer *Wissen Sie, Gnädigste, ganz so blöd bin ich nicht. Um nicht zu merken, wenn Sie wieder mal der Hafer sticht.* Stoppok, Krank, Madame
jemanden wie den letzten Dreck behandeln *Sie hat mich behandelt wie den letzten Dreck. Mir blieb nichts anderes übrig, ich musste weg.* Stoppok, Nachtzug
Keine Macht für niemand. (Ton Steine Scherben 1972) *Keine Macht für niemand. Klingt der Scherben Ton. Keine Macht, keine Macht.* Westernhagen, Keine Macht
keinen Finger krumm machen *Die machen keinen Finger krumm, die verdienen jeden Tag tausendmal mehr als ich.* Stoppok, Hart sein
kleines Licht *Mein Vater ist irgend so'n kleines Licht bei 'ner Bank. Meine Mutter putzt Treppen.* Westernhagen, Dass du mich verlässt
kreuz und quer *Die ganze Kohle überweist er - hin und her - kreuz und quer und ganz viel Geld in die Dritte, Vierte oder Fünfte Welt.* Lindenberg, Commander Superfinger
Land der Dichter und Denker (19. Jh., Ursprung unbekannt) **Variante** *Ach, du mein schauriges Vaterland, du Land der Richter und Lenker!* Wecker, Ach, du mein schauriges Vaterland
landauf, landab *hörts ihr uns alle mitanander ned schrein, landauf, landab, alle, die noch am Lebn sind.* Wecker, Der Baum
Liebe ist stärker als der Tod. (Altes Testament) **Variante** *Sag ihr, hier sei alles im Lot und je länger man kaut, desto süßer das Brot. Irgendwas ist immer, irgendwas ist immer und Liebe ist kälter als der Tod.* Element of Crime, Liebe ist kälter als der Tod
lügen wie gedruckt *Ich lüg wie gedruckt, du liest in mir wie im Buch, sagst „Dank dir für den Besuch!", du hast von mir jetzt echt genug.* Fettes Brot, Bring mich nach Haus
Mann, ist der Dickmann. (Werbespruch für Dickmann 1985) *Bist du fett, bist du ein Blickfang. Mann ist der dick, Mann.* Fettes Brot, Defintion von fett
Meine Ruh ist hin, mein Herz ist schwer, ich finde sie nimmer und nimmermehr. (Goethe, Faust 1808) *Die größte Liebe meines Lebens kann ich nicht vergessen, ich kann tun, was ich will, doch ich finde keine Ruh'.* Lindenberg, Die größte Liebe
Mia san mia. *Mir san die Freistaatbuam, und mir san mir.* Wecker, Bayernpower

mit dem Feuer spielen *Der Mensch spielt zu gern mit dem Feuer. Warum, das ist ihm selber schleierhaft. Menschen sind nicht zu beneiden mit ihrer Neugier, ihrer Leidenschaft.* Stoppok, Flügel
mit seinem Latein am Ende sein *Es gibt nun mal Momente, wo man am Ende mit seinem Latein ist.* Stoppok, Gesellschaftsspiele
Mutter Natur *Schön, wenn man liebt, was Mutter Natur einem gibt.* Element of Crime, Ein Hotdog unten am Hafen
Na, denn man tau. **Variante** *etzt nach all den Jahren taut er auf, ich kann's noch immer nicht verstehen.* [...] *Mein Freund der Kühlschrank, ja dann man tau.* Stoppok, Der Kühlschrank
nicht ganz dicht sein *Ihr glaubt wohl, ich bin nicht ganz dicht!* Lindenberg, Der Malocher
nichts Gutes verheißen *Wie er mit schrägem Blick, der schon nichts Gutes verhieß, den Bierstand in Richtung WC verließ, unterwegs mit 'nem Mülleimer zusammenstieß und sich die Augenbraue piercte mit nem Schaschlikspieß.* Stoppok, Cool durch Zufall
ohne mit der Wimper zu zucken *Hast du den Trottel gesehen, der eben, ohne auch nur mit einer Wimper zu zucken, verkehrtherum in die Einbahnstraße fuhr?* Element of Crime, Bevor ich dich traf
ohne Punkt und Komma reden **Variante** *Jetzt wird Machmut aus Aleppo vor die Kamera gestellt. Er redet mit Punkt und Komma von dem Bombenhagel und dem Tod seiner Omma.* Stoppok, Mein Herz hat damit nix zu tun
Öl ins Feuer gießen *Deine Tränen sind noch einmal richtig Öl im Feuer meiner Wut.* Element of Crime, Finger weg von meiner Paranoia
Reden ist Silber, Schweigen ist Gold **Variante** *Schweigen ist feige. Reden ist Gold.* Westernhagen, Schweigen ist feige
Regentropfen, die an mein Fenster klopfen, die sagen dir ... (Tangoschlager 1939) **Variante** *Regentropfen, die an mein Fenster schlagen, werden mir immer wieder nur das eine sagen: Du kommst nie mehr zurück.* Stoppok, Nie mehr zurück
rien ne va plus *Shit ... jetzt läuft ja hier wohl gar nichts mehr - rien ne va plus.* Lindenberg, Commander Superfinger
schwarz fahren *Getrampt oder mit'm Moped oder schwarz mit der Bahn, immer bin ich dir irgendwie hinterher gefahr'n.* Lindenberg, Cello
schwarz sehen *Fragt die Polizei mal nach 'nem Alibi, du siehst immer alles so schwarz. Bleib' ich mal weg bis 6 Uhr in der Früh, du machst immer alles gleich schlecht.* Stoppok, Alles so schwarz
schwer wie Blei *Wenn die Menschen Flügel hätten, flögen sie aus ihren Städten fort, hoch in die Wolken und ohne in Wort. Aber Menschen haben Beine, schwer wie Blei und meistens keine Zeit.* Stoppok, Flügel
seine sieben Sachen packen *Im Gepäck nicht mehr als sieben Sachen, auf dem Kompass nichts als geradeaus.* Element of Crime, Immer unter Strom
sentimentaler Hund *Und was für'n sentimentaler Hund, du weintest heimlich im Kino.* Lindenberg, Brief an den Jungen, der ich vor 30 Jahren war

sich die Hörner abstoßen *Als er mit seinem ersten Mädchen im Arm und mit Heiratsplänen nach Hause kam, nahm sein Vater ihn dezent zur Seite und sagte: Stoß dir erst mal die Hörner ab, Junge.* Lindenberg, Der sizilianische Wolf
sich einen abfrieren *Hey Maria, mach die Tür auf, ich frier mir hier draußen einen ab! Jetzt lass mal wieder gut sein, die Zeit für uns wird knapp.* Stoppok, Hey Maria
sich zum Narren machen *Habe ich die Wahl. Soll ich mich zum Narren machen. Wenn man mich bezahlt.* Westernhagen, Ich will es wissen
starker Tobak *Heiliger Strohsack, das ist für manchen starker Tobak.* Fettes Brot, Da draußen; Stoppok, Gute-Laune-Blues
Straße der Verdammten (Filmtitel 1955) *Ich mach' jetzt endlich alles öffentlich und erzähle, was ich weiß. Auf der Straße der Verdammten, die hier Bremer Straße heißt.* Element of Crime, Delmenhorst
trübe Tassen *Profilclowns auf allen Kanälen, die mich genau in dem Moment so grausam quälen, die lustigen Luschen, die trüben Tassen, worüber die lachen, das ist nicht zu fassen.* Stoppok, Gute-Laune-Blues
über Leichen gehen *Baby, gehn ma zua. Das ist halt das Schöne am deutschen Verbund: wir gehen über Leichen und die Andern vor die Hund.* Wecker, Clevermänner, Eastlandrunner
Über sieben Brücken musst du gehen. (Karat 1978) *Ich geh' über sieben Berge und über sieben Brücken und hüpf' noch kurz durch's Minenfeld und dann bin ich auch schon da - in der jungen Welt.* Lindenberg, Der Generalsekretär
um den heißen Brei reden *Ich rede nicht gern um den heißen Brei: Ich wollte euch nie erziehen.* Wecker, An meine Kinder
unter Strom stehen *Ich liebe hohe Spannung und stehe meistens unter Strom.* Lindenberg, Bis ans Ende der Welt **Viel Feinde, viel Ehr.** (Georg von Frundsberg 1473-1528) *Und du auf dem fliegenden Pferde. Rufst stolz: Viele Feinde, viel' Ehr.* Westernhagen, Alleine
von der Hand in den Mund leben *Sie lebten nicht schlecht von der Hand in den Mund, in Frieden und Freuden und mit einem Hund.* Stoppok, Adam und Eva
vor die Hunde gehen *Baby, gehn ma zua. Das ist halt das Schöne am deutschen Verbund: wir gehen über Leichen und die Andern vor die Hund.*Wecker, Clevermänner, Eastlandrunner
Warum ist die Banane krumm? Ja, wenn die Banane grade wär', dann wär's keine Banane mehr *Die Banane ist krumm, das ist doch klar, denn wenn die Banane grade wär', dann wär' sie keine Banane mehr.* Lindenberg, Bananenrepublik
Was kostet die Welt? *Freunde, was kostet die Welt? Eins fünfzig! Mmh, bezahl ich, ist ja wahrlich günstig.* Fettes Brot, Dionysos
weg vom Fenster sein *Die Rock'n'Roll-Gespenster sind weg vom Fenster, die Arie ist angesagt.* Lindenberg, Elli Pyrelli
wenn der Herrgott es will *Sie genossen ihr Charlie Chaplin Idyll, sollten Kinder kommen, wenn der Herrgott es will.* Stoppok, Adam und Eva
wie aus dem Ei gepellt *Wieder alles renoviert hier. Wie aus dem Ei gepellt. Kann man sich nicht vorstelln, dass das einem nicht gefällt.* Stoppok, Frisch renoviert

Wie man in den Wald hineinruft, so schallt es heraus. *Wie man's in den Wald hineinruft, so schallt's heraus. Wir stellen uns mit Flüstertüten vor's Alsterhaus.* Fettes Brot, Bundeskanzler *Ich steh jeden Morgen früh auf und ruf laut in den Wald hinein. Irgendwann hat mir mal jemand erzählt, so schallts auch wieder raus.* Fettes Brot, Echo
wilde Ehe *ihre Hoffnung war unendlich, ihre Ehe war wild.* Stoppok, Adam und Eva
Wollt ihr den totalen Sieg? (Goebbels 1943) *"Deutschland vor, noch ein Tor, und wieder zeigen wir's der ganzen Welt!" Die Nation ist im Rausch, alle wollen den totalen Sieg.* Lindenberg, Bei uns in Spanien
zwischen den Stühlen sitzen **Variante** *So hängt er mit den Gefühlen zwischen den Stühlen.* Lindenberg, Ali
zwischen Hoffen und Bangen *Du hast mir Treue geschworen. Zwischen Hoffen und Bangen. Wurde ich geboren.* Westernhagen, Clown

Sarah Broll / Roman Schneider

Empirische Verortung konzeptioneller Nähe/Mündlichkeit inner- und außerhalb schriftsprachlicher Korpora

Abstract

Linguistische Studien arbeiten häufig mit einer Differenzierung zwischen gesprochener und geschriebener Sprache bzw. zwischen Kommunikation der Nähe und Distanz. Die Annahme eines Kontinuums zwischen diesen Polen bietet sich für eine Verortung unterschiedlichster Äußerungsformen an, inklusive unkonventioneller Textsorten wie etwa Popsongs. Wir konzipieren, implementieren und evaluieren ein automatisiertes Verfahren, das mithilfe unkorrelierter Entscheidungsbäume entsprechende Vorhersagen auf Textebene durchführt. Für die Identifizierung der Pole definieren wir einen Merkmalskatalog aus Sprachphänomenen, die als Markierer für Nähe/Mündlichkeit bzw. Distanz/Schriftlichkeit diskutiert werden, und wenden diesen auf prototypische Nähe-/Mündlichkeitstexte sowie prototypische Distanz-/Schrifttexte an. Basierend auf der sehr guten Klassifikationsgüte verorten wir anschließend eine Reihe weiterer Textsorten mithilfe der trainierten Klassifikatoren. Dabei erscheinen Popsongs als „mittige Textsorte", die linguistisch motivierte Merkmale unterschiedlicher Kontinuumsstufen vereint. Weiterhin weisen wir nach, dass unsere Modelle mündlich kommunizierte, aber vorab oder nachträglich verschriftlichte Äußerungen wie Reden oder Interviews vollkommen anders verorten als prototypische Gesprächsdaten und decken Klassifikationsunterschiede für Social-Media-Varianten auf. Ziel ist dabei nicht eine systematisch-verbindliche Einordung im Kontinuum, sondern eine empirische Annäherung an die Frage, welche maschinell vergleichsweise einfach bestimmbaren Merkmale („shallow features") nachweisbar Einfluss auf die Verortung haben.

Keywords: Mündlichkeit, Schriftlichkeit, Nähe, Distanz, Textsorten, Empirik, Features, Machine Learning

1 Motivation

Natürlicher Sprache begegnen wir üblicherweise in mündlich gesprochener (phonischer) oder geschriebener (graphischer) Form. Letztere nutzt analoge respektive digitale Kommunikationsmedien, erlaubt eine Trennung von Produktions- und Rezeptionsphase und mithin mannigfaltige Formen der Überarbeitung (Fehlerkorrektur, Stilprüfung etc.). Gesprochener Diskurs dagegen findet überwiegend spontan und ungeplant statt. Damit liegt nahe, „dass das gesprochene und das geschriebene Deutsch unterschiedliche Präferenzen bei der Wahl der Mittel aus dem Systeminventar des Deutschen haben"

(Eichinger, 2017, S. 292). Äußerungen beider Modalitäten unterscheiden sich hinsichtlich Repertoire und Verwendung von syntaktischen Konstruktionen oder Vokabular, stilistischen Merkmalen – etwa Neubildungen oder Anakoluthen – und einigem mehr. Typisch lautliche Ausdrucksmöglichkeiten und Strategien auf der einen Seite kontrastieren mit einer (zumindest oft angenommenen) größeren Nähe zu (ebenfalls oft auch nur vermeintlichen) Sprachstandards und Konventionen.

Diese mediale Charakterisierung erscheint auf den ersten Blick als unkompliziert, erweist sich aber spätestens dann als unterkomplex, sobald alternative Kommunikationswege ins Spiel kommen: Songtexte beispielsweise lassen sich als medial mündlich kategorisieren, wenn sie live vorgetragen bzw. rezipiert werden, und als medial schriftlich, wenn wir uns auf ihre Darstellung in Liederbüchern o.Ä. beziehen. Gleiches gilt für viele andere kommunikative Formen wie Vorträge, Reden, Interviews usw., die geplant oder auch spontan stattfinden und vorab bzw. anschließend verschriftlicht werden.

Eine Überwindung des medialen Binarismus erlaubt der von Koch und Oesterreicher (1985, 2007) eingeführte und mittlerweile in der Linguistik weithin etablierte Ansatz eines multidimensionalen Kontinuums zwischen den beiden Polen 'Sprache der Nähe' und 'Sprache der Distanz'. Bei der Einordnung konkreter Äußerungen stehen hier nicht die Eigenschaften des Vermittlungsmediums im Vordergrund, sondern Kommunikationsbedingungen wie raumzeitliche Trennung, Reflektiertheit, Vertrautheit der Kommunikationspartner oder Affektivität. Aus diesen ergeben sich dann ggf. Präferenzen für die mediale Realisierung. Koch/Oesterreicher liefern zwar keine umfassende Systematik, welche sprachlichen Merkmale genau mit welchem Wirkungsgrad für eine Verortung im Nähe-Distanz-Kontinuum ausschlaggebend sind, gehen aber auf einige relevante universale und einzelsprachliche Eigenschaften ein, etwa morphosyntaktische Phänomene, lexikalische Vielfalt, Gliederungssignale, Verwendung von Partikeln usw. Eine klare und eindeutige Einordnung bleibt in vielen Fällen schwierig (vgl. z.B. Biber und Conrad (2019)).

Mit der vorliegenden Studie möchten wir keinen Beitrag leisten zu Debatten über Angemessenheit, Unzulänglichkeiten oder Modifikationen des Nähe-Distanz- bzw. Mündlich-Schriftlich-Ansatzes[1] (vgl. hierzu z. B. Feilke und Hennig (2016)), sondern die gewinnbringende Umsetzbarkeit seiner empirischen Algorithmisierung für die automatische Textklassifikation großer Datenmengen demonstrieren, basierend auf der Annahme eines konzeptionellen Kontinuums zwischen Nähesprache (mit einer Präferenz zur Mündlichkeit/Oralität) und Distanzsprache (mit einer Präferenz zur Schriftlichkeit/Literalität).

Digitale Infrastrukturen für Sprachressourcen spiegeln die medialen Pole wider: Die *Datenbank Gesprochenes Deutsch (DGD)* (Schmidt, 2017) als größte Sammlung gesprochener deutscher Sprache enthält ausschließlich Audioinhalte (und deren Transkripte), die mehr oder weniger typische mündliche Kommunikationssituationen repräsentieren. Im Gegensatz dazu umfasst das *Deutsche Referenzkorpus (DeReKo)* (Kupietz, Lüngen, Kamocki & Witt, 2018) medial schriftliche Texte, die dem Anschein nach in verschiede-

[1] In eine ähnliche Richtung zielen Unterscheidungen zwischen informeller und formeller Sprache oder zwischen Alltags-/Gebrauchs- und Bildungssprache; auf mögliche terminologische Überschneidungen und Unschärfen soll hier ebenfalls nicht eingegangen werden.

nen Bereichen der Nähe-Distanz-Kontinuumsskala angesiedelt werden können: Fach- und Publikumspresse, Belletristik, Redetranskripte, Soziale Medien etc. Die genaue Verortung ist keinesfalls trivial und konventionelle Metadaten helfen dabei nicht immer weiter: Zeitungen und Zeitschriften zum Beispiel umfassen zwar zuvorderst Beiträge, die vermutlich intuitiv als Distanzsprache (konzeptionell schriftlich) eingeordnet werden würden; sie enthalten aber ebenfalls Interviews, Gespräche und Diskussionen, die – ungeachtet der üblichen redaktionellen Nachbearbeitung und Fehlerkorrektur – tendenziell vermehrt Eigenschaften konzeptioneller Mündlichkeit aufweisen.

Offen bleibt angesichts der hohen Variationsbreite und Dynamiken bereits innerhalb etablierter Ressourcen die Klassifizierung korpuslinguistischer Spezialsammlungen wie der oben angesprochenen Songtexte: Tendieren diese zur Oralität oder zur Literalität oder sind sie irgendwie hybrid – und falls ja, aufgrund welcher Merkmale und mit welcher Gewichtung? Zuverlässig lassen sich solche Aussagen bestenfalls auf Einzeltextebene treffen, nicht pro Medientextsorte.[2] Für umfangreiche Datensamples existiert nichtsdestotrotz das Desiderat einer soliden automatisierten Verortung anhand empirisch ermittelbarer Kriterien.

Vor diesem Hintergrund verfolgt die vorliegende Studie drei Ziele: (a) Evaluation methodisch fundierter, datengesteuerter Klassifikationen konzeptioneller Mündlichkeit bzw. Schriftlichkeit auf einer belastbaren Textbasis (b) Operationalisierung der quantitativen Verortung auch nicht-eindeutiger Daten im Nähe-Distanz-Kontinuum (c) statistisch valide Identifikation wirkungsmächtiger sprachlicher Einflussfaktoren und Merkmale.

2 Related Work

Eine Reihe moderner Grammatik- bzw. Variationsbeschreibungen skizzieren Merkmale des gesprochenen Deutsch bzw. gehen auf umgangssprachliche Besonderheiten ein, exemplarisch seien Zifonun, Hoffmann und Strecker (1997) und Barbour und Stevenson (1998) genannt. Ágel und Hennig (2006a, 2006b) beziehen sich auf dem Weg zu einer Theorie des Nähe- und Distanzsprechens auf verschiedene von Koch/Oesterreicher angeführte Merkmale und konzipieren darüber hinausgehend eine differenzierte Systematik, die Nähemerkmale aus universalen Parametern der Nähekommunikation ableitet.

Wenige empirische Studien haben diese umfassenden Vorarbeiten bislang praktisch aufgegriffen und korpuslinguistisch evaluiert. Für das Deutsche berechnen Ortmann und Dipper (2019) eine Reihe einfacher sprachlicher Merkmale und nutzen diese zur Ermittlung des Grads an Oralität in Zeitungsartikeln (extrahiert aus Tüba-D/Z- und Tiger-Baumbanken), in Reden und Vorträgen (aus dem Gutenberg-DE-Korpus) sowie in weiteren Samples monologisch und dialogisch angelegter Diskurse (Predigten, Filmuntertitel, Gesprächs- und Chatprotokolle) aus öffentlichen verfügbaren Quellen bzw. Spezialkorpora. Das STTS-getaggte Korpus umfasst insgesamt ca. 2,5 Millionen Token; die 17 für die automatische Ermittlung herangezogenen Merkmale sind typisiert in die Kategorien Komplexität, Bezugnahme/Deixis, Syntax und Satztyp. Ortmann

[2] Und selbst dies gestaltet sich angesichts potenziell verschiedenartiger Textpassagen nicht durchgehend unproblematisch; vgl. (Dürscheid, 2016, S. 55).

und Dipper (2020) wenden den statistischen Ansatz mit Erfolg auch auf historische Texte an. Die kombinierten Erkenntnisse fließen in das Design eines übergreifenden, linguistisch motivierten Oralitäts-Maßes ein. [3]

Flankierend zu diesen Genre- und Textsorten-übergreifenden Arbeiten finden auf einzelne Medientypen fokussierte Auswertungen statt. Werner (2021) kontrastiert in einer multidimensionalen Registeranalyse (englischsprachige) Songtexte mit anderen Textsorten und weist trotz üblicherweise geplanter Textproduktion verschiedene mündlich/konversationelle Charakteristika nach. Androutsopoulos (2003) und Schlobinski (2005) beschreiben linguistische Features in Online-Sprache. Speziell für den Bereich der computervermittelten Kommunikation (Computer-Mediated Communication, CMC) untersucht Rehm (2002) die Verteilung eines kleinen Sets mutmaßlicher Merkmale konzeptioneller Mündlichkeit auf universitären Webseiten. Storrer (2000) thematisiert internetbasierte Kommunikationsformen wie E-Mail, Online-Foren und -Chats unter Bezugnahme auf entsprechende quantitative Erhebungen zur Einordnung in das Nähe-Distanz-Kontinuum. Für die Auswertung eines Chat-Korpus unterscheidet Kilian (2001) 14 linguistische Merkmale konzeptioneller Mündlichkeit, Cotgrove (2017) erweitert dieses Featureset im Rahmen einer Untersuchung von 600 deutschsprachigen Online-Kommentaren zu Youtube-Musikvideos.

3 Datengrundlage

3.1 Stratifikation des Untersuchungskorpus

Das Design unseres Untersuchungskorpus soll die beiden angenommenen Pole „konzeptionelle Mündlichkeit“ bzw. „konzeptionelle Schriftlichkeit“ unter Heranziehung einer aussagekräftigen Belegmenge authentischer Sprachsamples abdecken. Zudem dient es der experimentellen Verortung weiterer Sprachdaten aus verschiedenartigen Kommunikationskontexten. Zu diesen Zwecken fächern sich die Primärdaten auf in 14 gleich große Subkorpora mit einem Gesamtumfang von ca. 28 Millionen Wort-Token (vgl. Tabelle 1). Gemeinsame Datengrundlage sind die *Datenbank Gesprochenes Deutsch (DGD)* (Schmidt, 2017) und das *Deutsche Referenzkorpus (DeReKo)* (Kupietz et al., 2018) als jeweils umfassendste deutschsprachige Sammlung ihrer Art, sowie das Songkorpus (Schneider, 2020). Berücksichtigt werden ausschließlich komplette, ungekürzte Texte. Soweit mithilfe von Metadaten realisierbar, wurde bei der Stichprobenziehung auf eine diachron und regional ausgewogene Mischung geachtet.

Folgende Subkorpora enthalten **prototypisch konzeptionell mündliche Texte** aus den *DGD*-Gesprächskorpora:

(1) Deutsch Heute (DH): Regional stratifizierte gebrauchssprachliche Sammlung. Sie „umfasst alle Nationen und Regionen Mitteleuropas, in denen Deutsch heute als Amts- und Unterrichtssprache verbreitet ist“ (Kleiner, Berend, Knöbl & Brinckmann, 2014, S. 184). In unser Untersuchungskorpus fließen daraus nur Aufnahme-

[3] COAST (Conceptual Orality Analysis and Scoring Tool); online unter `https://github.com/rubcompling/COAST`.

transkripte freier Rede ein, zumeist biografische Erzählungen mit wechselnden Themenschwerpunkten. Vorleseaufgaben, die ebenfalls zum DH-Inventar zählen, werden explizit ausgefiltert.

(2) Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK): Transkribierte Audioaufzeichnungen authentischer Spontansprache, „die verschiedenste Interaktionstypen aus den Bereichen privater (z. B.Tischgespräche, Telefongespräche, Spielinteraktionen, Gespräche bei privaten Aktivitäten), institutioneller (z. B. schulischer Unterricht, Verkaufsgespräche, Fahrschulstunden, berufliche Gespräche, universitäre Prüfungsgespräche) und öffentlicher Kommunikation (z. B. Podiumsdiskussion, Schlichtungsgespräch) abdecken (Schmidt, 2018, S. 216)".

Folgende Subkorpora enthalten **prototypisch konzeptionell schriftliche Texte** aus den DeReKo-Schriftsprachkorpora:

(3) Zeitschriften: Zeitlich (diachron) ausgewogenes Sample der Jahrgänge 1970 bis 2021 des Nachrichtenmagazins *Spiegel*. Angesichts der über diesen Zeitraum erwartbaren Autorenfluktuation und dem damit verbunden vernachlässigbaren Risko unregelmäßiger Häufungen („Clumpiness") einzelner Autoren sowie der ebenfalls gegebenen Streuung der regionalen Autorenherkunft verzichten wir auf die Heranziehung weiterer Pressetitel. Zur Vermeidung möglicher Überschneidungen mit dem Subkorpus 'Interview' (s.u.) werden keine als 'Gespräch' oder 'Interview' deklarierten Beiträge einbezogen.[4]

(4) Zeitungen: Zufällige Auswahl von Inhalten des *Mannheimer Morgen*, zeitlich ausgewogen stratifiziert über die zurückliegenden vier Jahrzehnte. Analog zum Zeitschriftensubkorpus wird auf eine weitergehende Diversifizierung nach Pressetitel verzichtet.

Das Untersuchungskorpus weist damit eine weitestgehend ausgeglichene Menge an mutmaßlich konzeptionell schriftlichen Tokendaten und konzeptionell mündlichen Tokendaten auf (jeweils ca. 4 Millionen). Wohl wissend, dass ein einzelner Text grundsätzlich sowohl typische Nähe- als auch typische Distanzmerkmale aufweisen kann, lautet die Prämisse unserer Studie: Die Zuordnung der jeweils kompletten o.g. Texte zu einem der beiden Pole erlaubt das Trainieren statistischer Modelle für Nähe/Mündlichkeit bzw. Distanz/Schriftlichkeit.

Darüber hinaus umfasst unser Untersuchungskorpus verschiedene ad hoc **nicht eindeutig zuordenbare Textsammlungen** mit ebenfalls je 2 Millionen Token, die sich mutmaßlich in unterschiedlichen Bereichen des Mündlich-Schriftlich-Kontinuums bewegen:

[4] Zur Veranschaulichung des quantitativen Hintergrunds: Sämtliche Jahrgänge der originalen *Spiegel*-Daten zusammengenommen umfassen ca. 210 Millionen Token, davon entfallen ca. 23,5 Millionen Token auf Gespräche/Interviews. In unser Untersuchungskorpus werden aus den verbleibenden Daten ca. 40.000 Token pro Jahr aufgenommen.

(5) Belletristik: Berücksichtigt wird Unterhaltungsliteratur (Romane und Erzählungen) des 20. und 21. Jahrhunderts. Das datengebende DeReko-Korpus enthält Texte diverser Schriftsteller im Umfang von knapp 10 Millionen Token, verfasst zwischen 1997 und 2012. Daraus werden für jedes Jahr komplette Texte mit insgesamt 2 Millionen Token zufällig extrahiert.

(6) DGDMISC: In dieses Subkorpus fließen weitere Aufnahmen gesprochener deutscher Sprache aus der *Datenbank Gesprochenes Deutsch* ein. Es enthält Auszüge zehn heterogener *DGD*-Korpora, z. B. 'Biographische und Reiseerzählungen', 'Berliner Wendekorpus', 'Dialogstrukturen', 'Elizitierte Konfliktgespräche', 'Gesprochene Wissenschaftssprache', König-Korpus, Pfeffer-Korpus.

(7) Interviews: Das Subkorpus speist sich aus drei DeReKo-Quellen: Zum einen enthält es *Stern*- und *Zeit*-Interviews aus den Jahren 1992-1994, weiterhin diverse zwischen 1999 und 2003 geführte politische Interviews und schließlich zeitlich gleichmäßig verteilte und zufällig ausgewählte *Spiegel*-Interviews 1970-2021.

(8) Liveticker: Berücksichtigt werden Live-Berichterstattungen aus der Domäne 'Fußball' zwischen 2006 und 2016. Das Subkorpus enthält zu gleichen Teilen „in Echtzeit geschriebene und im Internet publizierte Reportagen zu laufenden Spielen" (Meier-Vieracker, 2018, S. 3) aus den Korpora zur Fußballlinguistik sowie vom Fachportal *kicker.de*.

(9) Reden: Das Subkorpus versammelt zufällig ausgewählte und zeitlich ausgewogen verteilte Plenarprotokolle des Landtags von Rheinland-Pfalz aus den Jahren von 1996 bis 2012.

(10) SocialMedia: Abgedeckt werden interaktive Kommunikationskanäle in Form von Online-Kurznachrichten. Das Subkorpus enthält zu gleichen Teilen eine zufällige Auswahl 2021 aufgezeichneter Twitter-Tweets sowie Beiträge aus dem Dortmunder Chat-Korpus (Beißwenger, 2013) aus den Jahren 1998 bis 2006.

(11) SocialMedia-L: Das Subkorpus enthält längere Kurznachrichten zur Abschwächung kürzebedingter Effekte. Es basiert auf den selben Quellen wie die vorstehende Auswahl; im Unterschied dazu werden Texte allerdings nicht zufällig ausgewählt, sondern pro Jahr diejenigen mit der höchsten Tokenanzahl priorisiert. Die Anzahl der Einzeltexte reduziert sich dadurch ungefähr um den Faktor 5.

(12) Songtexte: Das Subkorpus enthält deutschsprachige Songtexte der zurückliegenden fünf Jahrzehnte, untergliedert in autorenspezifische, thematische sowie popularitätsbasierte Subarchive (Charts); siehe (Schneider, 2019a). Abgedeckt werden ost- und westdeutsche Künstler sowie unterschiedliche Genres wie Hiphop, Liedermacher, Pop und Rock (Schneider, 2019b).

(13) Wikipedia Nutzerdiskussionen (WikiDisk): Berücksichtigt werden die Online-Diskussionsseiten der deutschsprachigen Wikipedia-Community (Margaretha &

Lüngen, 2014). Abgedeckt wird der Zeitraum zwischen 2002 und 2017 mit vom Umfang her jeweils ähnlichen jährlichen, zufällig extrahierten Anteilen.

(14) Wissenschaftsschriftsprache (WissSchrift): Das Subkorpus enthält arbiträr ausgewählte Artikel aus *Gingko (Geschriebenes ingenieurwissenschaftliches Korpus)* (Schirrmeister, Rummel, Heine, Suppus & Mendoza Sánchez, 2021), die zwischen 2007 und 2016 in den Fachzeitschriften 'Automobiltechnische Zeitschrift (ATZ)' und 'Motortechnische Zeitschrift (MTZ)' erschienen sind.

Einzeltextvolumina der Subkorpora unterscheiden sich aus naheliegenden Gründen mitunter erheblich: So sind z. B. die 2 Millionen Token in der Belletristik-Sammlung mit durchschnittlich 57.887 Token pro Text bereits bei 36 Texten erreicht, im Social-Media-Subkorpus – das viele extrem kurze Tweets und Chats mit durchschnittlich nur 36 Token pro Text enthält – erst bei 55.041 Texten.

Subkorpus	Texte	Token	Token/Text
DH	667	2.003.377	3.004
FOLK	648	2.001.021	3.088
Zeitschrift	3.011	2.003.071	665
Zeitung	8.505	2.000.598	235
Belletristik	36	2.083.944	57.887
DGDMISC	1.061	2.000.630	1.886
Interviews	1.010	2.002.315	1.983
Liveticker	1.444	2.001.636	1.386
Reden	49	2.006.773	40.955
SocialMedia	55.041	2.000.012	36
SocialMedia-L	12.349	2.008.262	163
Songtexte	7.455	2.002.538	269
WikiDisk	4.051	2.000.159	494
WissSchrift	1.199	2.000.314	1668

Tabelle 1: Umfang und Untergliederung des Untersuchungskorpus. Die ersten vier Subkorpora dienen als Trainingsdaten für die Klassifikatoren.

Sämtliche Primärdaten sind angereichert um Lemmaangaben sowie Wortklassenannotationen nach dem für das Deutsche einschlägigen Stuttgart-Tübingen-Tagset (STTS). Die DGD-basierten Subkorpora verwenden das für gesprochene Sprache modifizierte STTS 2.0 (Westpfahl, Schmidt, Jonietz & Borlinghaus, 2017), Songtexte eine nochmals für spezielle Phänomene -– insbesondere zusammengezogene Wörter — erweiterte Spezifikation (Schneider, 2022, S. 45).

3.2 Das Featureset

Die von Ágel und Hennig (2006a, 2006b) konzipierte Systematik, die linguistisch fundiert eine Begründung des Nähestatus einzelner Merkmale liefert, lässt sich ohne tiefgehende computerlinguistische Analyse und Annotation der zur Disposition stehenden Textdaten schwerlich operationalisieren. Wir orientieren uns deshalb zunächst am von Ortmann und Dipper (2019) verwendeten Featureset und erweitern dieses um zusätzliche linguistisch motivierte Merkmale bzw. lassen diejenigen Features außen vor, die sich für das Untersuchungskorpus nicht berechnen lassen. Letzteres betrifft satzbasierte Merkmale wie Satzlänge, Satzeinleiter und Satztyp ('Fragesatz', 'Ausrufesatz'), weil verschriftete mündliche Daten zumeist ohne Interpunktionszeichen daherkommen. In der Konsequenz kann unsere Modellierung als eine Art Machbarkeitsstudie betrachtet werden, die opportunistisch zunächst mit einer Reihe verfügbarer Merkmale arbeitet in der Hoffnung auf empirisch fundierte Hinweise zu deren Relevanz und Gewichtung.[5]

In der nachfolgenden Übersicht unterscheiden wir Merkmale, die ausschließlich unter Rückgriff auf die Textoberfläche generiert werden (Tabelle 2), von solchen, die für die Identifizierung von Belegen ergänzend den Output (maschineller) Tagger heranziehen (Tabelle 3) und damit u.a. von deren Methodik, Annotationsinventar und Güte abhängen.

Durch das Heranziehen unterschiedlicher Maße mit variierenden methodischen Details zielen wir auf die Erarbeitung konvergierender Evidenz bei der optimalen konzeptuellen Einstufung (*Permutation Feature Importance* und *T-Test*) bzw. informationeller Mehrwerte hinsichtlich der Wirkungsrichtung (Δ):

Die *Permutation Feature Importance* gibt die Random Forest-Schätzung der mittleren Abnahme an Genauigkeit bei Weglassen eines Features wieder.

Wir prüfen mit dem *Welch-T-Test*, ob sich die Datenmittelwerte eines Features für beide Klassen (konzeptionell mündlich bzw. konzeptionell schriftlich) unterscheiden, mit einer Irrtumswahrscheinlichkeit von 0.05 (*), 0.01 (**) und 0.001 (***).

Δ charakterisiert in Vorzeichenform die Wirkungsrichtung eines Features, basierend auf der Differenz zwischen seinen Mittelwerten für konzeptionell mündliche vs. konzeptionell schriftliche Texte (+ für höheren Mittelwert bei mündlichen Texten bzw. − für höheren Mittelwert bei schriftlichen Texten).

3.2.1 Merkmale ohne Tagging-Informationen

Folgende **lexikalische Merkmale**, die sich ohne Zuhilfenahme computerlinguistischer Annotationen direkt auf der Textoberfläche bestimmen lassen, fließen in unsere Klassifizierung ein:

- Lexikalische Vielfalt wird bereits von Koch und Oesterreicher (1985, S. 454) als Unterscheidungskriterium angeführt und beruht auf der Annahme einer tendenziell höheren lexikalischen Varianz in Distanztexten. Hier wie bei allen statistischen Merkmalen gilt: Selbstverständlich sind Kommunikationssituationen (=

[5] Wir danken Mathilde Hennig für wertvolle Anregungen zur linguistischen Einordnung vieler dieser Merkmale.

Merkmal	Typ	Beschreibung	Δ	T-Test
STTR	lexikalisch	Lexikalische Vielfalt (Standardisierte Type-Token Ratio	–	***
MATTR	lexikalisch	Lexikalische Vielfalt (Moving-Average Type-Token Ratio)	–	***
MLTD	lexikalisch	Lexikalische Vielfalt (Measure of Textual Lexical Diversity)	–	***
PRON1st_wf	lexikalisch	Verhältnis Personalpronomina 1. Pers. zu allen Wörtern	+	***
PRON2nd_wf	lexikalisch	Verhältnis Personalpronomina 2. Pers. zu allen Wörtern	+	***
bloss	lexikalisch	Verhältnis Gradpartikel *bloß* zu allen Wörtern	+	***
lediglich	lexikalisch	Verhältnis Gradpartikel *lediglich* zu allen Wörtern	–	***
kriegen	lexikalisch	Verhältnis Verb *kriegen* zu allen Wörtern	+	***
word_mean	morphologisch	Wortlänge (Mittelwert)	–	***
word_med	morphologisch	Wortlänge (Median)	–	***
elision_ART	morphologisch	Verhältnis indefinite Artikel mit Elision zu Formen ohne Elision	+	***
elision_VA	morphologisch	Verhältnis Hilfsverben mit Elision zu Formen ohne Elision	+	***
VERBshort	morphologisch	Verhältnis gekürzter Verbformen nach *ich* zu allen Verbformen nach *ich*	+	***
contracted	morphologisch	Verhältnis Wortkontraktionen zu allen Wörtern	+	***
PTK_MOD	pragmatisch	Verhältnis Modalpartikeln zu allen Wörtern	+	***
repetition	pragmatisch	Verhältnis Wortwiederholungen zu allen Wörtern	+	***
stretch	pragmatisch	Verhältnis Wörter mit Buchstaben-Reduplikation zu allen Wörtern	–	**

Tabelle 2: Klassifikations-Features ohne Tagging-Informationen.

Texte) denkbar, in denen das Merkmal ohne Effekt bleibt oder sogar in eine andere Richtung weist. Anstelle einfacher Type-Token-Ratios (TTR), die stark mit Textgrößen korrelieren und deshalb wenig aussagekräftig scheinen, evaluieren wir drei differenziertere Maße: (a) *Standardized Type-Token Ratio (STTR)* zerteilt Texte in aufeinander folgende Fenster von 100 Wörtern und berechnet deren TTR-Gesamtmittelwert; (b) *Moving-Average Type-Token Ratio (MATTR)* von Covington und McFall (2010) arbeitet ebenfalls mit Wortfenstern, verschiebt diese allerdings sequenziell und begegnet damit dem Problem übrig bleibender Segmente mit weniger als 100 Wörtern; (c) *Measure of Textual Lexical Diversity (MLTD)* schließlich ist ein von Mccarthy und Jarvis (2010) beschriebenes komplexes Maß, das in mehreren Berechnungsschritten und -richtungen einen vergleichsweise robusten Vielfaltswert kalkuliert. Unvorteilhaft bleibt, dass sich keines dieser Maße uneingeschränkt für sehr kurze Texte empfiehlt.

- Mit den Maßen *PRON1st_wf* und *PRON2nd_wf* messen wir die Verwendung der ersten und zweiten Person in Texten und kalkulieren relative Frequenzen der Personalpronomen *ich, mich, mir, wir, uns* bzw. *du, dich, dir, ihr, euch.* Motiviert ist das Maß vom mutmaßlich größeren Bedarf in Nähetexten; vgl. u.a. Ágel/Hennig oder Biber, Johansson, Leech, Conrad und Finegan (1999).

- Stilistisch könnten Gebrauchsfrequenzen der Gradpartikeln *bloß* vs. *lediglich* auf einen der beiden Pole hindeuten: „Vor allem mündlich wird *bloß* vorgezogen, insbesondere in informellen Kontexten. *Lediglich* findet sich vorzugsweise in geschriebener Sprache“ (Zifonun et al., 1997, S. 877).

- Ähnliches gilt für die Verwendung von *kriegen* als Auxiliar des Dativpassivs; diese „gehört heute fast ausschließlich in die mündliche Umgangssprache“ (Zifonun et al., 1997, S. 1829). Wir mutmaßen eine insgesamt häufigere Verwendung der Wortform *kriegen* in Nähe/Mündlichkeit und messen die relativen Frequenzen in der Hoffnung auf klassifikatorischen Ertrag.

Weiterhin untersuchen wir **morphologische Merkmale** im Sinne von Features, die auf Phänomene der strukturellen Morphologie referieren:

- Längenverteilungen spielen eine prominente Rolle in der Sprachstatistik und bei der Formulierung quantitativer Sprachregularitäten. Wir betrachten Wortlängen in der Annahme, dass morphologisch komplexere - und damit längere – Wörter als elaborierter und damit tendenziell eher der Schriftsprache zuzuordnen sind. Berechnet werden Mittelwert (*word_mean*) sowie Median (*word_med*); zur Plausibilität und Verwendung für unser Sprachmodell vgl. Abschnitt 4.3.

- Elisionen als sprachökonomisch motivierte Vokalauslassungen im Wortinnern oder am Wortende treten „in mündlicher Sprache weit häufiger auf als in der – in dieser Hinsicht konservativeren – Schriftsprache“ (Grammis, 2020). Zur statistischen

Abbildung modellieren wir drei Merkmale: (a) *Elision_ART* berechnet das Verhältnis indefiniter Artikel mit Elision zu Formen ohne Elision, also z. B. *ne* bzw. *'ne* vs. *eine*; (b) *Elision_VA* steht für das Verhältnis von Hilfsverben mit Elision zu Realisierungen ohne Lautausfall unter Berücksichtigung sämtlicher Personen und Zeitformen also z. B. *hab, hab', hatt, hatt', werd, werd'* usw. vs. *habe, hatte, werde* usw. (c) *VERBshort* drückt das Verhältnis gekürzter Verbformen nach *ich* zu allen Verbformen nach *ich* aus. Berücksichtigt werden also nicht auf *-e* endende Realisierungen wie *ich geh* vs. *ich gehe*.[6]

- Zusammenziehungen aufeinander folgender Wörter gelten als ebenfalls charakteristisch für den mündlichen Diskurs und entsprechen wie auch Elisionen dem „auf das Verfahren der Sprecheinheitenbildung rückführbar[en]" Merkmal 'phonisches Wort' bei Ágel und Hennig (2006a, S. 60). Unser Merkmal *contracted* misst das Verhältnis von Wortkontraktionen – also *gibts, gibt's, haste, kannste, machste* usw. – jeweils zu allen Textwörtern.

Schließlich evaluieren wir **pragmatische Merkmale**, die Sprachdiskurse strukturieren, Einzelaspekte verstärken oder Stimmungen ausdrücken:

- Modal-/Abtönungspartikeln wie *auch, bloß, denn, doch, eben, eh* usw. „operieren auf Erwartungen und Einstellungen" der Kommunikationspartner und können gesprächssteuernd wirken, indem sie dazu beitragen, „Äußerungen in den jeweiligen Handlungszusammenhang zu integrieren" (Grammis, 2018). Ob sie sich damit für eine Verortung im Nähe-Distanz-Kontinuum eignen, überprüfen wir anhand des Merkmals *PTK_MOD*. Mangels passgenauer STTS-Annotation operieren wir dabei notgedrungen listenbasiert auf der Wortoberfläche, was eine Abgrenzung zu anderen Verwendungen verhindert und die Merkmalsschärfe mindert.

- Wortwiederholungen können als rhetorische Figur zur Aussageverstärkung eingesetzt werden; vgl. 'holistische Gefühlsäußerung durch Reduplikation' bei Ágel/Hennig. Mithilfe des Merkmals *repetition* berechnen wir die relativen Häufigkeiten von mehrfachen Wortnennungen wie in *immer immer wieder* oder *sehr, sehr gut*.[7]

- Reduplikationen einzelner Buchstaben zur Emulierung emotionaler Prosodie finden sich prominent in Verschriftlichungen konzeptioneller Mündlichkeit; vgl. (Rehm, 2002), (Schneider, 2022). Unser Merkmal *stretch* misst relative Häufigkeiten solcher 'Stretchwörter' wie *soooo, suuuper, tschüßiiiii* oder *Jetzt geht's lo-oos*.

[6] Eichinger (2017, S. 312) weist darauf hin, „dass hier im Modus des Sprechens Formen erscheinen, die wie eine Abschleifung der geschriebenen Form durch die (normale) Geschwindigkeit des Sprechens (Allegro-Sprechen) erscheinen, es aber nicht sind oder zumindest nicht in jedem Fall sein müssen".

[7] Nicht berücksichtigt werden Wiederholungen von Wortfolgen wie *O Gott, o Gott*.

3.2.2 Merkmale mit Tagging-Informationen

Einige der vorstehend genannten Features lassen sich alternativ zur Erkennung auf der Sprachoberfläche auch (und ggf. sogar exhaustiver bzw. exakter) unter Zuhilfenahme computerlinguistischer Annotationen identifizieren. Andere sind zwingend auf Angaben zu Wortklasse (POS) oder Lemma angewiesen. In unsere Untersuchung fließen deshalb auch Merkmale ein, die auf solchen maschinell erstellten Tagging-Informationen basieren.

Dies betrifft folgende **lexikalische Merkmale**:

- Bereits Halliday (1985) nimmt eine geringere lexikalische Dichte in Gesprochenem an. Für einzelne Texte berechnen wir dafür das Merkmal *lexDens* als Verhältnis von Inhaltswörtern (Autosemantika: Nomina, Vollverben, Adjektive, Adverbien) zu allen Textwörtern.

- Demonstrativpronomina können im Diskurs eingesetzt werden, um auf Personen, Gegenstände oder Sachverhalte im Wahrnehmungsfeld zu verweisen. Eine durch verstärkten Gebrauch eventuell verbundene Hinweisfunktion auf Nähesprache kodiert das Merkmal *DEM* und nutzt für die Erkennung das STTS-Tag *PDS*. Flankierend berechnen wir als Feature *DEMshort* das Verhältnis kurzer Demonstrativpronomina wie *der, die, das, den, dem* usw. zu allen Demonstrativpronomina, also incl. der Langformen *dies, diese, dieser, diesen, dieses*; zur Unterscheidung ziehen wir zusätzlich die Lemmata heran (*die, d* vs. *diese, dies, die, d*).

- *PRON1st* und *PRON2nd* kodieren analog zu *PRON1st_wf* und *PRON2nd_wf* relative Frequenzen der Personalpronomina. Die Vorkommen ermitteln wir hier auf Basis der POS-Tags (*PPER*) sowie der Lemmata (*ich, wir* bzw. *du, ihr*).

- Das Merkmal *V_N* dient einer statistischen Annäherung an Verbal- bzw. Nominalstil. Ersterer gilt als lebendiger und eher umgangssprachlich anzutreffen, letzterer findet sich eher in sprachökonomisch optimierten (fachlichen) Distanztexten. Stark vereinfachend untersuchen wir hierfür das Verhältnis von Vollverben zu Nomina in den einzelnen Texten.

Ergänzende **morphologische Merkmale** mit Tagging-Informationen beschränken sich auf die konkatenative Morphologie ausgewählter Wortklassen:

- Die Merkmale *autosem_mean* bzw. *autosem_med* bestimmen Mittelwert bzw. Median der Wortlänge sämtlicher Autosemantika eines Texts. Auf diese Weise konzentrieren wir uns auf Wortklassen, die morphologische Komplexität durch Komposition oder Flexion überhaupt erst ermöglichen. Synsemantika – meist ohne Längenvariation – beeinflussen hierbei nicht den statistischen Aussagewert.

Als **pragmatische Merkmale** im Kommunikationszusammenhang kommen hinzu:

- Interjektionen als Merkmale mündlicher Kommunikation können spontane Empfindungen ausdrücken (*ächz, au, seufz* usw.), als Pausen-/Überbrückungs- (*äh,*

hm usw.) oder Aufmerksamkeitsmarker (*hey, ey* usw.) dienen und vieles mehr. Für die Berechnung des Features *INTERJ* beschränken wir uns auf Ein-Wort-Interjektionen.

- Antwortpartikeln sind typische hörerseitige Diskursbeiträge. Für das Merkmal *PTKANT* filtern wir nicht nach Wortform oder Lemma, sondern nutzen die STTS-Wortklassenannotationen. Neben *ja* und *nein* werden auch Formen wie *jaaa, nee, gewiss, bitte, danke* etc. abgedeckt.

Auch wenn sich womöglich aussagekräftigere Merkmale wie fehlerhafter Satzbau oder Selbstkorrekturen ohne tiefergehendes syntaktisches Tagging nicht algorithmisiert identifizieren lassen, können basierend auf den vorhandenen Annotationsebenen doch einige **syntaktische Merkmale** bestimmt werden:

- Passivstil deutet tendenziell auf Konstruiertheit und damit Distanzverortung eines Texts hin. Für das Feature *passiv* zählen wir – massiv vereinfacht und als experimentelle Annäherung – auf der Oberfläche erkennbare *werden*-Passive und berücksichtigen hierfür Partizip II-Verbformen unmittelbar nach dem passivbildenden Hilfsverb (realisiert als *werde, wirst, wird, werden, werdet, wurde, wurdest, wurden, wurdet*) bzw. mit maximal einem Zwischenwort; auch Belege mit umgekehrter Anordnung ohne optionale Lücke fließen in die Berechnung ein: *Der Bürgermeister wurde gestern abgewählt* oder *Das Problem soll geklärt werden.*

- Spontansprache gilt als syntaktisch weniger komplex als geplante Sprache. Das Feature *subord* dient der Identifizierung von Hypotaxen, also der Unterordnung von Nebensätzen in Satzgefügen. Wir setzen hierzu – wiederum stark vereinfachend – vom Tagger als unterordnende Konjunktionen erkannte Vorkommen von *um, anstatt, weil, dass* usw. in Beziehung zu allen Vollverben.

4 Textklassifikation

Wir trainieren binäre Klassifikatoren zur Unterscheidung konzeptioneller Nähe/Mündlichkeit bzw. Distanz/Schriftlichkeit auf Einzeltextebene. Neben der Klassifikationsgüte bewerten wir die Relevanz der oben eingeführten Merkmale. Anschließend verorten wir auch nicht-eindeutige Texte im Kontinuum.

Hierzu werden verschiedene Klassifikatoren auf einem Teil der vier in Tabelle 1 erstgenannten Subkorpora trainiert (Abschnitt 4.1). Für DH und FOLK wird dabei konzeptionelle Mündlichkeit und für die beiden DeReKo-Subkorpora (Zeitschriften und Zeitungen) konzeptionelle Schriftlichkeit angenommen; jeder Einzeltext ist ein Datenpunkt. Damit führen wir die statistischen Berechnungen auf Basis authentischer, breit gestreuter und gleichzeitig zeitgenössischer Datensamples durch, um die Modellierung der beiden Pole auf ein solides Fundament zu stellen. In Abschnitt 4.3 ermitteln wir redundante Variablen, die dann für die nachfolgenden Schritte ausgeschlossen werden.

Merkmal	Typ	Beschreibung	Tagging	Δ	T-Test
lexDens	lexikalisch	Verhältnis Autosemantika (ADJ.*, ADV, N.*, VV.*) zu allen Wörtern	POS	+	***
DEM	lexikalisch	Verhältnis Demonstrativpronomina (PDS) zu allen Wörtern	POS	+	***
DEMshort	lexikalisch	Verhältnis PDS-Kurzformen zu allen Demonstrativpronomina	POS, Lemma	+	***
PRON1st	lexikalisch	Verhältnis Personalpronomina (PPER) 1. Pers. zu allen Wörtern	POS, Lemma	+	***
PRON2nd	lexikalisch	Verhältnis Personalpronomina (PPER) 2. Pers. zu allen Wörtern	POS, Lemma	+	***
V_N	lexikalisch	Verhältnis Vollverben zu Nomina	POS	+	***
autosem_mean	morphologisch	Länge Autosemantika (Mittelwert)	POS	–	***
autosem_med	morphologisch	Länge Autosemantika (Median)	POS	–	***
INTERJ	pragmatisch	Verhältnis Ein-Wort-Interjektionen zu allen Wörtern	POS	–	***
PTK_ANT	pragmatisch	Verhältnis Antwortpartikeln zu allen Wörtern	POS	–	***
passiv	syntaktisch	Verhältnis partizipiale Verbformen vor oder hinter *werden*-Hilfsverb zu allen Verbformen	POS	–	***
subord	syntaktisch	Verhältnis unterordnender Konjunktionen (KOUS, KOUI) zu Vollverben	POS	+	***

Tabelle 3: Klassifikations-Features mit Tagging-Informationen.

Auf dieser Basis trainieren und evaluieren wir in Abschnitt 4.4 Klassifikatoren mit verschiedenen Feature-Teilmengen. Zwei dieser Klassifikatoren setzen wir schließlich in Abschnitt 4.5 für die Einordnung von Texten ein, die – wie z. B. Songtexte – intuitiv nicht eindeutig einer polaren Konzeption zugeordnet werden können und mutmaßlich irgendwo im Kontinuum zwischen Oralität und Literalität liegen.

4.1 Einteilung der Datensets

Die durchschnittlichen Primärtextgrößen unterscheiden sich zwischen DGD- und DeReKo-basierten Subkorpora um den Faktor 5 bis 13, so dass trotz gleicher Gesamttokenzahl grundsätzlich wesentlich mehr konzeptionell schriftliche Texte als konzeptionell mündliche Texte für das Training des Klassifikators zur Verfügung stehen. Da das Trainieren auf eine hohe generelle Genauigkeit (*accuracy*) abzielt, besteht hier die Gefahr, dass die Accuracy für die Minderheitenklasse (hier *specificity*) wesentlich schlechter ist als die für die Mehrheitsklasse (hier die *sensitivity*). Dem begegnen wir durch partielle Reduktion der Trainingsdaten (*downsampling*): Es werden aus der Menge der konzeptionell schriftlichen Texte zufällig die gleiche Anzahl an Texten gewählt, wie sie maximal für konzeptionelle Mündlichkeit bereitstehen.[8].

Konzeption	Texte
schriftlich	11516
mündlich	1315

Tabelle 4: Verteilung der Texte nach konzeptioneller Verortung bei gleicher Tokenzahl.

Die verbleibenden Daten werden im Verhältnis von 80% zu 20% in ein Development- und ein Validierungsset aufgeteilt. Das Developmentset teilen wir wiederum im 80/20-Verhältnis in ein Trainings- und ein Testset (*Out-of-Bag*-Stichprobe (OOB) zur unvoreingenommenen Bewertung z.B. von Parameteranpassungen beim Trainieren der Modelle). Tabelle 5 verdeutlicht die Aufteilung.[9]

4.2 Software und statistischer Ansatz

Die vorliegende Studie implementiert ein überwachtes Lernverfahren („supervised learning“) in Form eines Random-Forest-Algorithmus. Random Forest ist eine nichtparametrischer Klassifikationsmethode, die Ergebnisse verschiedener – zufällig und unkorreliert generierter – Entscheidungsbäume („decision trees“) kombiniert und ihre

[8] *Upsampling/bootstrap sampling* der Minderheitenklasse ist zwar grundsätzlich auch möglich, aber der Einfluss der unabhängigen Variablen ließe sich dann nicht zuverlässig bestimmen (Strobl, Boulesteix, Zeileis & Hothorn, 2007)

[9] Da die Gütekriterien auf den Trainingsdaten gemeinhin immer gut sind, wird die Güte auf Grundlage des Testsets bewertet und der Klassifikator entsprechend entwickelt. Um ein Overfitting auf das Testset zu vermeiden, wird nach Abschluss der Entwicklung die Güte des fertigen Klassifikators mit dem Validierungsset ermittelt.

Konzeption	Trainingsset	Testset	Validierungsset	
schriftlich	839	213	263	1.315
mündlich	844	208	263	1.315
	1.683	421	526	2.630

Tabelle 5: Anzahl der Texte nach Konzeptionsklassen pro Datenset.

Zuordnungen auf Basis dieser Einzelentscheidungen vornimmt. Das Verfahren wird als Stichprobennahme ohne Zurücklegen („subsampling without replacement“) realisiert, weil sich damit der Effekt konfundierender Variablen auf das Importance Measure verringert: Unser Vorgehen vermeidet, dass einige uninformative Variablen aufgrund ihrer Struktur häufiger einfließen als andere (vgl. Strobl et al., 2007).

Wir nutzen R in der Version 4.2.0 (R Core Team, 2022) und *cforest* aus dem *party*-Package (Hothorn, Buehlmann, Dudoit, Molinaro & van der Laan, 2006; Strobl, Boulesteix, Kneib, Thomas & Zeileis, 2008; Strobl et al., 2007) mit Default-Parametern. Abbildungen und Tabellen sind mit *ggplot2* (Wickham, 2016) und unter Zuhilfenahme von *xtable* (Dahl, Scott, Roosen, Magnusson & Swinton, 2019) erstellt. Für die *Variable Importance Measures* (VIMs) kommt das Package *permimp* (Debeer, Hothorn & Strobl, 2021) zum Einsatz. Die Güte-Werte wurden mit Hilfe von *caret* (Kuhn, 2022) extrahiert.

4.3 Ausschluss redundanter Variablen

Tokenlängen liegen als Durchschnittswerte (`word_mean` / `autosem_mean`) und als Mediane (`word_median` / `autosem_mean`) vor. Mediane sind robuster bei Ausreißern, allerdings gibt es Informationsverluste in Bezug auf Extrempunkte. Durchschnittswerte sind datengetreuer, andererseits können hier Ausreißer den Wert verzerren. Eine Klassifizierung mit beiden Längenmaßen erscheint unnötig. Um zu entscheiden, welche Variablen final Verwendung finden sollen, trainieren wir einen ersten Random Forest mit allen Features.

Auf Basis der *Permutation Importance* (*Mean Decrease Accuracy*, siehe Abbildung 1) zeigen sich dabei die Durchschnittswerte (`autosem_mean` und `word_mean`) als aussagekräftiger. Die Importance bleibt auch bei bei mehrmaligem Trainieren stabil.

Bestätigend beobachten wir, dass die Güte eines Klassifikators mit ausschließlich diesen beiden Prädiktoren am besten ist (Tabelle 6) und es eine Verbesserung sowohl zu Klassifikatoren mit jeweils nur einem Prädiktor als auch zu solchen mit anderen Variablenkombinationen gibt. Die Unterschiede sind allerdings nicht groß und betragen meist $< 1\,\%$, mit Ausnahme eines Klassifikators (`word_med`), der ca. 6,4 % weniger Testset-Daten richtig und ca. 13,7 % der konzeptionell mündlichen Daten falsch klassifiziert.

4.4 Vergleich verschiedenartiger Klassifikatoren

4.4.1 Einbezogene Features

Insgesamt trainieren wir Modelle für drei Klassifikatoren:

- Ein Klassifikator K_{Gesamt}, der – unter Ausschluss redundanter Features (s. Unterabschnitt 4.3) – alle Variablen beinhaltet.
- Ein Klassifikator $K_{Tagging}$ nur mit Variablen, die auf Tagging-Informationen basieren (s. Tabelle 3). Fehlannotationen können sich dabei auf die Performanz des Klassifikators auswirken.[10]
- Ein Klassifikator $K_{Oberfläche}$ nur mit Variablen, für deren Berechnung keine Tagging-Informationen genutzt werden, um den Einfluss eventueller Annotationsfehler abzuschätzen.

UV	Acc.	Acc.P.	B.Acc.	F1	Prec.	Sens.	Spec.
word_mean	0,9810	< 0,001	0,9810	0,9812	0,9812	0,9812	0,9808
word_med	0,9216	< 0,001	0,9315	0,9160	0,8451	1,0000	0,8631
autosem_mean	0,9834	< 0,001	0,9834	0,9836	0,9859	0,9813	0,9855
autosem_med	0,9834	< 0,001	0,9834	0,9836	0,9859	0,9813	0,9855
autosem_mean + word_mean	0,9857	< 0,001	0,9858	0,9860	0,9906	0,9814	0,9903
autosem_mean + word_med	0,9834	< 0,001	0,9834	0,9836	0,9859	0,9813	0,9855
autosem_med + word_med	0,9834	< 0,001	0,9834	0,9836	0,9859	0,9813	0,9855
autosem_med + word_mean	0,9834	< 0,001	0,9834	0,9836	0,9859	0,9813	0,9855

Tabelle 6: OOB-Güte für Klassifikatoren mit permutierten `autosem-` und `word-` Features.

4.4.2 Klassifikationsgüte der polaren Texte

Tabelle 7 illustriert, dass die Daten im Validierungsset beinahe ausnahmslos korrekt eingestuft werden. Am besten schneidet $K_{Oberfläche}$ ab, der Unterschied zu den beiden anderen Klassifikatoren ist jedoch kaum spürbar (0,0038 Accuracy-Punkte). Hier könnte sich der Umstand auswirken, dass ohne Rückgriff auf maschinelle Annotationen berechnete Variablen störunanfälliger sind.

[10] Der auf standardnahe Sprache trainierte Tagger weist Schwächen etwa bei der Beurteilung von Wortzusammenziehungen, Interjektionen und Named Entities in 'standardferneren' Texten auf.

Klassifikator	Acc.	Acc.P.	B.Acc.	F1	Prec.	Sens.	Spec.
Gesamt	0,99429	$< 0{,}001$	0,9943	0,9943	0,9962	0,9924	0,99618
Tagging	0,99429	$< 0{,}001$	0,9943	0,9943	0,9962	0,9924	0,99618
Oberfläche	0,99809	$< 0{,}001$	0,9981	0,99809	0,9962	1,0000	0,99621

Tabelle 7: Güte der Klassifikatoren auf dem Validierungsset.

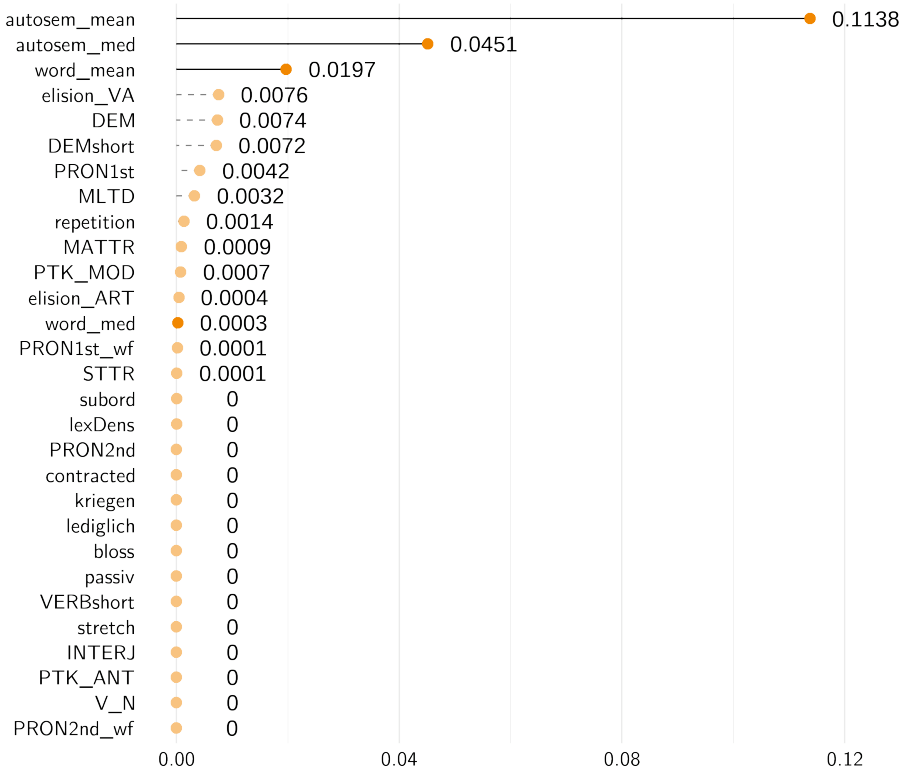


Abbildung 1: Permutation Importance für alle untersuchten unabhängigen Variablen. Die farblich hervorgehobenen Variablen sind untereinander redundant.

Die Accuracy (Acc.) ist bei allen Klassifikatoren hochsignifikant besser als die *No Information Rate*, also die beste Schätzung, wenn keine über die Gesamtverteilung der Klassen hinausgehenden Informationen vorliegen (Acc.P. mit $p < 0{,}001$). Der Accuracy-Durchschnitt (*Balanced Accuracy*, B.Acc.) fällt durchgehend nochmals minimal besser aus; dieser Unterschied lässt sich darauf zurückführen, dass die Anzahlen der Texte mit schriftlicher oder mündlicher Konzeption im Validierungsset nicht perfekt ausgeglichen sind (vgl. Tabelle 5).

Die F1-Werte, das harmonische Mittel aus *Precision* (der Anteil der tatsächlich schriftlichen Texte an allen als schriftlich klassifizierten Texte) und *Recall* (gleich *Sensitivity*, der Anteil aller richtig als schriftlich klassifizierten Texte an allen tatsächlich schriftlichen Texte unabhängig von ihrer Klassifikation), unterscheiden sich ebenfalls nur minimal von Acuracy bzw. Balanced Acuracy.

Die Accuracy-Werte legen nahe, dass alle drei Klassifikatoren zuverlässig arbeiten. Die Modelle unterscheiden sich nur geringfügig darin, wie zuverlässig sie Texte sprachlicher Konzeption (*Sensitivity* und *Precicion*) und Texte mündlicher Konzeption richtig klassifizieren (*Specificity*). Die *Precision* (Prec.) fällt bei allen Klassifikatoren gleich aus, jedoch ist sie bei K_{Gesamt} und $K_{Tagging}$ besser als die *Sensitivity* (Sens.). Bei $K_{Oberfläche}$ dagegen ist die *Sensitivity* besser als die *Precision*.

Die *Specificity* ist bei K_{Gesamt} und $K_{Tagging}$ gleich, bei $K_{Oberfläche}$ einen Hauch besser.[11] Von $K_{Oberfläche}$ werden Texte schriftlicher Konzeption gar nicht (*Sensitivity* = 1) und Texte mündlicher Konzeption (minimal) seltener fehlklassifiziert (*Specificity*).

4.5 Verortung der nicht-polaren Texte

Wir nutzen K_{Gesamt} und den insgesamt leistungsstärksten Klassifikator $K_{Oberfläche}$ für die experimentelle Verortung konzeptioneller Mischfälle auf Textebene – also derjenigen Texte, die wir nicht einem der beiden Pole zuschlagen und deren Einordnung im Kontinuum insofern ungesichert ist. Außer den 10 bislang nicht klassifizierten Subkorpora aus Tabelle 1 beziehen wir auch die beim Downsampling herausgefallenen schriftsprachlichen Texte ein. Dabei betrachten wir, ob die oben beschriebenen minimalen Güteunterschiede einen wahrnehmbaren Einfluss auf die Klassifikation nehmen.

Tabelle 8 ordnet die Subkorpora absteigend anhand des Anteils der von K_{Gesamt} als konzeptionell schriftlich eingestuften Texte. Wie erwartbar liegen die (zusätzlichen) Zeitungs- und Zeitschriftentexte mit über 99% am oberen Ende der Skala, werden allerdings noch übertroffen durch die wissenschaftssprachlichen Texte (ebenfalls intuitiv erwartbar) und Reden (nicht überraschend in Anbetracht der für Plenarprotokolle üblichen Vor- bzw. Nachbearbeitung). Die Klassifizierung der Zeitungs- und Zeitschriftentexte durch $K_{Oberfläche}$ fällt geringfügig eindeutiger aus, was sich mit der besseren *Sensitivity* (siehe Tabelle 7) des Modells deckt. Ein Unterschied der Sensitivität um nur 0,0076 Punkte (0,76 Prozentpunkte) macht hier demnach einen Accuracy-Unterschied von 0,2 % (Zeitungen) bzw. 0,5 % (Zeitschriften) aus.

[11] Dieser Unterschied hat bei der Klassifikation von konzeptionell mündlichen Texten einen Einfluss, wie in Abschnitt 4.5 deutlich wird.

	Gesamt			Oberfläche		
	schriftlich	mündlich	%	schriftlich	mündlich	%
WissSchrift	1.199	0	100 %	1.199	0	100 %
Rede	49	0	100 %	49	0	100 %
Zeitung	7.499	25	99,7 %	7.513	11	99,9 %
Zeitschrift	2.662	15	99,4 %	2.675	2	99,9 %
Liveticker	1.403	41	97,2 %	1.444	0	100 %
WikiDisk	3.846	205	94,9 %	4.001	50	98,8 %
Interview	955	55	94,6 %	978	32	96,8 %
SocialMedia	51.909	3.125	94,3 %	51.045	3.989	92,8 %
SocialMedia-L	10.773	1.576	87,2 %	11.469	880	92,9 %
Belletristik	28	8	77,8 %	31	5	86,1 %
Songs	3.965	3.490	53,2 %	2.255	5.200	30,2 %
DGDMISC	324	737	30,5 %	303	758	28,6 %

Tabelle 8: Textklassifikation mit Gesamt-Modell und Oberflächen-Modell. Die %-Spalten geben die prozentualen Anteile der als konzeptionell schriftlich eingestuften Texte an.

Inhalte des Subkorpus `DGDMISC` werden - ebenfalls erwartbar - als vorrangig mündlich klassifiziert, auch hier arbeitet $K_{Oberfläche}$ etwas eindeutiger als K_{Gesamt}. Der Unterschied in der *Specificity* um nur 0,003 Prozentpunkte macht hier einen 1,9-prozentigen Unterschied bei der Zuordnung aus.

Bemerkenswert sind die Klassifikationsergebnisse für einige andere Subkorpora, z.B.:

- Bei Interviews, also verschriftlichten Gesprächen, würde man intuitiv von einer tendenziell mündliche(re)n Konzeption ausgehen. Beide Klassifikatoren verorten allerdings den ganz überwiegenden Großteil dieser Texte als konzeptionell schriftlich.

- Der Einfluss der unterschiedlichen Textlängen in den beiden SocialMedia-Subkorpora fällt bei K_{Gesamt} größer aus als bei $K_{Oberfläche}$: Zweiterer klassifiziert die Inhalte beider Korpora fast gleich häufig als konzeptionell schriftlich bzw. mündlich, ersterer klassifiziert weniger SocialMedia-L-Texte (87,2 %) als SocialMedia-Texte (94,3 %) als konzeptionell schriftlich.

- Die beiden Modelle beurteilen Songtexte in Teilen unterschiedlich. K_{Gesamt} verortet die Datenpunkte (= Texte) weitestgehend ausgeglichen (3490 mündlich, 3965 schriftlich). $K_{Oberfläche}$ klassifiziert Songtexte mehrheitlich (69,8 %) als konzeptionell mündlich.

5 Diskussion

In diesem Abschnitt zeigen wir Gründe für unterschiedliche Klassifikationen auf und beleuchten überraschende Zuordnungen im Detail.

5.1 Falsche Zuordnungen im Validierungsset

Zunächst erörtern wir aus qualitativer Perspektive verschiedene Fehlklassifikationen im Validierungsset. Damit soll ein Gefühl für die Wirkungsweise der Klassifikatoren vermittelt sowie ein möglicher qualitativer Analyseansatz demonstriert werden.

Die Klassifikatoren haben im Validierungsset nur sehr wenige Texte nicht wie erwartet eingeordnet (vgl. Tabelle 7): K_{Gesamt} 3 von 526, $K_{Oberfläche}$ 1 von 526. Wir identifizieren folgende Faktoren, die Fehlklassifikationen begünstigen:

- Es gibt zum einen unter den einflussreichen Features mit den größten PI-Werten (Permutation Importance) Fälle, in denen der fehlklassifizierte Text innerhalb einer Spanne liegt, in der sich Werte konzeptionell mündlicher und schriftlicher Texte überschneiden.
- Zum anderen kann ein fehlklassifizierter Text innerhalb der Zielgruppe, der er eigentlich zugeordnet werden sollte, in Bezug auf ein (einflussreiches) Merkmal ein statistischer Ausreißer (Outlier) sein.
- Gleichzeitig kann er allerdings auch innerhalb der anderen Gruppe in Bezug auf das gleiche Merkmal in die Quartilsspanne fallen (wie beispielsweise Text 75645 bei `elision_VA`, der als Outlier unter den konzeptionell schriftlichen Texten in die Interquartilsspanne der konzeptionell mündlichen Texte fällt und vom gemischten Klassifikator fälschlicherweise als konzeptionell mündlich klassifiziert wird; siehe Abbildung 2).

5.1.1 Fehlklassifikationen des Gesamtmodells

K_{Gesamt} stuft Text 75645[12] fälschlicherweise als konzeptionell mündlich und 00361 bzw. 00159 fälschlicherweise als konzeptionell schriftlich ein. $K_{Oberfläche}$ schlägt Text 00260 unerwartet mündlicher Nähe zu. Nachfolgend inspizieren wir diese Texte genauer.

Ein Blick in die Primärinhalte enthüllt Text 75645 als kurzen Feuilleton-Zeitungsartikel mit popkulturellem Bezug. Darin finden sich keine komplexen Wörter und die durchschnittliche Wortlänge fällt entsprechend kurz aus. Dafür erkennt der Klassifikator fälschlicherweise eine auf Mündlichkeit hindeutende Hilfsverbelision (tatsächlich handelt es sich um die englische Wortform 'is') sowie mehrere mit Tagging-Unterstützung identifizierte kurze Demonstrativpronomina (Feature `DEMshort`). Der Oberflächen-Klassifikator plädiert auf konzeptionelle Distanz/Schriftlichkeit.

[12] Die originalen Text-Siglen sind hier aus Platzgründen gekürzt.

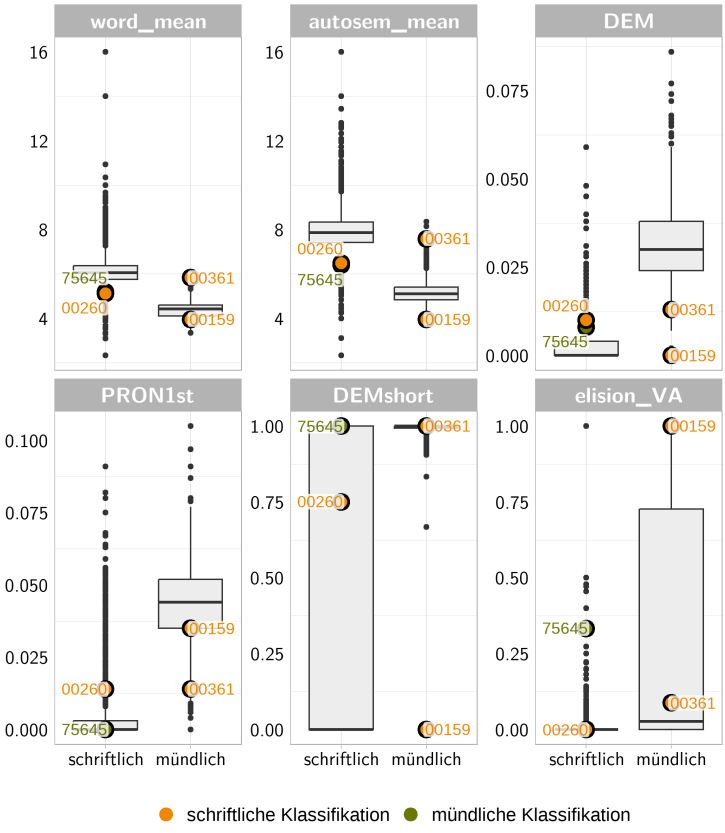


Abbildung 2: Boxplots für die sechs einflussreichsten Features des Gesamtmodells K_{Gesamt} mit farblich markierten Fehlklassifikationen.

FOLK-Transkript 00361 gibt eine längere wissenschaftliche Podiumsdiskussion wieder, 00159 die knappe Anmeldung an der Pforte eines Polizeireviers; beide Texte enthalten vergleichsweise wenig typische Nähe-/Mündlichkeitsmerkmale. Auch hier beurteilt $K_{Oberfläche}$ im Gegensatz zu K_{Gesamt} korrekt.

Abbildung 2 visualisiert die Situation. Sie zeigt die sechs für die Klassifikation maßgeblichsten Merkmale, d.h. Features mit der höchsten PI (vgl. dazu Abbildung 1).

Text 75645 liegt bei den einflussreichen Wortlängen-Features am unteren Ende des ersten Quartils für konzeptionell schriftliche Texte. Es besteht bei beiden Features eine Überschneidung des ersten Quartils für konzeptionell schriftliche Texte mit dem vierten Quartil der konzeptionell mündlichen Texte – und der fehlklassifizierte Text fällt genau in diese Überschneidung. Bei Merkmal `DEM` liegt der Text in der Überschneidung des vierten Quartils der konzeptionell schriftlichen Texte und des ersten Quartils der konzeptionell mündlichen Texte.

Wie aber an der Position des von K_{Gesamt} korrekt klassifizierten Zeitschriften-Texts 00260 deutlich wird, kann diese Ambiguität nicht der alleinige Grund für die Fehlklassifikation sein. Die Werte von `elision_VA` und `DEMshort` geben Aufschluss: Bei Merkmal `elision_VA` ist der Text ein potenzieller Outlier unter den konzeptionell schriftlichen Texten, liegt aber noch bequem im dritten Quartil der konzeptionell mündlichen Texte. Ebenso liegt der Text bei Merkmal `DEMshort` im Vergleich zu allen konzeptionell schriftlichen Texte am Maximum – wobei der Mittelwert beim Minimum liegt –, im Vergleich zu konzeptionell mündlichen Texten jedoch beim Mittelwert.

Im Gegensatz zu den Wortlängen-Werten von Text 75645 sind diese für den fälschlicherweise als konzeptionell schriftlich eingeordneten Text 00361 eindeutiger, denn der Text ist für die mündlichen Nähe-Texte ein Outlier, fällt jedoch ungefähr auf den Mittelwert der schriftlichen Distanz-Texte. Schließlich hat der ebenfalls als konzeptionell schriftlich klassifizierte Text 00159 einen für mündliche Nähe-Texte kleinen Anteil an Demonstrativpronomina und verkürzten Demonstrativpronomina (`DEM` und `DEMshort`) und ist damit ein potentieller Outlier; für konzeptionelle Distanz/Schriftlichkeit sind die Werte dagegen durchschnittlich.

5.1.2 Fehlklassifikation des Oberflächenmodells

Der von $K_{Oberfläche}$ unerwartet als Nähetext eingestufte Zeitschriftenartikel 00260 enthält bei näherer Betrachtung tatsächlich mehrere auf Mündlichkeit hindeutende Merkmale. Es handelt sich um eine mit direkter Rede angereicherte Buchbesprechung.

Für die Zuordnung scheinen insbesondere die Features `word_mean`, `PTK_MOD` und `PRON1st_wf` verantwortlich: Bei `word_mean` liegt der Artikel in der Überschneidung des ersten Quartils der konzeptionell schriftlichen Distanztexte und des vierten Quartils der konzeptionell mündlichen Nähetexte. Bei `PTK_MOD` fällt eine Überschneidung des vierten Quartils der konzeptionell schriftlichen Distanztexte und des ersten Quartils der konzeptionell mündlichen Nähetexte auf. Darüber hinaus ist der Artikel bei `PRON1st_wf` ein potenzieller Outlier unter den konzeptionell schriftlichen Distanztexten, nicht aber bei konzeptionell mündlichen Nähetexten.

Damit ist speziell bei den drei Oberflächen-Features mit der höchsten Permutation Importance (PI; vgl. Abbildung 4) eine gewisse Ambiguität gegeben.

5.2 Geringer Einfluss einzelner Merkmale

Einige Features haben einen sehr geringen oder gar keinen Einfluss auf die Textklassifikation. Deutlich wird das an einer niedrigen bzw. auf 0 lautenden Permutation Importance (PI; vgl. Abbildung 3). Im Falle der *Standardized Type-Token Ratio (STTR)* lässt sich konstatieren, dass die konkurrierenden und für unterschiedliche Text- und Korpusgößen robusteren Wortschatz-Maße *MATTR* bzw. *MLTD* erfreulicherweise einen höheren Einfluss aufweisen und damit die lexikalische Vielfalt doch recht prominent im Modell abgebildet wird. In anderen Fällen lässt sich eine plausibler Nullwerte-Zusammenhang konstatieren: So werden beispielsweise in über 80 % aller Texte keine Passivkonstruktionen erkannt – vermutlich nicht zuletzt, weil nur sehr vereinfacht auf der Oberfläche analysiert wurde. In solchen Fällen kann der Klassifikator dann auch keine Grenze für die Aufteilung der Daten anhand des Passiv-Merkmals festlegen.

Subkorpus	*kriegen*	*bloß*	*lediglich*
FOLK	46,3%	8,02%	0%
DH	27,74%	10,5%	0,15%
DGDMISC	16,21%	11,5%	2,5%
Songs	11,59%	5,7%	0,32%
Interview	9,2%	5,54%	4%
Liveticker	3,19%	0,42%	6,44%
Zeitschrift	2,82	2,42%	4,64%
Email-L	2,22%	1,84%	1,86%
SocialMedia-L	0,83%	0,28%	0,4%
Email	0,6%	0,55%	0,4%
Zeitung	0,47%	0,36%	2,7%
WikiDisk	0,44%	0,17%	1,28%
SocialMedia	0,27%	0,09%	0,05%
Belletristik	0%	16,67%	2,78%
Rede	0%	0%	0%
WissSchrift	0%	0,33%	15,42%

Tabelle 9: Prozentuales Erscheinen dreier lexikalisch-stilistischer Merkmale in Subkorpora, absteigend angeordnet nach Abdeckung von *kriegen*.

Entsprechend gering ist generell das praktische Gewicht stilistischer Frequenzmaße wie *kriegen*, *bloß* und *lediglich* oder von Interjektionen, Kontraktionen, Stretchwörtern usw., weil viele Texte weder das eine noch das andere enthalten; vgl. Tabelle 9.[13]

[13] Diese Verteilung sagt selbstverständlich nichts über die eigentliche Nähe-Distanz-Klassifizierung aus, sondern dient bestenfalls als Indiz für die generelle Brauchbarkeit der Merkmale.

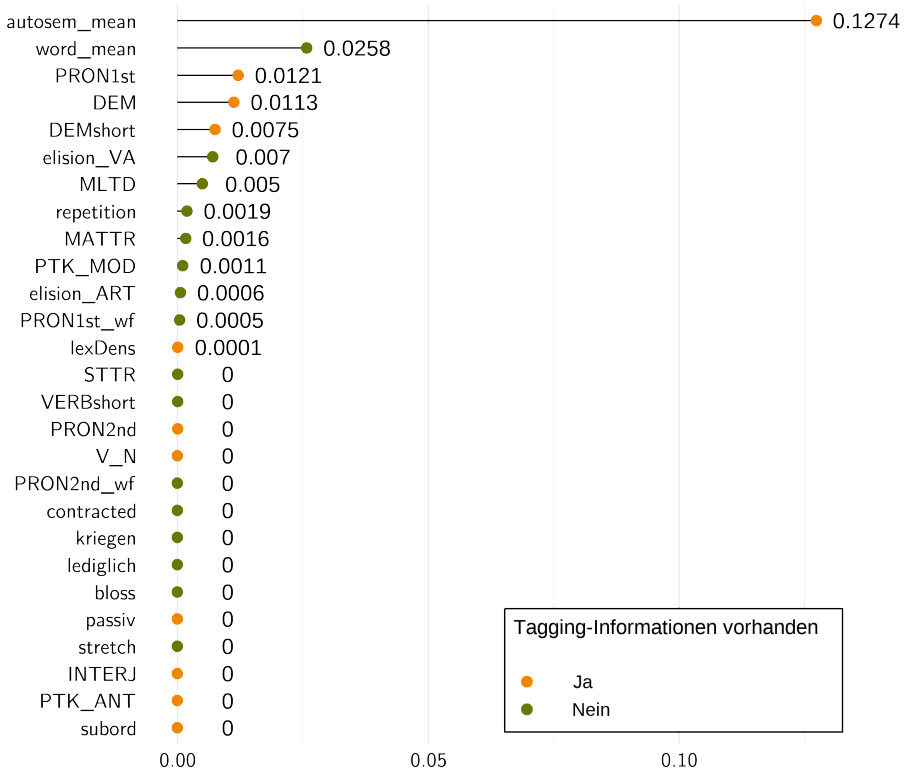


Abbildung 3: Permutation Importance (PI) für Features des Gesamt-Klassifikators.

Auch andere Merkmale können in den Daten diskriminieren, jedoch nicht so zuverlässig und durchgängig wie diejenigen mit hoher PI. Bei K_{Gesamt} spielen beispielsweise die Features `Verbshort` und `PRON2nd_wf` keine Rolle, bei $K_{Oberfläche}$ – der insgesamt weniger Features in Betracht zieht – nehmen sie dagegen nachweisbar Einfluss.

5.3 Nicht-polare Texte im Kontinuum

Nur in vergleichsweise wenigen Fällen unterscheiden sich die Klassifikationen der Modelle markant (vgl. Unterabschnitt 4.5); dies betrifft in erster Linie Songtexte. In anderen Fällen (Interviews, Social Media) hätte man intuitiv vielleicht eine höhere Tendenz zu Nähe/Mündlichkeit erwartet.

Für Interviews analysieren wir nachfolgend anhand der Permutation Importance, wie die mehrheitliche Zuordnung zu Distanz/Schriftlichkeit zustande kommt. Für die unterschiedlich langen Social-Media-Texte gehen wir kurz auf mögliche Gründe variierender Klassifikationen ein. Songtexte betrachten wir zudem aus diachroner und Musikgenre-Perspektive.

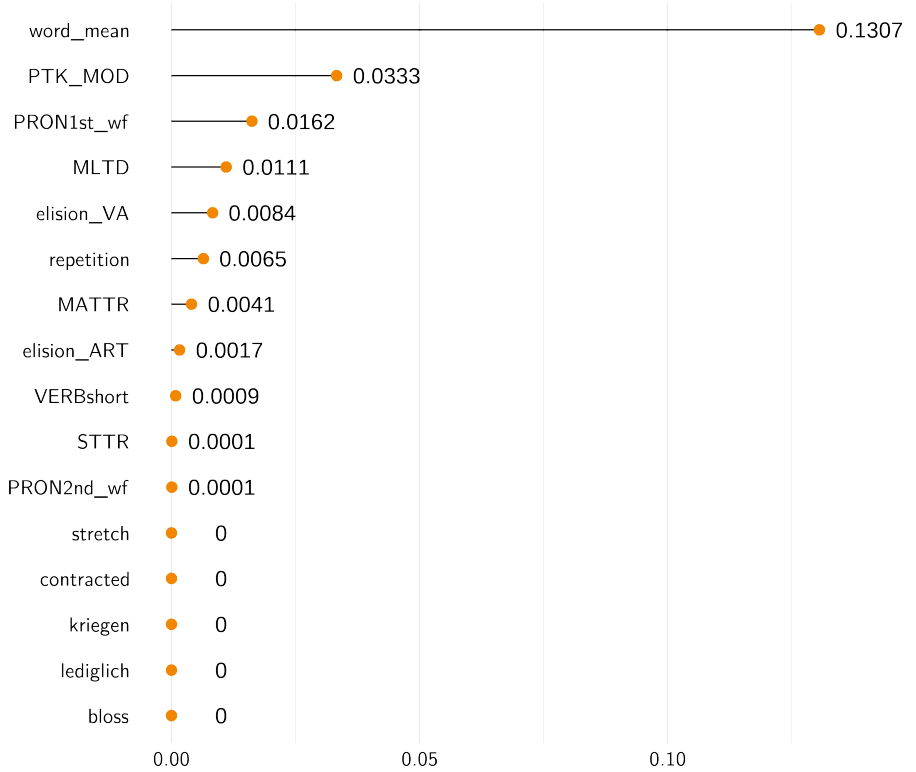


Abbildung 4: Permutation Importance (PI) für Features des Oberflächen-Klassifikators.

5.3.1 Interviews

Interviews im Untersuchungskorpus werden zu ca. 95% als konzeptionell schriftlich klassifiziert, wobei Tokenlänge-Merkmale wie auch bei anderen Textsorten beträchtlichen Einfluss nehmen (Abbildung 3 / Abbildung 4). Ein gezielter Blick auf die Merkmalsverteilung gibt Hinweise auf Hintergründe: Interviews ähneln in ihrer Verteilung der für

das Training eingesetzten Gruppe konzeptionell schriftlicher Texte. Dies verdeutlicht der Violinenplot (Abbildung 5): Je höher die Anzahl der Datenpunkte in einer Spanne, desto breiter ist die „Violine“. Sowohl bei den schriftlichen Distanztexten als auch bei Interviews häufen sich Werte zwischen 7 und 8 (`autosem_mean` (Durchschnitt 7,9/7,32; Median 7,87/7,37)) bzw. um 6 (`word_mean` (Durchschnitt 6,08 (schriftlich)/5,69 (Interview) ; 6,06/5,72)) herum. Bei konzeptionell mündlichen Texten liegen diese Werte eher bei 5 (Durchschnitt 5,15; Median 5,11) bzw. 4 (Durchschnnitt 4,37; Median 4,42).

Darüber hinaus fällt beim gemischten Klassifikator K_{Gesamt} ins Gewicht, dass eher wenige volle Demonstrativa verwendet und Hilfsverben selten elidiert werden. Dies scheint den größeren Anteil gekürzter Demonstrativa und den häufigen Gebrauch von Pronomina der ersten Person Singular zu überwiegen.

Beim Oberflächen-Klassifikator sorgen – neben der Wortlänge und den elidierten Hilfsverben – die seltene Verwendung von Modalpartikeln, eine höhere lexikalische Vielfalt (`MLTD`) sowie seltene unmittelbare Mehrfachnennungen von Wörtern für das tendenziell schriftliche Klassifikationsmuster. Eine plausible (hier nicht weiter beleuchtbare) Ursache liegt mutmaßlich in der für Printpublikationen üblichen Überarbeitung mündlich geführter Gespräche: Ziel solcher Revisionen ist üblicherweise nicht durchweg die exakte Sprecherwiedergabe, sondern eine gewisse mediale Anpassung. Dadurch lassen sich vergleichsweise wenige Wortwiederholungen und Elisionen erklären.

Unter den einflussreichsten Features, die eher in Richtung Nähe/Mündlichkeit steuern, fallen dagegen allein Pronomina der ersten Person ins Gewicht. Vor diesem Hintergrund erschließt sich sich die eindeutige Tendenz beider Klassifikatoren, Interviews mehrheitlich als Texte schriftlicher Konzeption zu klassifizieren.

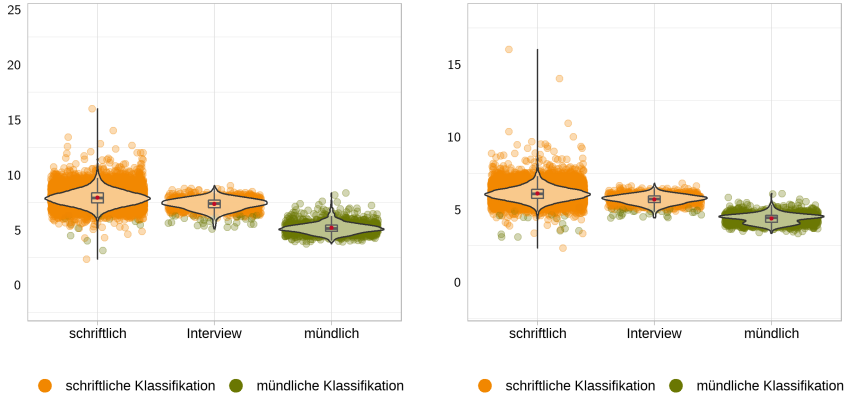
5.3.2 Social Media

Soziale Medien – Tweets und Chats – weisen die mit Abstand kürzesten Textlängen im Untersuchungskorpus auf. Inhalte aus SocialMedia-L mit etwas größeren Textlängen werden von K_{Gesamt} häufiger als konzeptionell mündlich klassifiziert als Inhalte ohne Mindestlänge (siehe Tabelle 8). $K_{Oberfläche}$ macht dagegen insgesamt keinen Unterschied zwischen den beiden Subkorpora.

Der Effekt hängt mutmaßlich damit zusammen, dass es in SocialMedia-L mehr Datenpunkte mit `DEM` > 0 gibt als in den Social-Media-Inhalten ohne Mindesttextlänge, also mehr Tweets und Chats, in denen Demonstrativpronomina überhaupt vorkommen. Da `DEM` für K_{Gesamt} eins der einflussreichsten Merkmale pro Nähe/Mündlichkeit ist (vgl. Abbildung 3), von $K_{Oberfläche}$ jedoch gar nicht beachtet wird, könnte dies die Unterschiede erklären.

Hinsichtlich der bei anderen Textsorten maximal einflussreichen Tokenlänge-Merkmale lassen sich in Abbildung 6 keine markanten Verteilungsunterschiede erkennen. Lediglich die Spannen sind in SocialMedia größer als in SocialMedia-L, Mediane und Mittelwerte unterscheiden sich kaum.

Eine mediale Aufgliederung der Social-Media-Daten zeigt markant unterschiedliche Ergebnisse für Tweets und Chats: Erstere werden unabhängig von Subkorpus oder



(a) `autosem_mean` **(b)** `word_mean`

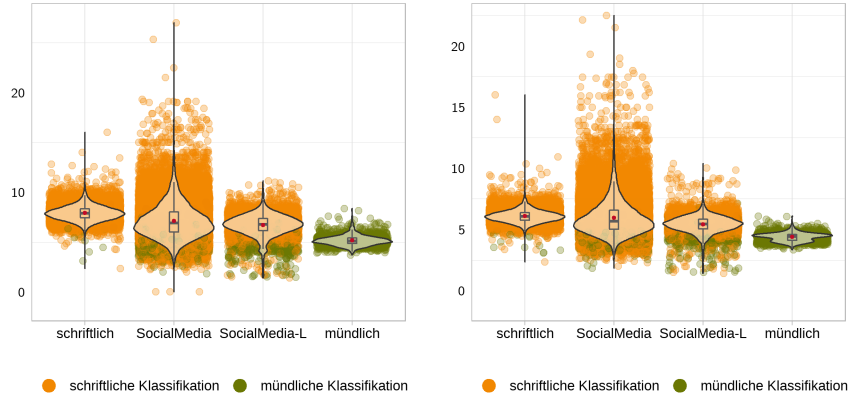
Abbildung 5: Verteilungen der Tokenlängen in Interviews. Es werden alle Datenpunkte (auch Trainingsdaten) abgebildet und entsprechend ihrer Verortung durch den Oberflächen-Klassifikator eingefärbt.

Klassifikationsmodell ganz überwiegend als konzeptionell schriftlich eingestuft. Letztere dagegen zu knapp unter 50% ($K_{Oberfläche}$) bzw. knapp über 50% (K_{Gesamt}) als konzeptionell mündlich. Unter der Prämisse, dass Chats häufiger unter Bekannten ausgetauscht, Tweets dagegen eher an offene Adressatenkreise gerichtet werden, verkörpern Nähe und Distanz die plausibleren Pol-Etiketten als Mündlichkeit und Schriftlichkeit.

5.3.3 Popsongs als „mittige Textsorte“

Songtexte werden von den beiden Klassifikatoren in unterschiedlichem Maß als konzeptionell mündlich bzw. schriftlich eingestuft (Tabelle 8): K_{Gesamt} verteilt beide Konzeptionen ungefähr gleich häufig, $K_{Oberfläche}$ verortet Songs signifikant häufiger als mündliche Nähetexte.

Wirkungsweisen einzelner Merkmale lassen sich auch hier am Beispiel von `DEM` herausstellen: K_{Gesamt}, der das Verhältnis von Demonstrativpronomina zu allen Wörtern eines Texts prominent für dessen Klassifizierung heranzieht, weist Songtexte ohne Demonstrativa (`DEM` $= 0$) verstärkt als konzeptionell schriftlich aus (Abbildung 7, links). Dies korreliert mit der aus den Violinplots ersichtlichen Dichteverteilung der Vergleichsgruppen: Einen Wert von 0 gibt es häufiger bei konzeptionell schriftlichen Texten, während ein höherer Wert häufiger in konzeptionell mündlichen Texten vorkommt. Im Gegensatz dazu, dass der PI-Wert von `DEM` bei K_{Gesamt} relativ hoch ist, wird dieses Merkmal von $K_{Oberfläche}$ nicht berücksichtigt, entsprechend häufiger kommt es dann zur Nähe/Mündlichkeits-Klassifizierung (Abbildung 7, rechts).



(a) `autosem_mean` **(b)** `word_mean`

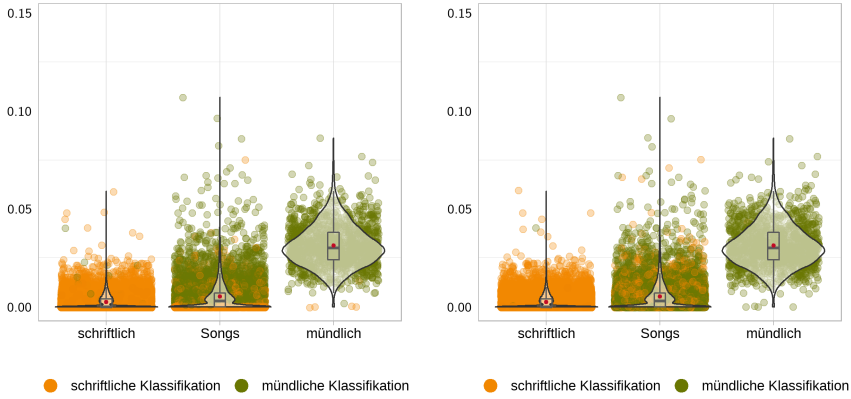
Abbildung 6: Violinplots der Tokenlänge-Merkmale in den Social-Media-Subkorpora. Daten sind auf Basis des Oberflächen-Klassifikators eingefärbt.

Trotz solcher Unterschiede verorten beide Modelle Popsongs (mit einer bewusst weiten Auslegung von „Pop“) unisono als „mittige“ Textsorte. In keinem anderen Subkorpus verteilen sich die betrachteten Merkmale ähnlich ausgewogen; nur ein einziges Merkmal (`kriegen`) findet sich überhaupt nicht im Subkorpus. Tabelle 10 informiert über minimale und maximale Merkmalswerte und gibt jeweils den Mittelwert über alle Songtexte an.

Für eine Plausibilitätsbewertung der Textklassifikationen lohnt sich ein Blick in die Primär- und Metadaten. Dabei fallen Besonderheiten einzelner Künstler bzw. musikalischer Genres (im Songkorpus als Archive recherchierbar; vgl. `https://songkorpus.de`.) ins Auge:

- Klassifikator K_{Gesamt}, der im Songkorpus insgesamt ca. 53% Nähetexte ermittelt, ordnet „traditionelle“ Singer/Songwriter signifikant häufiger unter Distanz/Schriftlichkeit ein.[14] Am deutlichsten trifft das auf Texte von Hannes Wader zu, die zwischen Poesie und erzählender Prosa pendeln (129 konzeptionell schriftlich, 67 konzeptionell mündlich). Ebenfalls als mehrheitlich schriftsprachlich werden die Werke von Reinhard Mey und Tocotronic eingeordnet und interessanterweise mit Herbert Grönemeyer ein Künstler, bei dem erklärtermaßen die Musik vor dem Text kommt, der also beim Komponieren zunächst mit Nonsensphrasen arbeitet und die finalen Texte erst zur fertigen Melodie formuliert.

[14] Die oben angesprochene Tendenz von $K_{Oberfläche}$, Songtexte eher als konzeptionell mündliche Nähetexte zu verorten, fällt hier übrigens schwächer aus, was die Ergebnisse von K_{Gesamt} für das Subsample stützt.



(a) Datenpunkte nach Klassifikationen des Gesamt-Klassifikators eingefärbt.

(b) Datenpunkte nach Verortung des Oberflächen-Klassifikators eingefärbt.

Abbildung 7: Violinplots für das Merkmal `DEM`. Songtexte, bei denen `DEM` gleich oder um 0 ist, werden vom Gesamt-Klassifikator als konzeptionell schriftlich eingestuft. Der Oberflächen-Klassifikator beurteilt sie als konzeptionell mündlich.

- Derselbe Klassifikator gruppiert auch ein Datensubset von 500 Songtexten aus der damaligen DDR, die ein breites Spektrum künstlerischen Schaffens in Ostdeutschland zwischen 1970 und 1990 abdecken, zu zwei Dritteln in die Kategorie Distanz/Schriftlichkeit. Interpretatorisch ließe sich hier vermuten, dass eventuell ein zentral verordneter „künstlerischer Anspruch“ mitspielt, vielleicht auch der Versuch, mit metaphorischen Mitteln Widerstand zu transportieren bzw. Zensurbeschränkungen durch elaborierte sprachliche Kniffe zu umgehen.

- Bemerkenswert erscheint die Einordnung von ebenfalls deutlich über 60% eines 500 Songs umfassenden Datensubsets „Neue Deutsche Welle (NDW)“ von Ende der 1970er bis Mitte 1980er Jahre als schriftsprachliche Distanztexte. Unter dem Gesichtspunkt, dass NDW als Gegenbewegung zu „emotionalen“ Mainstream-Genres entstand und nicht wenige NDW-Bands sprachliche Kühle und Minimalismus als Stilmittel einsetzten, bietet sich diese modellbasierte Einordnung als Ausgangspunkt für anknüpfende Fragestellungen an.

- Schlüssig erscheint aufgrund der seinerzeitigen Popularität von NDW eine insgesamt vermehrte Distanz/Schriftlichkeit in Chartsongs der 1980er Jahre. Abbildung 8 untermauert diese Vermutung. Das zeitlich stratifizierte Subsample enthält sämtliche in den Top-100-Singlecharts platzierten deutschsprachigen Songtexte seit 1970. Es umfasst nicht zwangsläufg gleich viele Inhalte pro Jahr: Da Hitpara-

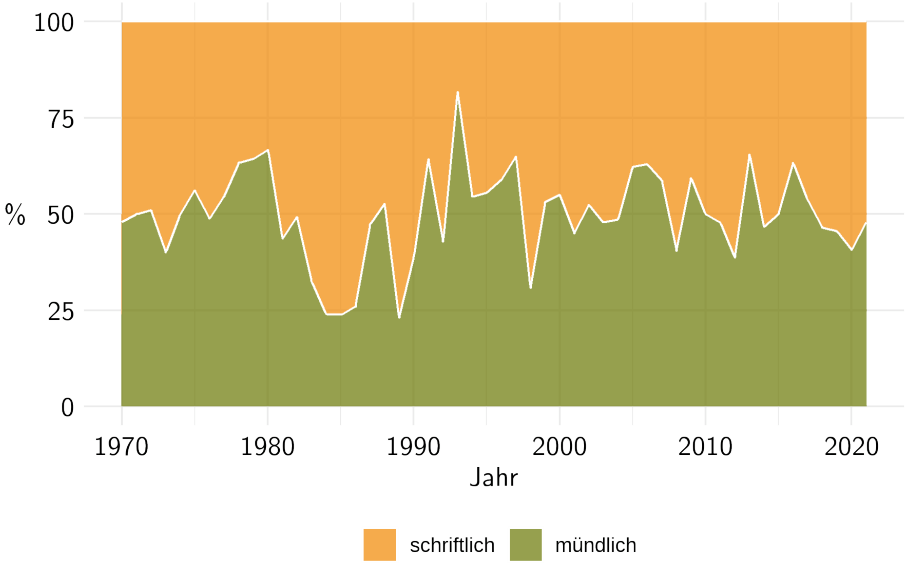


Abbildung 8: Nähe/Mündlichkeit und Distanz/Schriftlichkeit diachron in Chartsongs. Die grüne Fläche stellt den Anteil der als konzeptionell mündlich klassifizierten Texte pro Jahr dar, die orangene den der als konzeptionell schriftlich klassifizierten Texte.

den Moden und Trends reflektieren – also auch eine über die Jahre schwankende Popularität deutschsprachiger Popsongs – ist das Charts-Sample in diesem Sinne zwar repräsentativ, diachron aber nur bedingt ausgewogen.

- Ein umgekehrtes Bild und mit ca. 60% signifikant mehr Nähe als Distanz bietet zeitgenössischer Hiphop-Sprechgesang. Hierzu inspizieren wir ein Subsample mit 1000 Deutschrap-Songs (564 zu 436) sowie speziell Texte der Band 'Fettes Brot' (82 zu 53). Die Tendenz erscheint plausibel: Hiphop spricht in der Hauptsache ein junges Publikum an und gilt mit seinen Freestyle-Elementen, parataktischen Formen, Ellipsen usw. als vergleichsweise nahe an Umgangssprache. Interessanterweise ist Hiphop das einzige Musikgenre im Korpus, bei dem der Verzicht auf Tagging-Infos zu (marginal) weniger Nähe/Mündlichkeitsklassifikationen führt: 540 zu 459 ($K_{Oberfläche}$) vs. 563 zu 436 (K_{Gesamt}).

	Minimum	Maximum	Mittelwert
STTR	0	0,831	0,303
MATTR	0,025	0,970	0,328
MLTD	4,836	1357,235	115,693
PRON1st_wf	0	0,316	0,050
PRON2nd_wf	0	0,209	0,027
bloss	0	0	0
lediglich	0	0,004	0,000008
kriegen	0	0	0
word_mean	3,088	9,208	4,559
word_med	2	8	4,015
elision_ART	0	1	0,239
elision_VA	0	1	0,176
VERBshort	0	1	0,543
contracted	0	0,133	0,004
PTK_MOD	0	0,234	0,034
repetition	0	0,371	0,002
stretch	0	0,055	0,0002
lexDens	0,204	0,678	0,450
DEM	0	0,107	0,005
DEMshort	0	0,107	0,004
PRON1st	0	0,193	0,043
PRON2nd	0	0,202	0,023
V_N	0	51	0,799
autosem_mean	3,264	10,339	5,541
autosem_med	2	8,5	5,063
INTERJ	0	0,368	0,003
PTK_ANT	0	0,281	0,001
passiv	0	0,280	0,001
subord	0	2,182	0,140

Tabelle 10: Feature-Werte in Songtexten.

6 Zusammenfassung und Ausblick

Prototypische, konzeptionell mündliche Nähetexte lassen sich anhand unseres – für diese erste Annäherung notwendigerweise opportunistisch zusammengestellten – Featuresets sehr gut von konzeptionell schriftlichen Distanztexten unterscheiden. Wir haben hierfür Phänomene aufgegriffen, die in der sprachwissenschaftlichen Forschung als Markierer für Nähe/Mündlichkeit bzw. Distanz/Schriftlichkeit diskutiert werden, und die ohne tiefe syntaktische Analyse („deep parsing“) maschinell bestimmbar sind. Aufbauend auf der sehr guten binären Klassifikation haben wir die trainierten statistischen Modelle auf weitere Textsorten angewendet. Daraus ergibt sich keine unmittelbar linguistisch plausible Positionierung dieser Textsorten im Kontinuum. Allerdings lassen sich wertvolle Aussagen darüber ableiten, wie sich die herangezogenen Merkmale in situativ und medial heterogenen Sprachdaten verteilen und von den Klassifikatoren genutzt werden. Unser Verfahren ersetzt keinen systematischen theoretischen Überbau und keine ausdifferenzierte Methodik zur Beurteilung von Nähe-Distanz bzw. Mündlichkeit-Schriftlichkeit. Aber es illustriert die Praktikabilität empirischer Verfahren als evidenzbasierte Hilfestellung bei der Evaluation; die Auswahl bekannter Marker aus der linguistischen Forschung ermöglicht eine Einbettung statistischer Erkenntnisse in die Sprachtheorie.

Interessant erscheint die maschinelle Verortung von Popsongs als „mittige Textsorte“, die viele Nähe-Distanz-Merkmale vereint und den Gegenstandsbereich damit als hochattraktiv für die empirisch-deskriptive Sprachforschung ausweist.

Unsere breit stratifizierte Textsorten-Analyse untermauert bekannte Argumente, warum sich Nähe und Distanz als Pole zur Beschreibung von Äußerungsformen besser eignen als der mediale Dualismus Mündlichkeit/Schriftlichkeit: Wir konnten empirisch belegen, dass unsere Modelle mündlich kommunizierte Äußerungen wie Interviews oder Reden summa summarum vollkommen anders beurteilen als Telefon- oder Unterrichtsgespräche, die unter verschiedengestaltigen Kommunikationsbedingungen ablaufen und entsprechend divergierende Versprachlichungsstrategien wählen. Die Plausibilität des Nähe-Distanz-Konzepts wird unterstrichen durch die – nach Aufgliederung der Social-Media-Daten – aufgedeckten Klassifikationsunterschiede für Tweets und Chats.

Offenkundig besitzen Wortlänge-Features für sich genommen in den trainierten Modellen bereits eine hervorragende diskriminierende Vorhersagekraft; hier könnten Folgeuntersuchungen das Gewicht der übrigen Merkmale bei Wegfall dieser prominenten Markierer präziser ausloten. Dabei gälte es, weitere potenziell redundante Maße hinsichtlich ihrer wechselseitigen Abhängigkeit zu analysieren, mit dem Ziel, beispielsweise nur noch ein Maß für Aussagen zur lexikalischen Vielfalt (bislang: MLTD, MATTR, STTR) heranzuziehen. Einzelne Features bieten offenkundiges Optimierungspotenzial, etwa die Messung von Passivstil oder von Wortwiederholungen, bei denen außer Einzelwörtern auch Wortfolgen eine Rolle spielten sollten.

Ein naheliegendes Desiderat besteht in der Anreicherung unseres Textsortenspektrums um weitere bislang nicht berücksichtigte Textsorten. Auch die methodische Einbeziehung von Streuungs-/Dispersionsmaßen zur Bewertung der Verteiltheit von Merkmalen steht noch aus. Zur Überprüfung der Bewertungsgüte bzw. des Feature-Status bleibt ein breit

stratifizierter annotierter Goldstandard wünschenswert, in den Urteile linguistischer Experten einfließen. Damit verbindet sich die spannende Frage, wie sich eher qualitative Ansätze und Theorien mit maschinengestützten Verfahren abgleichen lassen. Vor diesem Hintergrund stellen wir unsere trainierten Klassifikationsmodelle wissenschaftsöffentlich zur Reproduktion, Evaluierung und Optimierung zur Verfügung.[15]

Literatur

Ágel, V. & Hennig, M. (2006a). Praxis des Nähe- und Distanzsprechens. In V. Ágel & M. Hennig (Hrsg.), *Grammatik aus Nähe und Distanz: Theorie und Praxis am Beispiel von Nähetexten 1650-2000* (S. 33-74). Tübingen: Niemeyer.

Ágel, V. & Hennig, M. (2006b). Theorie des Nähe- und Distanzsprechens. In V. Ágel & M. Hennig (Hrsg.), *Grammatik aus Nähe und Distanz: Theorie und Praxis am Beispiel von Nähetexten 1650-2000* (S. 3-31). Tübingen: Niemeyer.

Androutsopoulos, J. K. (2003). Online-Gemeinschaften und Sprachvariation : soziolinguistische Perspektiven auf Sprache im Internet. *Zeitschrift für germanistische Linguistik : deutsche Sprache in Gegenwart und Geschichte*, *31* (2), 173 – 197.

Barbour, S. & Stevenson, P. (1998). *Variation im Deutschen. Soziolinguistische Perspektiven.* Berlin, Boston: De Gruyter.

Beißwenger, M. (2013). Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik*, *41* (1), 161–164. Zugriff auf `https://doi.org/10.1515/zgl-2013-0009`

Biber, D. & Conrad, S. (2019). *Register, genre, and style* (2. Aufl.). Cambridge University Press. doi: 10.1017/9781108686136

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *The longman grammar of spoken and written english.* Harlow, UK.: Longman.

Cotgrove, L. A. (2017). *Is Computer-Mediated Communication "written Colloquial Speech" (Kilian 2001)? A Quantitative Study of German-Language YouTube Comments.* Nottingham: University of Nottingham. (MA Thesis)

Covington, M. & McFall, J. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*, 94-100. Zugriff auf `https://doi.org/10.1080/09296171003643098`

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A. & Swinton, J. (2019). xtable: Export Tables to LaTeX or HTML [Software-Handbuch]. Zugriff auf `https://CRAN.R-project.org/package=xtable`

Debeer, D., Hothorn, T. & Strobl, C. (2021). permimp: Conditional Permutation Importance [Software-Handbuch].

Dürscheid, C. (2016). *Einführung in die Schriftlinguistik. Mit einem Kapitel zur Typographie von Jürgen Spitzmüller* (5. Aufl.). Göttingen: Vandenhoeck & Ruprecht.

Eichinger, L. M. (2017). Gesprochene Alltagssprache. In Deutsche Akademie für Sprache und Dichtung/Union der deutschen Akademien der Wissenschaften (Hrsg.),

[15] `https://songkorpus.de/data/`

Vielfalt und Einheit der deutschen Sprache. Zweiter Bericht zur Lage der deutschen Sprache (S. 283 – 331). Tübingen: Stauffenburg.

Feilke, H. & Hennig, M. (Hrsg.). (2016). *Zur Karriere von ›Nähe und Distanz‹: Rezeption und Diskussion des Koch-Oesterreicher-Modells.* Berlin: De Gruyter.

Grammis. (2018). *Abtönungspartikeln.* Mannheim: Leibniz-Institut für Deutsche Sprache. Zugriff auf `https://grammis.ids-mannheim.de/systematische-grammatik/1322` (Grammatisches Informationssystem: Systematische Grammatik)

Grammis. (2020). *Elision.* Mannheim: Leibniz-Institut für Deutsche Sprache. Zugriff auf `https://grammis.ids-mannheim.de/terminologie/1169` (Grammatisches Informationssystem: Wissenschaftliche Terminologie)

Halliday, M. (1985). *Spoken and Written Language.* Geelong: Deakin University Press.

Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A. & van der Laan, M. (2006). Survival Ensembles. *Biostatistics*, *7* (3), 355–373.

Kilian, J. (2001). T@stentöne. Geschriebene Umgangssprache in computervermittelter Kommunikation. In M. Beißwenger (Hrsg.), *Chatkommunikation.* (S. 55–78). Stuttgart: Ibidem.

Kleiner, S., Berend, N., Knöbl, R. & Brinckmann, C. (2014). „Deutsch heute“: ein sprachgebietsweites Forschungsprojekt zur regionalen Variation in der gesprochenen deutschen Standardsprache. *Klagenfurter Beiträge zur Sprachwissenschaft*, *34-36*, 179 – 193. Zugriff auf `https://nbn-resolving.org/urn:nbn:de:bsz:mh39-28744`

Koch, P. & Oesterreicher, W. (1985). Sprache der Nähe — Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, *36* (1), 15–43. Zugriff auf `https://doi.org/10.1515/9783110244922.15`

Koch, P. & Oesterreicher, W. (2007). Schriftlichkeit und kommunikative Distanz. *Zeitschrift für germanistische Linguistik*, *35*, 346 - 375.

Kuhn, M. (2022). caret: Classification and regression training [Software-Handbuch]. Zugriff auf `https://CRAN.R-project.org/package=caret` (R package version 6.0-93)

Kupietz, M., Lüngen, H., Kamocki, P. & Witt, A. (2018). The German Reference Corpus DeReKo: New Developments – New Opportunities. In N. Calzolari et al. (Hrsg.), *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018).* Miyazaki, Japan: ELRA.

Margaretha, E. & Lüngen, H. (2014). Building linguistic corpora from Wikipedia articles and discussions. *Journal for Language Technology and Computational Linguistics (JLCL)*, *29* (2), 59–82. Zugriff auf `https://nbn-resolving.org/urn:nbn:de:bsz:mh39-33306`

Mccarthy, P. & Jarvis, S. (2010). Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, *42*, 381-92. Zugriff auf `https://doi.org/10.3758/BRM.42.2.381`

Meier-Vieracker, S. (2018). Fussball digital - korpuslinguistische Perspektiven auf die Sprache des Fußballs. *Sprachreport*, *34* (2), 1-9. Zugriff auf `https://nbn`

-resolving.org/urn:nbn:de:bsz:mh39-75266

Ortmann, K. & Dipper, S. (2019). Variation between different discourse types: Literate vs. oral. In *Proceedings of the sixth workshop on NLP for similar languages, varieties and dialects* (S. 64-79). Ann Arbor, Michigan: Association for Computational Linguistics. Zugriff auf https://aclanthology.org/W19-1407

Ortmann, K. & Dipper, S. (2020). Automatic orality identification in historical texts. In *Proceedings of the 12th language resources and evaluation conference* (S. 1293-1302). Marseille, France: European Language Resources Association. Zugriff auf https://aclanthology.org/2020.lrec-1.162

R Core Team. (2022). R: A Language and Environment for Statistical Computing [Software-Handbuch]. Zugriff auf https://www.R-project.org/

Rehm, G. (2002). Schriftliche Mündlichkeit in der Sprache des World Wide Web. In A. Ziegler & C. Dürscheid (Hrsg.), *Kommunikationsform E-Mail* (Bd. 7, S. 263–308). Tübingen: Stauffenburg.

Schirrmeister, L., Rummel, M., Heine, A., Suppus, N. & Mendoza Sánchez, B. (2021). Gingko – ein Korpus der ingenieurwissenschaftlichen Sprache. *Deutsch als Fremdsprache* (4). Zugriff auf https://doi.org/10.37307/j.2198-2430.2021.04.04

Schlobinski, P. (2005). Mündlichkeit/Schriftlichkeit in den Neuen Medien. In L. M. Eichinger (Hrsg.), *Standardvariation. Wie viel Variation verträgt die deutsche Sprache?* (S. 126 – 142). Berlin: de Gruyter.

Schmidt, T. (2017). DGD – die Datenbank für Gesprochenes Deutsch. *Zeitschrift für germanistische Linguistik*, *45* (3), 451-463. Zugriff auf https://doi.org/10.1515/zgl-2017-0027

Schmidt, T. (2018). Gesprächskorpora. In M. Kupietz & T. Schmidt (Hrsg.), *Korpuslinguistik* (S. 209-230). Berlin: De Gruyter.

Schneider, R. (2019a). *Corpus of Song Lyrics.* Mannheim. Zugriff am 19.03.2023 auf https://songkorpus.de

Schneider, R. (2019b). "Konservenglück in Tiefkühl-Town"– Das Songkorpus als empirische Ressource interdisziplinärer Erforschung deutschsprachiger Poptexte. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS)* (S. 229-236). Erlangen: German Society for Computational Linguistics (GSCL). Zugriff auf https://nbn-resolving.org/urn:nbn:de:bsz:mh39-93189

Schneider, R. (2020). A Corpus Linguistic Perspective on Contemporary German Pop Lyrics with the Multi-Layer Annotated "Songkorpus". In N. Calzolari et al. (Hrsg.), *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)* (S. 842-848). Paris: European Language Resources Association. Zugriff auf http://nbn-resolving.de/urn:nbn:de:bsz:mh39-100347

Schneider, R. (2022). Zwischen Schriftlichkeit und Mündlichkeit: Songtexte in der deskriptiven Sprachforschung. *Sprachreport*, *38* (1), 38-50. Zugriff auf https://nbn-resolving.org/urn:nbn:de:bsz:mh39-109499

Storrer, A. (2000). Schriftverkehr auf der Datenautobahn: Besonderheiten der schriftlichen Kommunikation im Internet. In G. G. Voš, W. Holly & K. Boehnke (Hrsg.), *Neue Medien im Alltag. Begriffsbestimmungen eines interdisziplinären*

Forschungsfeldes (S. 151-175). Opladen: Leske + Budrich.

Strobl, C., Boulesteix, A.-L., Kneib, T., Thomas, A. & Zeileis, A. (2008). Conditional Variable Importance for Random Forests. *BMC Bioinformatics published by Springer Nature*, *9* (307), 1–11. Zugriff auf https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, *8* (25). Zugriff auf https://doi.org/10.1186/1471-2105-8-25

Werner, V. (2021). Catchy and conversational? : a register analysis of pop lyrics. *Corpora : corpus-based language learning, language processing and linguistics*, *16* (2), 237–270. Zugriff auf https://fis.uni-bamberg.de/handle/uniba/51410

Westpfahl, S., Schmidt, T., Jonietz, J. & 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). Version 1.1* (Working Paper). Zugriff auf https://nbn-resolving.org/urn:nbn:de:bsz:mh39-60634

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis [Software-Handbuch]. New York: Springer-Verlag. Zugriff auf https://ggplot2.tidyverse.org

Zifonun, G., Hoffmann, L. & Strecker, B. (1997). *Grammatik der deutschen Sprache.* Berlin/New York: De Gruyter. Zugriff auf https://doi.org/10.1515/9783110872163

Korrespondenzanschrift

Sarah Broll
Leibniz-Institut für Deutsche Sprache
broll@ids-mannheim.de

Roman Schneider
Leibniz-Institut für Deutsche Sprache
schneider@ids-mannheim.de

Gertrud Faaß / Helmut Schmid

Segmentierungs- und Annotationsverfahren für die Texte Udo Lindenbergs: Apostrophe und andere Herausforderungen

Kurzfassung

In der Computerlinguistik ist eine kaskadische Prozessierung von Texten üblich. Dabei werden diese zuerst segmentiert (tokenisiert), d.h. Tokens und ggf. Satzgrenzen werden erkannt. Dabei entsteht meist eine Liste bzw. eine einspaltige Tabelle, die sukzessive durch weitere Prozessierungschritte um zusätzliche Spalten – also positionale Annotationen wie z.B. Wortarten und Lemmata für die Tokens in der ersten Spalte – ergänzt wird. Bei der Tokenisierung werden alle Spatien (Leerzeichen) gelöscht. Schon immer problematisch waren dabei Interpunktionszeichen, da diese äußerst ambig sein können, aber auch mehrteilige Namen, die Leerzeichen enthalten und eigentlich zusammengehören. Dieser Beitrag fokussiert auf den Apostroph, der in vielfältiger Weise in den Texten Udo Lindenbergs eingesetzt wird sowie auf mehrteilige Namen, die wir als Tokens erhalten möchten. Wir nutzen dafür das komplette Lindenberg-Archiv des *songkorpus.de*-Repositoriums, kategorisieren die auftretenden Phänomene, erstellen einen Goldstandard und entwickeln ein teils regel-, teils auf maschinellem Lernen basierendes Segmentierungswerkzeug, das insbesondere die auftretenden Apostrophe, aber auch - lexikonbasiert - mehrteilige Namen nach unseren Vorstellungen erkennt und tokenisiert. Im Anschluss trainieren wir den RNN-Tagger (Schmid, 2019) und zeigen auf, dass ein spezifisch für diese Texte angepasstes Training zu Genauigkeiten $\geq 96\%$ führt.
Dabei entsteht nicht nur ein Goldstandard des annotierten Korpus, das dem Songkorpus-Repositorium zur Verfügung gestellt wird, sondern auch eine angepasste Version des RNN-Taggers (verfügbar auf github), die für ähnliche Texte verwendet werden kann.

1 Einführung

In der Computerlinguistik ist die kaskadische Textverarbeitung seit Jahrzehnten eine übliche Methodik:[1] Der Text wird zuerst segmentiert (auch: tokenisiert), d.h. in ein Format gebracht, das ein Token pro Zeile zeigt, wobei mit dem Begriff Token entweder ein orthographisches Wort, eine Zahl oder ein Interpunktionszeichen gemeint ist. Alle Spatien werden üblicherweise dabei gelöscht; danach folgen weitere Verarbeitungsschritte, welche die resultierende Liste bzw. einspaltige Tabelle um weitere Spalten bzw. Annotationen auf Token-Ebene wie Wortart und Lemma ergänzt. Dadurch ergibt sich

[1] Zwar ist auch eine End-to-End-Verarbeitung möglich, die Tokenisierung, Tagging und Lemmatisierung in einem Schritt durchführt, bspw. mit einem Encoder-Decoder-Ansatz. Dies ist jedoch noch keine gängige Praxis.

eine dreispaltige Tabelle, die wiederum als Eingabedatei für weitere (z.B. syntaktische) Analysen oder auch direkt für korpuslinguistische Untersuchungen verwendet wird. Von manchen Segmentierungswerkzeugen werden auch Satzgrenzen erkannt und als strukturelle Merkmale des ursprünglichen Textes automatisiert eingefügt.

Diese kaskadische Verarbeitung ist sinnvoll, hat allerdings den Nachteil, dass sich Fehler in der Tokenisierung auf die weiteren Verarbeitungsschritte auswirken, denn was falsch tokenisiert ist, kann kaum korrekt mit einer Wortart annotiert werden. Hier zeigt sich, dass auftretende Interpunktion problematisch sein kann, denn sie ist in Teilen ambig, kann also mehrere Funktionen haben.

In diesem Artikel fokussieren wir uns auf den Apostroph, welcher in der uns vorliegenden Fassung der Texte einheitlich als gerader Apostroph (') auftritt. Wir vernachlässigen aber nicht ein weiteres Problem der Tokenisierung: Mehrwortlexeme und mehrteilige Namen, deren Bestandteile durch Bindestriche, aber auch durch Apostroph und/oder Leerzeichen getrennt werden.

Das Lindenberg-Korpus beinhaltet 2.827 Apostrophe (von 448.996 Zeichen insgesamt), die im Original teilweise als eigene Tokens, teilweise als Teil von fälschlicherweise als Tokens identifizierten Einheiten auftreten (z.B. *ich's*). Da wir eine möglichst hohe Korrektheit bei der späteren Wortartenannotation erzielen möchten, halten wir ein angepasstes Segmentierungsverfahren für notwendig.

Apostrophe haben in schriftlich konzipierten deutschen Standardtexten mehrere Einsatzmöglichkeiten, wie in (a) bis (c) geschildert:

(a) Zitate eingrenzen (*Sie bezeichnete den Artikel als 'groben Unfug'.*)
(b) Den Genitiv kennzeichnen bei Wörtern, die auf s enden (*Sokrates' Äußerung, ...*)
(c) Das Abkürzen von Jahreszahlen (*Im Frühjahr '21 kam der Schnee.*)

In eher mündlich konzipierten Texten finden sich darüber hinaus öfter Verwendungen wie in (d) bis (h) beschrieben. Hierbei sind dialektale Einsatzmöglichkeiten nicht berücksichtigt.

(d) Ersetzen eines Schwa am Ende eines Wortes (*Ich hab' das auch gelesen, Lass' mich in Ruh'*)
(e) Ersetzen eines Schwa innerhalb eines Verbs (*Wir sind verlor'n*)
(f) Verkürzen eines Artikels nach einer Präposition, wobei die Wörter miteinander verbunden werden (*Gekauft für'n Apfel ...*)
(g) Verkürzen eines Artikels, abgetrennt vom vorherigen Wort (*... und 'n Ei.*)
(h) Verkürzen eines expletiven *es* und Verbindung mit dem vorstehenden Wort (*Ja, gibt's denn so was?*)

Kaum eine der aktuellen Publikationen über Tagsets und ihre Erweiterungen beschreibt, wie die konkrete Tokenisierung in solchen Fällen zu erfolgen hat. Im Artikel "What is a word, What is a sentence? Problems of Tokenization" hatten Grefenstette und Tapanainen (1994) zwar bereits mögliche Herausforderungen und verschiedene Lösungen zu den hier im Fokus stehenden Apostrophen beschrieben, aber kein generelles

Handhabungskonzept vorgelegt. Bartz et al. (2013) erwähnen zwar teilweise die o.g. Fälle, gehen aber in ihrer Beschreibung aufkommender Segmentierungsprobleme von Tokenisierern für internetbasierte Kommunikation nicht weiter auf diese speziellen Fälle ein. Sie schlagen allerdings auch Lösungsmöglichkeiten für die dadurch aufkommende Problematik der kaskadischen Verarbeitung vor, denen wir in diesem Beitrag teilweise folgen.

Besonders viele Vorkommen von Apostrophen in verschiedenen Anwendungsfällen finden wir in den Texten des Songwriters und Interpreten Udo Lindenberg, die wir als Teil aus dem Songkorpus (*songkorpus.de*) zur Verfügung gestellt bekamen und auf die wir uns als typisches Beispiel eines Nicht-Standardtextes fokussieren. Trotz vieler Angebote an Tokenisierungswerkzeugen fand sich keines, das diese Texte adäquat und einheitlich tokenisiert. Daher beschreiben wir hier, Bartz et al. (2013) folgend, unseren eigenen zum Teil regel- und wissensbasierten korrigierenden Ansatz der Vereinheitlichung und optimalen Vorbereitung von Lindenberg-Texten für das Tagging, das auch ein Post-Editing umfasst, wie auch in Zinsmeister et al. (2014, S. 4101) angedacht.

In den Abschnitten 3 und 4 beschreiben wir unsere Anpassungen der bereits im Songkorpus mit Wortarten annotierten Fassung der Texte, die sich unter anderem aus der neuen Tokenisierung ergeben. Ziel ist die Erstellung eines neuen Goldstandard-Korpus, mit dessen Hilfe ein Tagger trainiert werden kann. In Abschnitt 5 evaluieren wir drei verschiedene Versionen des RNNTaggers (Schmid, 2019):

1. Die online verfügbare Originalversion des deutschen RNNTaggers, die nur auf dem Tiger-Korpus (Brants et al., 2004) trainiert wurde;
2. eine Version, die auf dem Tiger-Korpus und dem Trainingsteil des ursprünglichen Lindenberg-Korpus trainiert wurde;
3. eine dritte Version, die auf dem Tiger-Korpus und dem Trainingsteil des von uns überarbeiteten Lindenberg-Korpus trainiert wurde.

Die Evaluationsergebnisse zeigen, dass sich die Qualität der Wortartannotation deutlich verbessert, wenn der Text optimal tokenisiert wird und der Tagger auf einem Trainingskorpus aus der Zieldomäne trainiert wird. Wir zeigen auf, dass Tokenisierung immer noch ein nicht-triviales Problem darstellt, das bei der computerlinguistischen Verarbeitung eine wichtige Rolle spielt.

Während das Originalkorpus nach unserer Zählung 91.223 Tokens, davon 1.991 als Eigennamen erkannte Tokens (918 Types) umfasst, beinhaltet der neue Goldstandard 89.791 Tokens, davon 1.901 als Eigennamen erkannte, z.T. mehrteilige Tokens (866 Types). Die geringere Anzahl an Tokens insgesamt im Goldstandard erklärt sich aus der Zusammenschreibung von Tokens mit Apostroph durch den neuen Tokenisierungsansatz, erläutert in Tabelle 2.

2 Tokenisierung: Besonderheiten

2.1 Interpunktion

Liedtexte werden oft zumindest teilweise ohne Interpunktion verfasst. In ihrer üblicherweise vorliegenden Form zeigen Zeilenumbrüche stattdessen an, dass ein (Teil-)Satz endet. Darauf begründet beschreibt Schneider (2020, S. 843) eine TEI P5-konforme Annotation der gegebenen Liedzeilen, die es ermöglicht, Satzgrenzen zu erkennen. Die Lindenberg-Texte wurden entsprechend aufbereitet und liegen im songkorpus-Repositorium mit z.T. ergänzter Interpunktion vor.

Punkt-Disambiguierung ist, wie bereits in Grefenstette und Tapanainen (1994) geschildert, immer noch ein Problem: So erkennen einige aktuell verfügbare Tokenisierer wie z.B. der ASV Segmentizer[2] Punkte, die eigentlich Teil von Abkürzungen sind (wie in *Dr.* oder *o.k.*) nicht als solche, trennen diese ab und verhindern so eine korrekte Annotation.

Dazu treten in den vorhandenen Texten Sequenzen von zwei oder drei Punkten (oder auch das Tripledot-Zeichen …) auf, manchmal gefolgt von einem Satzendepunkt, manchmal nicht. Falls kein Satzendepunkt folgt, sollte man davon ausgehen, dass das Tripledot-Zeichen die Satzgrenze darstellt, es muss also vom vorgehenden Wort getrennt werden, auch wenn es diesem ohne Leerzeichen folgt. Auch für diese spezifische Verwendung empfiehlt sich ein regelbasiertes Post-Editing der tokenisierten Fassung, weil die Anzahl der Vorkommen im Korpus zu gering ist, um eine Maschine das Phänomen mit ausreichender Qualität lernen zu lassen.

2.2 Apostrophe als Herausforderung für die Tokenisierung

Die nachfolgenden Beispiele (1) bis (11) aus dem Lindenberg-Korpus verdeutlichen die dortige Verwendung von Apostrophen, die wir in Tabelle 1 tiefer analysieren bzw. kategorisieren werden. Zur Fokussierung wurden die Beispiele teilweise auf das Wesentliche verkürzt. Wo immer möglich erfolgt eine Zuordnung zu den oben genannten bekannten Phänomenen (a) bis (h).

1. *Die kämpfen für's ewige Gestern.* (f)
2. *Für 'ne Woche nach Wien.* (g)
3. *Die Frau schreibt'n läppischen Abschiedsbrief.*
4. *Ich baute 'ne Mauer um mein Herz.* (g)
5. *Auf der Erde gibt's sowas noch.* (h)
6. *Ich denk' immer nur an dich.* (d)
7. *Was sonst so um mich 'rum passiert.*
8. *Da kommt man nicht so ohne weit'res rein.* (e)
9. *Es sind des Haifisch's Flossen rot.*
10. *Im Sommer '84 kommen wir vorbei.* (c)
11. *Einen Groupie hab'n die auch.*

[2] verfügbar auf `https://weblicht.sfs.uni-tuebingen.de`.

Wir können die aufgeführten Beispiele in Tabelle 1 klar in 11 Kategorien verorten. Spalte 2 in Tabelle 1 zeigt auf, dass Segmentierung nicht unabhängig von der anschließenden Annotation mit Wortarten betrachtet werden kann. Schließlich können wir keine Tokens definieren, für die das Tagset im nächsten Verarbeitungsschritt keine Annotierungsmöglichkeit bietet. Wir müssen uns auch wie im Fall 9 daran halten, dass wir, den Grundsätzen der Korpuslinguistik folgend, auch leicht abweichende Varianten von der Standardschreibweise prozessieren sollten, wollen also auch Fälle wie *Haifisch's* verarbeiten, auch wenn eigentlich *Haifisches* bzw. *Haifischs* erwartet werden würde.

Wir verwenden bei unserem Vorgehen das ursprüngliche STTS-Tagset 1.0 (Schiller, Teufel & Stöckert, 1999) vor allem, weil wir in den vorliegenden Texten keine Fälle sehen, die eine Erweiterung benötigen (siehe auch Abschnitt 3.1). Dazu gibt es bis heute wenige Tagger, die für spätere Fassungen dieses Tagsets trainiert wurden. Auch das zugrundeliegende Korpus verwendet diese Annotation.

Nr.	Kategorie	Beschreibung	Beispiel
1	APPR'ART	Ein verkürzter Artikel folgt APPR ohne Leerzeichen	... *für's ewige* ...
2	APPR_'ART	Ein verkürzter Artikel folgt APPR nach einem Leerzeichen	... *für 'ne* ...
3	^APPR'ART	Ein gekürzter Artikel folgt einer anderen Wortart als APPR ohne Leerzeichen (wobei das Zeichen ^ negiert)	... *schreibt'n* ...
4	^APPR_'ART	Ein gekürzter Artikel folgt einer anderen Wortart als APPR nach einem Leerzeichen	... *baute 'ne* ...
5	POS'PPER	Ein gekürztes Pronomen (*es*) folgt einer beliebigen Wortart ohne Leerzeichen	... *gibt's* ...
6	POS'	Verkürztes Wort (Schwa in Endstellung oder Endung gelöscht)	... *denk'* ..., ... *is'* ...
7	'POS	Verkürztes Wort (meist Silbe *he-* aus ADV gelöscht)	... *mich 'rum* ...
8	^VTeil1'Teil2	Verkürztes Wort (kein Verb, Schwa im Wort gelöscht)	... *weit'res* ...
9	GEN's	Verkürzte Genitiv-Form	... *Haifisch's* ...
10	'YEAR	Verkürzte Jahreszahlen	*'84*
11	V'n	Verkürzte Verbform (Schwa im Wort gelöscht)	... *hab'n* ...

Tabelle 1: Kategorisierung von Apostrophenvorkommen in Lindenberg-Texten

2.3 Mehrwortlexeme, die Apostrophe, Bindestriche und/oder Leerzeichen enthalten

Beim Auftreten von Komposita, deren Bestandteile durch Bindestriche getrennt werden (aus den Lindenberg-Texten: *Schlager-Fuzzi, Gerade-aus-Maus, Schicki-Micki-Mode-Huhn, 08/15-Casanovas*), können sich Tokenisierungs- bzw. Taggingprobleme ergeben, wie der ASV-Tokenisierer zeigt, der im Wort *U-Boot* drei Tokens erkennt. Bekannt ist das außerdem bereits für Komposita, deren Bestandteile durch Leerzeichen getrennt sind (*Apollo 13, Rolling Stones, St. Tropez*) bzw. Kombinationen aus beiden (*Never Come Back-Airline, D-471 81 61*). Zuletzt – und hier gibt es bei den Songtexten besonders viele Fälle – wird die Bezeichnung *Rock'n'Roll* (auch *Rock'n Roll* oder *Rock 'n' Roll*) in den Texten Lindenbergs nicht nur auf verschiedenste Weisen geschrieben, sondern auch mit weiteren Wörtern ergänzt, wodurch neue Komposita gebildet werden (*Rock'n'Roll-Gespenster, Rock 'n' Roll Zigeuner*).

Man kann von einem Tokenisierer nicht erwarten, dass er diese Schreibweisen alle voneinander unterscheiden lernt und Namen oder englische Abkürzungen zweifelsfrei identifizieren kann. Daher werden Leerzeichen üblicherweise beim Tokenisierungsvorgang gelöscht, die Bestandteile der Namen voneinander getrennt und erst in einem weiteren Verarbeitungsschritt (der sog. *Named Entity Recognition*) wieder zusammengefügt. Ähnlich gehen wir vor, im Einzelnen beheben wir das Problem wie folgt:

1. Wir tokenisieren den Text zuerst regulär mit dem Tokenisierer ohne externe Wissensbasis.
2. Wir annotieren die Tokenliste mithilfe eines passenden Taggers bzw. lassen mit einem entsprechenden Werkzeug Namen erkennen.
3. Die so erkannten Namen werden von einem Skript gesammelt und in eine Liste geschrieben.
4. Die Liste der mehrteiligen Namen wird von Hand überprüft und korrigiert.
5. Anschließend wird der Text unter Berücksichtigung der erstellten Liste vom Tokenisierer erneut tokenisiert und steht dem Tagger korrigiert zur Verfügung.

Der von uns verwendete RNN-Tagger (Schmid, 2019) konnte auch ohne besonderes Training diese mehrteiligen Tokens zuverlässig als Eigennamen erkennen. Die Erweiterung der Eigennamen-Liste des Tokenizers um die neuen Eigennamen ist derzeit noch nicht automatisiert und muss ggf. für andere Korpora erneut durchgeführt werden.

3 Tokenisierung und anschließende Annotation mit Wortarten (Tagging)

3.1 Analyse der Fälle 1 bis 5 nach Tabelle 1

Bartz et al. (2013) schlagen Wortartannotationen von umgangssprachlich kontrahierten Formen (dort: Phänomentyp III.2) vor. Sie berücksichtigen allerdings nicht die hier vorliegende vielfältige Verwendung von Apostrophen. Wenn wir z.B. “APPR'ART” (wie in “für'n”) als ein Token erkennen und dieses mit KTRAPPRART annotieren würden, wäre keine Gleichbehandlung mit “APPR 'ART” (mit trennendem Leerzeichen) möglich.

Wir gehen nicht davon aus, dass die Verwendung des Apostrophs bzw. des Leerzeichens in den Liedtexten andere als phonetische Gründe hat und entscheiden uns dafür, bei der Tokenisierung alle mit einem Apostroph verkürzten Artikel einheitlich zu behandeln, sie also als eigenständige Tokens zu segmentieren und mit dem Tag "ART" zu annotieren. Im Test mit verschiedenen Tokenizern, die in WebLicht (Hinrichs, Zastrow & Hinrichs, 2010) angeboten werden, zeigt sich, dass auch diese eine Abtrennung des Artikels in fast allen Fällen korrekt durchführen (die Ergebnisse finden sich in Tabelle 3).

Verkürzte Pronomina (Fall 5), die auf andere Wörter folgen, werden von uns genauso behandelt wie verkürzte Artikel: sie sind eigenständige Tokens und werden ggf. von dem vorhergehenden Token abgetrennt.

3.2 Analyse der Fälle 6-10 nach Tabelle 1

Die Fälle 6 bis 11 beschreiben verkürzte Wortformen. Hier darf der Wortteil ab dem Apostroph keinesfalls abgetrennt werden und auch der Apostroph muss am bzw. im Wort verbleiben. Die meisten Tokenizer aus WebLicht (siehe Abschnitt 4.2) erkennen übrigens Tokens, die mithilfe von Apostrophen intern verkürzt werden, korrekt.

3.3 Analyse des Falls 11 nach Tabelle 1

Fall 11 ist problematisch. In der ambig erscheinenden Form *hab'n* könnten zwei Verkürzungen beinhaltet sein: *hab 'n* (*habe ein/einen*), es könnte sich aber auch um ein einfaches, intern verkürztes Verb handeln: *haben*. Eine Einzelfallentscheidung muss der Tokenisierer treffen, eventuell muss diese Entscheidung jedoch kontextabhängig im Rahmen eines manuellen Post-Editing korrigiert werden.

3.4 Zusammenfassung

Tabelle 2 zeigt unsere Wunsch-Tokenisierung der mit Apostroph versehenen Tokens, die sich aus der bisherigen Argumentation ergibt. Fälle 11A und 11B werden im Rahmen des Posteditings manuell unterschieden, bevor die Wortartannotation durchgeführt wird.

4 Vorgehen bei der Tokenisierung

4.1 Regelbasierte Anpassung der bestehenden Tokenisierung

Für die weitere Verarbeitung wurden uns die von Schneider (2020, S. 843-844) beschriebenen Liedtexte von Udo Lindenberg mit z.T. eingefügten Satzzeichen (Kommata, Punkte und Fragezeichen) untokenisiert sowie tokenisiert und mit Wortarten sowie Lemmata annotiert zur Verfügung gestellt. So hatten wir die Möglichkeit, weitere Tokenizer anzuwenden und deren Ergebnis mit der bereits bestehenden Fassung zu vergleichen. Allerdings fanden sich in den bereitgestellten Dateien auch für uns nicht akzeptable Tokenisierungen, die wir regelbasiert mithilfe von Skripten sowie - in ambigen Fällen

Nr.	Kategorie	Token1	Token2
1	APPR'ART	*für*	*'s*
2	APPR 'ART	*mit*	*'ner*
3	VVFIN'ART	*schreibt*	*'n*
4	VVFIN'ART	*baute*	*'ne*
5	PWAV'PPER	*wo*	*'s*
6	VVFIN'	*denk'*	
7	POSS'ADV	*mich*	*'rum*
8	gekürztes ADJ	*weit'res*	
9	Genitiv's	*Haifisch's*	
10	'Jahr	*'84*	
11A	V'n	*hab'n*	
11B	V'n	*hab*	*'n*

Tabelle 2: Default-Tokenisierung der Apostrophvorkommen in Lindenberg-Texten

- manuell angepasst haben. Dadurch entstand ein Goldstandard, der für die weiteren Verarbeitungsschritte verwendet wurde.

Um die Evaluation der in WebLicht vorhandenen Tokenisierer zu vereinfachen, wurden Sätze mit den als problematisch bekannten Phänomenen aus dem Goldstandard extrahiert und diese Testsuite (siehe Tabelle 5 im Anhang 1) in WebLicht geladen. Diese Testsuite umfasst alle oben beschriebenen Phänomene. Unsere Default-Tokenisierung umfasste 457 Tokens mit insgesamt 68 Fällen, überwiegend Apostroph-Einsätze, aber auch Komposita und Tripledots-Verwendungen. Alle Tokenizer in WebLicht (siehe Tabelle 6 in Anlage 2) erkennen jedoch eine höhere Anzahl, wobei der ASV Tokenizer mit den meisten (555) Tokens alle Apostrophe abtrennt. Dies ist sicherlich der Tatsache geschuldet, dass das Werkzeug eigentlich als Satz-Segmentierer eingesetzt werden sollte und nur eine einfache Tokenisierungkomponente enthält[3]. Allerdings wird er augenscheinlich gleichwertig zu den weiteren Tokenizern in WebLicht zur Verwendung angeboten. Auch die anderen Phänomene werden von diesem Tokenizer nicht identisch zu unserem Default prozessiert. Sehr problematisch sind Segmentierungen wie z.B. die Sequenz *geseh* | ' | *n* (die senkrechten Striche symbolisieren die erkannten Tokengrenzen). Damit werden auch nicht mit eine Wortart annotierbare Tokens (Einzelbuchstaben) erkannt. Der ASV-Segmentierer wird daher als nicht geeignet angesehen, diese Art von Texten zu prozessieren.

Die beiden SFS-Tokenizer haben alle für uns relevante Phänomene gleich gehandhabt und zum Beispiel verkürzte Verben sehr gut erkannt, dafür jedoch Artikel nicht von vorstehenden Präpositionen abgetrennt. Sie trennen außerdem abkürzende Apostrophe

[3]Siehe Erläuterungen auf `https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Tools_in_Detail#Tokenizers`.

Ergebnis	SoJaMo	sfs-OpenNLP	sfs-Tübingen	BBAW	Blingfire	ASV-Segmentation
wie Default	42	47	47	56	11	0
nicht wie Default	26	21	21	12	57	68

Tabelle 3: Ergebnisse der WebLicht-Tokenizer

am Anfang von Sätzen wenn Großschreibung folgt, wie in *'Ne dunkle Wand*. Fast dieselben Abweichungen zeigen sich auch bei SoJaMo. Blingfire erkannte außerdem Komposita wie *U-Boot* nicht. Tabelle 3 zeigt die Zahl der aufgetretenen Abweichungen aufsummiert[4].

4.2 Tokenisierung mit statistischer Desambiguierung

Da die vorhandenen Tokenisierer die Songtexte noch nicht zufriedenstellend tokenisieren konnten, haben wir einen neuen Tokenisierer mit einer statistischen Desambiguierungskomponente entwickelt, der auf Ideen in (Schmid, 2000) basiert. Der Tokenisierer verarbeitet seine Eingabe in folgenden Schritten:

- Ersetzung von (Folgen von) Leerzeichen, Tabulatoren und anderen "Whitespace"-Symbolen durch Tokengrenzen
- Abtrennung von Satzzeichen, Klammern und Anführungszeichen am Beginn und Ende jedes Tokens als separate Tokens
- statistische Desambiguierung zwischen Satzpunkten, Abkürzungspunkten und Ordinalzahlpunkten wie in (Schmid, 2000) beschrieben
- statistische Desambiguierung von Apostrophen wie unten beschrieben
- Erkennung von Mehrwort-Einheiten durch eine einfache Longest-Match-Suche mit einer gegebenen Liste von Mehrwortausdrücken.

4.2.1 Statistische Desambiguierung von Apostrophen

Wenn ein Apostroph innerhalb eines Tokens auftritt, betrachten wir folgende Möglichkeiten (wobei der Unterstrich für ein Leerzeichen steht):

- Der Apostroph ersetzt "e" wie in *ander'n*.
- Der Apostroph ersetzt "e_" wie in *möcht'so*.
- Der Apostroph ersetzt "_e" wie in *gibt's*.
- Der Apostroph ersetzt "ei" wie in *irgend'nem*.
- Der Apostroph ersetzt "_ei" wie in *für'n Ei*.
- Der Apostroph ersetzt "_eine" wie in *für'n Apfel*.

[4] Dies liegt darin begründet, dass eine komplette Übersicht der Einzelergebnisse den Rahmen dieses Beitrags sprengen würde (gedruckt umfasst sie 18 Seiten). Die Einzelergebnisse sind jedoch bei den AutorInnen dieses Beitrags verfügbar

- Der Apostroph ersetzt "_de" wie in *Er ist über'n Berg.*
- Der Apostroph ersetzt "_da" wie in *über's Ziel.*
- Der Apostroph ersetzt "e_e" wie in *hab's.*
- Der Apostroph ersetzt nichts wie in *Smog'n'Roll.*[5]

Wir desambiguieren zwischen diesen Ersetzungsoperationen, indem wir den Apostroph im Token t nacheinander durch jede der obigen Zeichenfolgen – inklusive den Apostroph selbst – ersetzen und die Zeichenfolge r mit der höchsten Aposteriori-Wahrscheinlichkeit $p(r|t)$ laut folgender Gleichung auswählen:

$$p(r|t) = \frac{p_{lm}(t/r)\; p_{op}(r \rightarrow \text{'})}{\sum_s p_{lm}(t/s)\; p_{op}(s \rightarrow \text{'})} \tag{1}$$

Dabei bezeichnet t/r das Ergebnis der Ersetzung des Apostrophs im Token t durch r. Bspw. ergibt *hab's/e_e* den String *habe_es*. Das Sprachmodell $p_{lm}(t/r)$ liefert die Wahrscheinlichkeit des Wortes (bzw. Wortpaares) t/r, und $p_{op}(r \rightarrow \text{'})$ liefert die Wahrscheinlichkeit dafür, dass der Teilstring r in t/r durch den Apostroph ersetzt wird (statt unverändert zu bleiben).[6]

Für das Token *schreib'n* bekommen wir unter anderem die möglichen Ersetzungen *"schreiben"* und *"schreib ein"* mit den Bewertungen p_{lm}(schreiben) p_{op}(e $\rightarrow$ ') bzw. p_{lm}(schreib ein) p_{op}(_ei $\rightarrow$ ').

Wenn die wahrscheinlichste Ersetzungsoperation $\hat{r} = \arg\max_r p(r|t)$ mit einem Leerzeichen beginnt (bspw. _ei), fügen wir vor dem Apostroph eine Tokengrenze hinzu (bspw. *für 'n*). Wenn die wahrscheinlichste Ersetzungsoperation mit einem Leerzeichen endet (e_), fügen wir nach dem Apostroph eine Tokengrenze hinzu (*möcht' so*). Im Fall der Ersetzung e_e verdoppeln wir den Apostroph und fügen dazwischen eine Tokengrenze ein (*hab' 's*). In allen anderen Fällen bleibt das Token unverändert.

4.2.2 Sprachmodell

Für das Sprachmodell $p_{lm}(.)$ verwenden wir ein wortbasiertes Markowmodell erster Ordnung (Bigramm-Modell), welches nicht zwischen groß- und kleingeschriebenen Buchstaben unterscheidet. Das Bigramm-Modell wird mit interpolierter Kneser-Ney-Glättung "on-the-fly" auf den zu tokenisierenden Daten (ohne Verwendung der Goldstandard-Annotation) trainiert. Daher ist dieses Verfahren weniger geeignet, um kurze Texte zu tokenisieren. Kurze Texte können jedoch mit längeren Textstücken gleicher Art zusammengefasst werden, um bessere Ergebnisse zu erzielen.

[5] Man kann hier natürlich argumentieren, dass es sich im Englischen durchaus um eine Verkürzung von *and* zu *'n'* handelt. Mit fremdsprachlichen Verkürzungen befassen wir uns jedoch im Rahmen dieser Arbeit nicht.

[6] $p_{op}(r \rightarrow \text{'})$ ist keine Wahrscheinlichkeitsverteilung über die verschiedenen Ersetzungen r. Es gilt also in der Regel: $\sum_r p_{op}(r \rightarrow \text{'}) \neq 1$. Unser statistisches Modell ist ein Noisy-Channel-Modell, welches erst mit dem Sprachmodell p_{lm} eine Phrase t/r generiert (bspw. **habe_es**) und dann zufällig mit Wahrscheinlichkeit $p_{op}(r \rightarrow \text{'})$ entscheidet, ob der Teilstring r in t (hier: e_e durch einen Apostroph ersetzt (und damit "verrauscht") wird oder nicht. Bei der Desambiguierung versuchen wir zu rekonstruieren, wie die unverrauschte Zeichenfolge wahrscheinlich aussah.

4.2.3 Training

Unser statistisches Modell umfasst die Parameter $p_{lm}(.)$ und $p_{op}(.)$ und wird mit dem Expectation-Maximization-Algorithmus (EM-Algorithmus) trainiert. Dazu werden zunächst die zu verarbeitenden Texte mit dem Tokenisierer von Schmid (2000) vorläufig tokenisiert. Auf den tokenisierten Texten wird dann das Sprachmodell $p_{lm}(.)$ trainiert, um es zu initialisieren. Die Wahrscheinlichkeiten der verschiedenen Ersetzungsoperationen $p_{op}(.)$ werden mit 0,01 initialisiert mit der Ausnahme $p_{op}(\text{'} \rightarrow \text{'}) = 0,0001$. Diese Ausnahme dient dazu, (echte) Ersetzungen zunächst zu präferieren.[7] Dann führen wir zwei[8] EM-Iterationen durch. In jeder EM-Iteration berechnen wir zunächst nach Formel 1 für jedes apostrophierte Token t die Aposteriori-Wahrscheinlichkeit jeder möglichen Ersetzungs-Operation r.

Für jede Operation r summieren wir dann die Aposteriori-Wahrscheinlichkeiten $p(r|t)$ über alle apostrophierten Tokens t des Textes wie folgt:

$$f_{r,t} = F_t \, p(r|t)$$

F_t ist hier t's Häufigkeit im Text. Mit den so erhaltenen erwarteten Häufigkeiten schätzen wir die Ersetzungswahrscheinlichkeiten neu:

$$p_{op}(r \rightarrow \text{'}) = \frac{\sum_t f_{r,t}}{\sum_t f_{r,t} + F_{t/r}}$$

$F_{t/r}$ ist hier die Häufigkeit des Tokens (oder Token-Paares) t/r (bspw. *habe_es*) im Text.

Analog werden die Parameter des Sprachmodelles neugeschätzt. Für jedes Wortpaar (s,t) mit $t = t_1\text{'}t_2$, subtrahieren wir 1 von seiner Texthäufigkeit $h_{s,t}$ und addieren für jede mögliche Ersetzung r ohne Leerzeichen den Wert $p(r|t)$ zur Häufigkeit $h_{s,t_1 r t_2}$. Für jede Ersetzung $r = r_1\_r_2$ mit Leerzeichen addieren wir den Wert $p(r|t)$ zur Häufigkeit $h_{s,t_1 r_1}$ und zur Häufigkeit $h_{t_1 r_1, r_2 t_2}$. Dann schätzen wir die Parameter des Sprachmodelles mit der Kneser-Ney-Methode (Kneser & Ney, 1995) neu:

$$\begin{aligned} h_t &= \sum_s h_{s,t} \\ p(t) &= \frac{h_t - \delta_1}{\sum_{t'} h_{t'}} \\ p(t|s) &= \frac{h_{s,t} - \delta_2}{\sum_{t'} h_{s,t'}} \\ \alpha(s) &= 1 - \sum_t p(t|s) \end{aligned}$$

[7] Noch bessere Ergebnisse werden erzielt, wenn die Wahrscheinlichkeiten der Ersetzungsoperationen " da" und " eine" mit 1 initialisiert werden.

[8] Mehr EM-Iterationen haben zu etwas schlechteren Ergebnissen geführt.

$$
\begin{aligned}
k_t &= |\{(s,t)|h_{s,t} > 0\}| \quad \text{// Kneser-Ney-Methode} \\
p_{bo}(t) &= \frac{k_t}{\sum_{t'} k_{t'}} \\
p(s,t) &= p(s)\ (p(t|s) + \alpha(s)\ p_{bo}(t))
\end{aligned}
$$

δ_1 und δ_2 sind hier Discounts, die nach der Formel $N_1/(N_1 + 2N_2)$ berechnet werden, wobei N_i die Zahl der Wörter/Wortpaare mit Häufigkeit i ist. Discounting bei den Unigramm-Wahrscheinlichkeiten $p(t)$ hat sich in Experimenten als vorteilhaft erwiesen.

4.2.4 Evaluation

Wir evaluierten den statistischen Tokenisierer auf dem Goldstandard-Korpus, das 89.791 Tokens umfasst. Der Tokenisierer machte dabei insgesamt 52 Fehler, die sich in folgende Klassen einteilen lassen:

- In 11 Fällen wurde ein Satzpunkt nicht abgetrennt, weil der nachfolgende Satz mit einem Kleinbuchstaben begann.

  Beispiel: *... du bist verloren und ganz alleine hier oben. so 'n Gefühl, das hab' ich manchmal ...*
- In 8 Fällen wurde *'s* nicht abgetrennt, weil das vorausgehende Wort ausschließlich mit nachfolgendem *'s* auftrat.

  Beispiele: *solln's klappt's versuchen's*
- In 6 Fällen wurde *'n* nicht abgetrennt, weil der Apostroph auch für *e* hätte stehen können.

  Beispiele: *mal'n war'n wär'n echt'n*
- In 5 Fällen wurde *'s* nicht von dem Wort *krieg* abgetrennt, weil das Wort *Krieges* relativ häufig auftrat und Groß-/Kleinschreibung nicht unterschieden wird.
- In 5 Fällen wurde nach dem Ausdruck *D-471 81 61* der nachfolgende Satzpunkt nicht abgetrennt. Diese Fehler waren durch eine Lücke im Programmcode verursacht.
- In 4 Fällen wurden Tokenisierungsfehler durch fehlende Leerzeichen in der Eingabe verursacht.

  Beispiele: *... oder -denn wissen-*
- 3 Tokenisierungsfehler lassen sich auf fremdsprachliche Einsprengsel zurückführen: Der Ausdruck *Harry's* wurde fälschlich zerlegt und die Ausdrücke *C'est* und *That's* wurden fälschlich nicht zerlegt.
- In 3 Fällen wurde vor dem Apostroph abgetrennt, weil die ungekürzte Wortform nicht auftrat.

  Beispiele: *zwei'n seid'nen*
- Die restlichen 7 Fehler lassen sich keiner größeren Klasse zuordnen.

5 Tagging

Für die Wortart-Annotation und Lemmatisierung verwenden wir den RNNTagger (Schmid, 2019), der auf rekurrenten neuronalen Netzen basiert.

Der RNNTagger wurde für Deutsch und viele andere Sprachen trainiert und ist für nicht-kommerzielle Zwecke frei verfügbar[9]. Für unsere Evaluationen haben wir den Tagger auf unterschiedlichen Korpora neu trainiert und verglichen.

5.1 Evaluation

Für die Evaluation teilten wir das Lindenberg-Korpus in Trainingsdaten, Entwicklungsdaten und Testdaten auf. Wir wählten die ersten 75.189 Tokens als Trainingsdaten, die nächsten 7.269 Tokens als Entwicklungsdaten und die restlichen 7.333 Tokens als Testdaten.

Da bei der Erstellung des Goldstandard-Korpus einige Fehler in der Wortart-Annotation des Originalkorpus korrigiert wurden, haben wir diese Korrekturen auf das Originalkorpus übertragen, um faire Ergebnisse zu erhalten. In Fällen, wo ein apostrophiertes Token fälschlich nicht aufgespalten worden war, wurde das Token mit einem Doppeltag annotiert. In Fällen, in denen ein Token fälschlich in mehrere kleinere Tokens aufgespalten worden war, annotierten wir alle Tokens bis auf das erste mit dem neuen Tag *PART*.

Um den Einfluss der unterschiedlichen Tokenisierungen auf die Taggingergebnisse zu untersuchen, führten wir Experimente durch, bei denen wir den RNNTagger entweder nur auf dem Tigerkorpus (Tiger) oder zusätzlich auf dem Trainingsteil des Lindenberg-Korpus mit der originalen (Tiger+original) oder der verbesserten Tokenisierung des Goldstandards (Tiger+goldstandard) trainierten. Jede Tagger-Version wurde viermal mit den Standard-Hyperparametern und unterschiedlichen Startwerten des Zufallszahlengenerators trainiert. Nach jeder Trainingsepoche wurde das Taggermodell auf den Entwicklungsdaten evaluiert. Das Modell aus der besten Trainingsiteration wurde jeweils auf den Testdaten evaluiert. Für jede Taggervariante gab es somit 4 Evaluationsergebnisse. Die Tokenisierung der Trainings-, Entwicklungs- und Testdaten war in jedem Experiment einheitlich, d.h. entweder immer die Originaltokenisierung oder immer die neue Goldstandard-Tokenisierung.

Tabelle 4 zeigt die erzielten Genauigkeiten. Der Tagger, der nur auf dem Tiger-Korpus trainiert wurde, erzielte mit der Goldstandard-Tokenisierung im Mittel deutlich bessere Ergebnisse (+0.6%) als mit der Original-Tokenisierung. Zusätzliches Training auf dem Lindenberg-Korpus verbesserte die Ergebnisse sogar um +3,0% mit der Goldstandard-Tokenisierung (Tiger+goldst.) und um +3,7% mit der Original-Tokenisierung (Tiger+original).

Überraschenderweise erzielte der Tagger, der zusätzlich auf dem Korpus mit der originalen Tokenisierung trainiert wurde (Tiger+original), ebenso gute Ergebnisse wie der Tagger, der zusätzlich auf den Goldstandard-Daten trainiert wurde (Tiger+goldst.). Der

[9] `https://www.cis.lmu.de/~schmid/tools/RNNTagger`

Trainingsdaten	Entwicklungsd.	Testdaten	Genauigkeit in % (Mittel, Max., Min., Stdabw.)			
Tiger	original	original	92,58	92,80	92,44	0,18
Tiger	goldst.	goldst.	93,17	93,33	92,96	0,15
Tiger+original	original	original	96,24	96,33	96,15	0,09
Tiger+goldst.	goldst.	goldst.	96,15	96,37	96,03	0,15

Tabelle 4: Genauigkeit der verschiedenen Tagger-Varianten

Genauigkeitsunterschied von 0,09% war statistisch nicht signifikant.[10] Der RNNTagger ist also in der Lage, auch falsch tokenisierte Texte korrekt zu annotieren, wenn er die Problemfälle im Training kennenlernen konnte.

Eine weitere, allerdings weniger umfassende Evaluation haben wir mithilfe der Songtexte der Band *Fettes Brot* durchgeführt, um zu prüfen, inwieweit sich Verbesserungen durch die Verwendung der neu entwickelten Segmentierung – allerdings ohne manuelle Eingriffe für die Namenserkennung – ergeben. Im Original-Korpus ist diese Sammlung mit 59.269 Tokens verzeichnet, sie beinhaltet 1.309 Apostrophe. Das unveränderte Tokenisierungswerkzeug aus (Schmid, 2019) erzeugt 78.360 Tokens, das hier neu entwickelte 78.028 Tokens. Bereits mit der bisherigen Version des Segmentierers (Schmid, 2019) werden die folgenden Phänomene korrekt erkannt:

- Löschen des Schwa in oder am Ende eines Worts, wie in *hör', hör'n, Fernseh'n*
- Abtrennen der gekürzten unbestimmten Artikel *ein, einer, einem, einen* wie in *'n Knall, 'ner Weile, aus'm Häuschen, durch 'en Reifen*

Die neue Version des Segmentierers erkennt zusätzlich die folgenden Phänomene

- Abtrennen des Personalpronomens / des explikativen *es*, wie in *gibt's, geht's, habt's*: 222 Vorkommen
- Abtrennen des gekürzten unbestimmten Artikels *ein* wie in *und'n Kind* : 78
- Abtrennen des gekürzten bestimmen Dativartikels *dem* wie in *auf'm Dreier*

Satz (12) unten veranschaulicht, wie der Tokenisierer in manchen Fällen den Apostroph verdoppelt (12a). In Satz (13) ist uns die Verwendung des ersten *'n* zwar nicht ganz klar[11], das neue Werkzeug erkennt jedoch zumindest den Artikel korrekt als eigenständiges Token.

12 *Und würd's nochmal von vorne losgehen, wärst du wieder dabei.*
12a *Und würd' **'s** nochmal von vorne losgehen, wärst du wieder dabei.*
13 *..., was'n das für'n Name?*
13a *..., was **'n** das für **'n** Name?*

[10] Der t-Test lieferte einen p-Wert von 0.37.
[11] Bei dem Ausdruck *was'n* könnte es sich um eine Verkürzung des Ausdruckes *was ist denn* handeln.

6 Mögliche weitere Arbeiten

Wir zeigen in diesem Beitrag auf, dass die Texte Udo Lindenbergs eine besondere computerlinguistische Prozessierung benötigen, da insbesondere der Apostroph dort ambig auftritt. Segmentierungswerkzeuge, die von WebLicht derzeit zur Verfügung gestellt werden und die auf Standard-Texten trainiert wurden, können diese Texte nur eingeschränkt und z.T. überhaupt nicht handhaben.

Es ist also durchaus noch nötig, aber selbst bei diesem eher kleinen Korpus auch möglich, Segmentierungswerkzeuge spezifisch für diese Textart zu entwickeln unter Berücksichtigung des später anzuwendenden Tagsets für die Wortartenannotierung. Wir haben ein solches Werkzeug vorgestellt, welches mit dem EM-Algorithmus direkt auf den zu verarbeitenden Texten trainiert wird und Apostrophe recht zuverlässig desambiguieren kann. Mit unserem Segmentierungswerkzeug gehen wir weiter als in (Schiller et al., 1999) beschrieben und ermöglichen es, mehrteilige Namen bereits bei der Tokenisierung als eigenständige Tokens zu identifizieren und sie damit während der weiteren Prozessierung mit den richtigen Wortart-Tags zu annotieren. Die erstellten Werkzeuge sind für andere Anwendungen frei unter `https://github.com/helgihu/German-Song-Tagger` verfügbar.

Wir erwarten, dass die vorgestellte Verarbeitungspipeline auch auf das Songkorpus als Ganzes anwendbar ist, sofern nicht unvorhergesehene neue Problemstellungen in Erscheinung treten. Die Tokenisierung des Gesamtkorpus kann in derselben Weise erfolgen wie beim Lindenberg-Teilkorpus. Wenn auf die Identifizierung noch unbekannter mehrteiliger Namen verzichtet werden kann, sind hier keine manuellen Eingriffe erforderlich. Wegen der Ähnlichkeit der Textsorten sollte die Tagging-Genauigkeit auch bei den Songtexten anderer Interpreten von dem zusätzlichen Training auf den manuell annotierten Lindenbergtexten profitieren, auch wenn dort die Steigerung der Genauigkeit etwas kleiner ausfallen dürfte.

Unsere Evaluation auf den Songtexten der Band *Fettes Brot* zeigte, dass sich die Behandlung von Apostrophen mit dem neuen Werkzeug eindeutig verbessert hat. Weitere, detailliertere Tests sowie die Erstellung eines Goldstandards wären allerdings nötig, um eine vollständige Evaluation dieser und andere Songtexte durchzuführen.

Literatur

Bartz, T., Beißwenger, M. & Storrer, A. (2013). Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *Journal for Language Technology and Computational Linguistics (JLCL)*, *28* (1), 157–198.

Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., ... Uszkoreit, H. (2004). Tiger: Linguistic interpretation of a german corpus. *Journal of Language and Computation*, *2*, 597-620.

Grefenstette, G. & Tapanainen, P. (1994). What is a word, what is a sentence? problems of tokenization. In *Proceedings of the 3rd International Conference on Computational Lexicography* (S. 79–87).

Hinrichs, M., Zastrow, T. & Hinrichs, E. (2010). WebLicht: Web-based LRT Services in a Distributional eScience Infrastructure. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)* (S. 489–493).

Kneser, R. & Ney, H. (1995). Improved backing-off for m-gram language modeling. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Bd. 1, S. 181–184).

Schiller, A., Teufel, S. & Stöckert, C. (1999). *Guidelines für das Tagging deutscher Textkorpora mit STTS.* Zugriff am 2022-08-29 auf `https://www.ims.uni-stuttgart.de/documents/ressourcen/lexika/tagsets/stts-1999.pdf`

Schmid, H. (2000). *Unsupervised learning of period disambiguation for tokenisation.* Zugriff am 2022-10-19 auf `https://www.cis.lmu.de/~schmid/papers/tokeniser.pdf`

Schmid, H. (2019, May). Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH).*

Schneider, R. (2020). A Corpus Linguistic Perspective on Contemporary German Pop Lyrics with the Multi-Layer Annotated « Songkorpus ». In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)* (S. 842–848).

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)* (S. 4097–4104).

Korrespondenzanschrift

Gertrud Faaß
Universität Hildesheim
Institut für Informationswissenschaft und Sprachtechnologie
gertrud.faass@uni-hildesheim.de

Helmut Schmid
Ludwig-Maximilians Universität München,
Institut für Informations- und Sprachverarbeitung
schmid@cis.lmu.de

7 Anhang 1

Testsuite für die WebLicht-Tokenizer

Nr.	Kategorie	TSNr.	Testsatz (TS)
1	APPR'ART	1	Die kämpfen für's ewige Gestern.
		2	Ich war nie der Typ für'n Liebesfilm.
		3	Mit'm U-Boot fahren wollte ich schon immer mal.
		4	Ich sitze an der Bar mit'nem Drink.
		5	Und die tanzt auf'm Tisch wie'n Gogogo-Girl.
2	APPR_'ART	6	Für 'ne Woche nach Wien.
		7	Mit 'nem Star oder mit 'nem Statist.
		8	Und hab was mit 'ner anderen Frau.
3	^APPR'ART	9	Dass es viel mehr war als nur so'n kleiner Flirt am Rand.
		10	Die Frau schreibt'n läppischen Abschiedsbrief.
		11	Au ja, 'n Teenie mit Doktortitel lila Strapse und drüber'n weißer Kittel
4	^APPR_'ART	12	Dann reißt er noch 'nen blöden Witz.
		13	Es geht doch hier nicht um 'ne schnelle sexuelle Nacht.
		14	Ich kenn' 'ne Lady.
		15	. 'Ne dunkle Wand und ein Riss geht durch die Zeit
		16	Au ja, 'n Teenie mit Doktortitel lila Strapse und drüber'n weißer Kittel
		17	Er wär 'n Astronaut.
		18	Ich baute 'ne Mauer um mein Herz.
		19	In der der Hand 'ne Cognacflasche und 'n Autogramm von Klaus Kinski.
		20	Doch er ist so 'ne Art Traum-Dompteur.
		21	Aus irgend 'nem blöden Grund rasen die aneinander vorbei.
		22	Als nur 'ne schnelle Romanze am Strand.
		23	Ist alles erst 'n paar Stunden her.

		24	Ich fuhr mit dem Auto n' bißchen zu schnelle.
5	POS'PPER	25	Auf der Erde gibt's doch sowas noch.
		26	Wo's dauernd auf die Fresse gibt.
		27	Wie's lächerlicher nicht mehr geht.
		28	Alles im Lot und wenn's untergeht.
		29	Und da gibts's auch Wahnsinnsfrauen.
6	POS'	30	Ich denk' immer nur an dich.
		31	Hab' nicht gewusst, dass das alles so stark ist.
		32	Bis ich absolut keine Luft mehr hab'.
		33	Denn da wär' beinah ein Schiffsunglück passiert.
		34	Oder wenn ich durchdreh', die letzten Scheine auf den Tresen hau'.
		35	Das lässt mich einfach nicht mehr in Ruh'.
		36	Ich hätt' dich beinah' nicht geseh'n.
7	'POS	37	Haben wir uns ein Glas Nost 'reingeknallt.
		38	Was sonst so um mich 'rum passiert.
8	V'n	39	Einen Groupie hab'n die auch.
		40	Ich muss Dich wiederseh'n.
		41	Und da darf keiner zwischensteh'n.
		42	Kannst' auch mal abschmier'n.
9	^V'n	43	Ich hab was mit einer ander'n Frau.
		44	Da kommt man nicht so ohne weit'res rein.
		45	Die krabbeln durch die seid'nen Betten.
10	GEN's	46	Es sind des Haifisch's Flossen rot.
11	'YEAR	47	Im Sommer '84 kommen wir vorbei.

Tabelle 5: Die Testsuite: z.T. verkürzte Sätze aus dem Lindenberg-Korpus

8 Anhang 1

WebLicht: Tokenizer
Informationen aus `https://weblicht.sfs.uni-tuebingen.de`

Tokenizer	Attribute	Description
SFS Tübingen	Desc:	Tokenizer/sentences from the OpenNLP project. The 'newlineBounds'
	parameter	treats newlines as a hard break (a sentence boundary).
	Creator(s):	SfS: Uni-Tuebingen
	Contact:	wlsupport@sfs.uni-tuebingen.de
	PID:	`http://hdl.handle.net/11858/00-1778-0000-0004-BA7B-4`
SoJaMo	Desc:	SoMaJo is a state-of-the-art tokenizer and sentence splitter for German web and social media texts. You can find more information : `https://github.com/tsproisl/SoMaJo`
	Creator(s):	SfS: Uni-Tuebingen
	Contact:	wlsupport@sfs.uni-tuebingen.de
	PID:	`http://hdl.handle.net/11022/0000-0007-E7D0-9`
CLAR: ASV	Desc:	The sentence segmentizer used by the Wortschatz project for German texts (also containing a very simple tokenizer)
	Creator(s):	CLARIN-D center, Natural Language Processing Group, University of Leipzig
	Contact:	clarin@informatik.uni-leipzig.de
	PID:	`http://hdl.handle.net/11022/0000-0000-94F4-5`

Blingfire	Desc:	Tokenizer/Sentencer from Microsoft, called BlingFire. It is designed for fast-speed and quality tokenization of natural language. For more information about the quality and efficiency of that tokenizer, please check out the corresponding github page: `https://github.com/microsoft/BlingFire`.
	Creator(s):	SfS: Uni-Tuebingen
	Contact:	wlsupport@sfs.uni-tuebingen.de
	PID:	`http://hdl.handle.net/11022/0000-0007-DA1F-2`
SFS-OpenNLP	Desc:	Tokenizer from the OpenNLP Project
	Creator(s):	SfS: Uni-Tuebingen
	Contact:	wlsupport@sfs.uni-tuebingen.de
	PID:	`http://hdl.handle.net/11858/00-1778-0000-0004-BA63-7`
BBAW	Desc:	detects word- and sentence boundaries in raw text using WASTE (`http://www.dwds.de/waste/`)
	Creator(s):	BBAW: Berlin-Brandenburg Academy of Sciences and Humanities
	Contact:	jurish@bbaw.de
	PID:	`https://hdl.handle.net/21.11120/0000-0008-3183-C`
IMS	Desc:	Czech, Slovenian, Hungarian, Italian, French, German, English tokenizer and sentence boundary detector
	Creator(s):	IMS: University of Stuttgart
	Contact:	clarin@ims.uni-stuttgart.de
	PID:	`http://hdl.handle.net/11022/1007-0000-0000-8E1F-F`

Tabelle 6: WebLicht Informationen über die dort verfügbaren Tokenizer

Akshay Mendhakar / Mesian Tilmatine

Automatic Authorship Classification for German Lyrics Using Naïve Bayes

Abstract

Text classification is a prevalent and essential machine-learning task. Machine learning classifiers have developed immensely since their inception. The naïve Bayes classifier is one of the most prominent supervised machine learning classifiers. In this experiment, we highlight the performance of Naïve Bayes for classifying of authors/artists on the German lyrics corpus ("Songkorpus") and compare the classification results with other classifier algorithms. The corpus of investigation consists of six artists with 970 songs in total. Bayes model evaluation measures revealed a precision of 0.91, recall of 0.94, and F1-measure of 0.9. Furthermore, the classification performance with other classifier algorithms did not reveal any statistically significant difference in performance. The results of the study add to the high volume of reports on the classification accuracy of Naive Bayes for the task of lyrical classification.

Keywords: German Lyrics, Text Classification, Naïve Bayes, Machine Learning

1 Introduction

Text mining methodologies have led rise to multiple applications such as text classification, regression, clustering, and association. In text classification, the desired categories are defined in advance, and records are classified into one or some among them (Kowsari et al., 2019). The popularity of text classification systems has grown drastically in the last two decades (Cichosz, 2014; Fell & Sporleder, 2014; Haggblade, Hong & Kao, 2011, Jiang et al., 2018; Kowsari et al., 2019). The application of text classification can be seen in use cases such as content moderation, sentiment classifier, product review classification, email spam classification etc. (Hu & Downie, 2009; 2010; Homem & Carvalho, 2011; Howard, Silla Jr & Johnson, 2011; Jiang et al, 2018.) The most common classifiers are Decision Tree, Perceptron, Naïve Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, and Artificial Neural Networks (Khan, Baharudin, Lee & Khan, 2010). In the past decade, research in the field of song classification has received little focus (Mandel & Ellis, 2005). This can be accredited to the lack of standardized lyrical and audio datasets over the internet. Even though researchers can gather data from websites such as *www.azlyrics.com*, *www.songlyrics.com*, *www.lyrics.com*, etc., the need for large standardized datasets remains a significant issue in song classification. Research in the field of song classification can be noted to identify the genre (Mayer, Neumayer & Rauber, 2008), performers (Pettijohn & Sacco Jr, 2009), sentiment of the song (Logan, Kositsky, & Moreno, 2004; Yang & Lee, 2009), progression of a performer's career (Gomaa, 2022), language usage and geographic distribution (Jin & Ryoo, 2014; Pettijohn & Sacco Jr, 2009) etc.

Audio-based classifications focus on features (spectral and rhythmic), tempo, pitch, rhythm, loudness, etc. Classifiers based on textual lyrics focus on text features such as tokes (words, phrases & sentences), word frequencies, morpho-syntactic structures, rhyme patterns, etc. The performance of audio-based classifiers and text-based classifiers for song corpus has been tested empirically. Research reports on the automatic identification of Frederick Chopin's piano pieces were found to have a classification accuracy of 70% (Davis, 2018). In text-based systems, an accuracy ranging from 50-70 % has been reported in sentiment analysis to discover natural genre clusters (Logan, Kositsky, & Moreno, 2004). In contrast, an accuracy of 76% has been reported for lyric-based song sentiment classification using the sentiment vector space model (Yang & Lee, 2009). Similarly, combined audio and text-based classification systems methods have been reported to yield an accuracy ranging from 48.37% to 66.32% (Mayer, Neumayer & Rauber, 2008).

Some researchers have pointed out that the accuracy of song classifiers highly depends on the type of classifiers. In the study by Khan et al. (2010), the accuracy of the classification changes depending on the classifier used, i.e., for Support Vector Machines, it was 67 % to 97 %. In contrast, for Neural Networks, it improved from 76 % to 100 %, depending on genre.

Automatic Authorship classification has a rich research history and developmental trend. The main idea behind authorship attribution is that texts written by different authors can be distinguished by measuring statistical text features (Stamatatos, 2009). This field has developed rapidly with the development of machine learning classification techniques. Depending on the number of target classifications used to classify the dataset, different approaches can be used to perform the classification task. Decision trees and support vector machines are commonly used for binary classification (Elaidi et al., 2018). This constraint makes it difficult to apply these methods to tasks with more than two target classifications. In terms of obtaining a general toolkit, the naive Bayes classifier seems better suited for broader classification goals (Yang, 2018).

This study aims to test whether a naive Bayesian classifier can correctly predict song authors/artists based on lyrics alone. The used corpus of song lyrics ("Songkorpus"; Schneider, 2020) contains multiple linguistically motivated annotation layers (including POS and lemmatizations), but for this study, we only included plain text. The Naive Bayes Classifier was chosen because it seems well-suited for small datasets: our subcorpus comprises 970 text samples divided into six categories. The following article is organized as follows: The next section briefly describes some theories behind naive Bayesian classifiers. Section 3 describes the methods and measures used in our study. Section 4 discusses the results, and section 5 draws conclusions and provides future directions.

2 Theoretical framework of Naïve Bayes

A naive Bayes classifier is a type of probabilistic classification mechanism based on the Bayesian theorem, a posthumous theory by Thomas Bayes (Bayes & Hume, 1763; Tabak,

2004). Derived from the concepts of inferential statistics, Bayes' theorem serves as the basis for multiple machine learning models. The theorem is based on the logical probability of an event occurring concerning other events or features (Lewis, 1998). Equation (1) shows the Bayesian rule with P(A) & P(B) denoting the probability of an event A and event B respectively. Similarly, P(A|B) denotes the probability of A concerning B and vice versa in P(B|A).

P(A|B) = P(B|A) P(A) / P(B) **(1)**

Further, if we try to generalize the equation (1) for a series event represented by x and y, the equation becomes (2).

P(yi | x1, x2, …, xn) = P(x1, x2, …, xn | yi) * P(yi) / P(x1, x2, …, xn) **(2)**

In this estimates the prior P (yi) from the dataset is a conditional factor of the class P(x1, x2, …, xn | yi). This estimation is unviable if the sample size is small. Therefore, the dataset has to include a large number of samples which helps in the estimation of different possible combinations of a given value to predict its possibility. In this situation, where the number of observations in the dataset is growing, the application of the Bayes Theorem becomes difficult. In the case of variables being conditionally independent given the class, the estimation of the variable-value data is represented by the equation (3)

$$P(\mathbf{x} \mid y) = \prod_{i=1}^{n} P(x_i \mid y) \quad \textbf{(3)}$$

Here, n represents the number of variables in the sample and x_i is the i^{th} value of the variable x. In situations where there are multiple classes, where we represent the number of classes with *k* and c_i as the i^{th} class equation (3) is represented by equation (4). Thus we represent a classifier that is linear in nature.

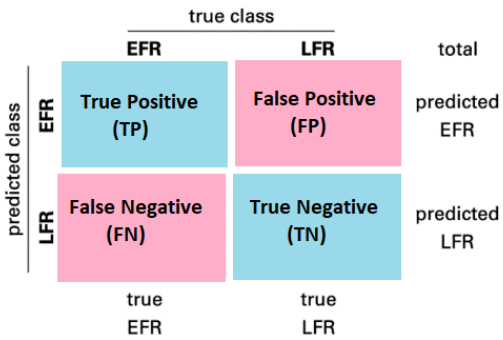
$$P(\mathbf{x}) = \prod_{i=1}^{k} P(c_i) P(\mathbf{x} \mid c_i) \quad \textbf{(4)}$$

When the dataset contains categorical variables, frequency counts play a vital role in the estimation of the probabilities of P(y) and $P(x_i \mid y)$. It involves methods like the Laplace estimation of the m-estimation method to measure the frequency. Further, this estimation can be compared and updated with new data as the training data is used as a single pass while training only. Therefore, this form of learning is supported by incremental learning. Similarly, when we look at numerical variables, discretization of the data is used. Therefore, the probability estimation is based on density estimation.

Naïve Bayes classifier utilizes the naïve Bayes theorem to solve a wide range of classification problems (Rish, 2001). Its computational efficiency and ease of implementation make it one of the most utilized supervised machine learning classification methods. Applications cover document classification (Ting et al., 2011), spam filtering (Metsis et al., 2006), content moderation (Risch, Ruff & Krestel, 2020), sentiment classification (Narayanan, Arora & Bhatia, 2013) and many more. The widespread acceptability of the naïve Bayes classifier is due to a wide range of factors such as its computational efficiency, low variance, its incremental learning abilities, strong aversion against missing values or high variability in the data.

Computational efficiency in modelling and prediction is an indisputable advantage over some other classification algorithms, which is due to its ability to parallelize data sets. i.e., the training time is linear for the number of training examples and the number of attributes and the classification time is linear for the number of attributes and is not affected by the number of examples studied. For the traits mentioned above, it would be helpful to add two more elements: resistance to over-equipping and the ability to manipulate multiple picks. naive Bayes operates on lower-order probability estimates derived from training data. They can easily be updated as new training data becomes available (Kohavi,1996). The classification results are prone to low variability with a high bias cost. It always uses all attributes for all predictions and is therefore relatively sensitive to noise in classified examples. Because it uses probability, it is also relatively insensitive to noise and missing values in the training data (Gama, Medas & Rodrigues, 2005).

In order to evaluate the performance of the naïve Bayes classifier model, we follow the results obtained from the confusion matrix (Figure 1). True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) allow to compute Precision (PR), Recall (RE), Accuracy (CA), Error rate (ER) and F1 measures. The formulas are displayed in figure 1.



$$PR = \frac{TP}{TP+FP}$$

$$RE = \frac{TP}{TP+FN}$$

$$CA = \frac{TP+TN}{TP+TN+FP+FN}$$

$$F_1 = \frac{2TP}{2TP+FP+FN}$$

Figure 1: Classical confusion matrix.

For any machine learning model, the F1 index is one of the most important metrics to determine its performance (Lipton, Elkan & Narayanaswamy, 2014). The value of the F1 index ranges from 0 to 1, where 0 is the worst possible score with poor classification. Another performance metric is the Receiver Operating Characteristics (ROC) curve (figure 2). It graphically determines classification performance of a binary classifier as a function of TP and FP measures.

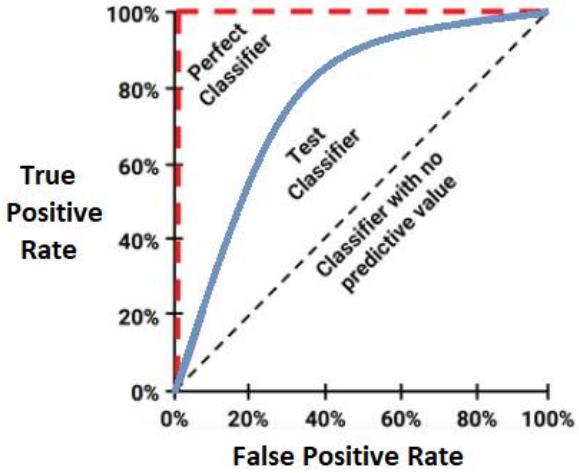


Figure 2: ROC curve.

In figure 2, the dotted diagonal signifies the zero thresholds or no classification. The blue line measures the true performance of the classifier with respect to the red dotted line which symbolizes perfect classification with 100% accuracy (100% true positives and 0% false negatives). In uniformly distributed datasets, measuring the accuracy of the classifier is enough to predict the classifier's performance. Whereas with imbalanced datasets, ROC AUC may be more significant. ROC AUC considers the trade-offs between precision and recall, while accuracy only determines how many predictions are correct. Generally, AUC is preferred over accuracy as it is a much better indicator of model performance.

3 Experiments

We conduct a series of experiments focusing on authorship attribution for song lyrics of Udo Lindenberg, Konstantin Wecker, Stoppok, Ulla Meinecke, Hannes Wader, and Fettes Brot. The results of naïve Bayes and other classification approaches are contrasted.

The first half of the experiment was focused on applying the naïve Bayes classifier to the songkorpus dataset. The publicly available repository provides detailed corpus statistics, as

well as visualizations on character, word, verse, song and corpus level. Our first step was data gathering and pre-processing from the songkorpus website. The data came in XML format and was further transformed to plain text. The processed text included semantic and structural information related to the artist and the song verses. All additional information was omitted. The plain textual data was subject to further linguistic analysis, using the Profiling UD tool (Brunato et al., 2020). Profiling UD extracts 130 computational linguistic parameters under raw textual features, morphosyntactic and syntactic parameters. A similar protocol as noted in (ref. Mendhakar, 2022) was used to extract and process the Linguistic features extracted from the tool. The extracted parameters were tabulated into an excel file. By feature reduction, pruning of the number of features was carried out. The resultant dataset consisted of 970 data points of 115 parameters categorized under six different artists. After initial dataset creation, randomization of rows was made. The basic demographics of the dataset created are highlighted in table 1. Table 1 represents the database representation of each artist.

Artist	Songs considered	Representation in the dataset
Fettes Brot	91	9.38%
Udo Lindenberg	316	32.58%
Ulla Meinecke	78	8.04%
Stoppok	77	7.94%
Hannes Wader	168	17.32%
Konstantin Wecker	240	24.74%

Table 1: Description of the dataset considered in the study.

All experiments were conducted on a system with an Intel Corei7 CPU at 2.4GHz, 8 GB of RAM, and 1 TB of secondary storage, running windows 10 and MATLAB 2021b. The dataset was loaded onto the machine learning toolkit of MATLAB software for further processing. The utility of MATLAB's machine learning toolkit was due to the capability of comparing multiple classifiers in one place and also due to its ease of implementation. The classical naïve Bayes classifier was designed with the preset features for classification. The developed dataset was split into a training and testing dataset. Two-thirds of the dataset was used to train the classifier and the rest of the data was used for its testing and validation. The split of the dataset was randomized to eliminate any artist bias. The parameters were tweaked and multiple runs were carried out to find the best possible classification accuracy.

In the second stage of the experiment, the classification accuracy of the naïve Bayes classifier was compared with other commonly used classifiers, such as logistic regression (LR), support vector machines (SVM), naive Bayes (NB), decision tree classifier (DTC), K-nearest neighbor (KNN), and neural networks (NNs). To improve the classification accuracy, hyper-parameter tuning and dimension reduction were employed. Additionally, multiple iterations of the classifier parameters were run by removing correlated features, using log probabilities in calculations, and parallelized calculations were performed. These additional steps were used in order

to identify the best possible classification results. The performance of each classifier is compared in the next section.

4 Results and discussion

When measuring performance of the Naive Bayes classifier, various iterations were carried out by implementing the Laplace smoothing, with the classifier's accuracy being the best at the estimator's value of 0.06. Figures 3 & 4 display the confusion matrix and ROC, respectively, of the best naïve Bayes classifier. Table 2 summarizes accuracy, error rate, precision, recall and F-measure across each class of the dataset.

Class	Accuracy(CA)	Precision (PR)	Recall (RE)	F1 Score
1	97.94 %	0.89	0.89	0.89
2	93.40 %	0.92	0.88	0.9
3	97.73 %	0.89	0.82	0.85
4	97.22 %	0.85	0.79	0.82
5	94.95 %	0.83	0.89	0.86
6	95.26 %	0.88	0.93	0.91
Total	96.08 %	0.88	0.87	0.87
Total with 0.06 Laplace	**97.03 %**	**0.91**	**0.94**	**0.90**

Table 2: Evaluation of the Naïve Bayes classifier.

In figure 4, the ROC curve of the classifier shows that the area under the curve (AUC) was 0.89, which is a very good classifier performance.

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	TPR	FNR
Class 1	81	5	0	3	1	1	89%	11%
Class 2	3	277	3	2	5	12	88%	12%
Class 3	1	4	64	1	2	0	82%	18%
Class 4	1	6	2	61	2	0	79%	21%
Class 5	4	14	5	4	149	3	89%	11%
Class 6	1	10	4	6	9	224	93%	7%

Figure 3: Confusion matrix plot of the Naïve Bayes classifier.

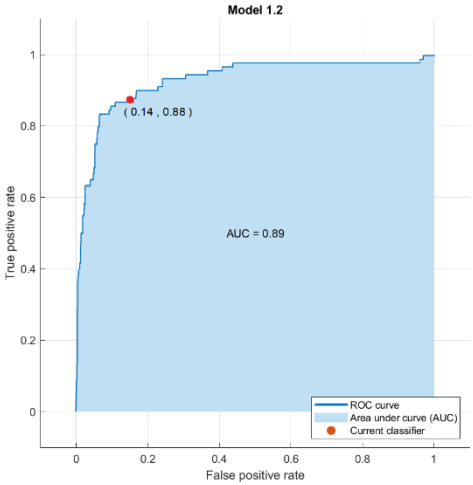


Figure 4: ROC curve of the Naïve Bayes classifier.

Even with the application of Laplace smoothing, the accuracy of the classifier did not change significantly. Therefore, the naïve Bayes classifier build in this experiment has an overall accuracy of 97.03 %

4.1 Comparing the performance of different classifiers

To rank the performance of our classifiers, we apply different classifiers to the dataset. Table 3 show that Naïve Bayes had the best performance and obtained better results for almost all metrics. The average accuracy rate of Naïve Bayes for the test set is 91% and its highest accuracy is 97%. However, the comparison of training and test set accuracies indicates that Naïve Bayes and especially RF suffer from overfitting problems. Decision Trees show the worst performance, deep learning algorithms like neural networks (ANN, CNN and LSTM) take the most execution time, with LSTM being the slowest. One outstanding point is that CNN performed very well and was much faster than ANN and LSTM.

Algo-rithms	Preci-sion (%)	Recall (%)	F1-Score (%)	Training Set Accu-racy (%)	Test Set Accuracy (%) (Avg/Highest)
Naïve Bayes	**91**	**94**	**90**	93	91/97
SVM	73	72	72	70	67/73
RF	80	79	79	100	76/79

ANN	77	78	77	83	82/82
LSTM	87	86	84	88	81/87
CNN	91	92	91	95	91/96

Table 3: Comparison of different classifiers of the study.

5 Conclusion

Creating a meaningful lyrics dataset is a tedious and time-consuming task. For example, guest appearances by other artists or two versions of the same song (e.g. studio version and live version) must be handled with care. By using the precompiled Songkorpus, we empirically tested the accuracy of different authorship classificators on a reliable dataset. The results of our best model seems promising and are in accordance with comparable research reports on naïve bayes classifiers (Rish, 2001; Dai et al., 2007; Labatut & Cherifi, 2012; Nitze, Schulthess & Asche, 2012; Altheneyan & Menai, 2014; Baron, 2016; Shih, Stow, & Tsai, 2019). It can be concluded from our experiments that the Naive Bayes classifier seems to be a good choice for authorship attribution of song lyrics, at least for the investigated singer-songwriter dataset. Since the used dataset is relatively small, it would be a reasonable choice to use our classifiers on bigger datasets in order to make better generalizations.

6 References

Altheneyan, A. S., & Menai, M. E. B. (2014). Naïve Bayes classifiers for authorship attribution of Arabic texts. *Journal of King Saud University-Computer and Information Sciences*, *26*(4), 473-484.

Baron, G. (2016). Comparison of cross-validation and test sets approaches to evaluation of classifiers in authorship attribution domain. In *International Symposium on Computer and Information Sciences* (pp. 81-89). Springer, Cham.

Bayes, T., & Hume, D. (1763). Bayes's Theorem. In *Proceedings ofthe British Academy* (Vol. 113, pp. 91-109).

Brunato, D., Cimino, A., Dell'Orletta, F., Venturi, G., & Montemagni, S. (2020). Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 7145-7151).

Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, *81*(3), 429-450.

Cichosz, P. (2014). Data mining algorithms: explained using R. John Wiley & Sons.

Cunningham, S. J., Bainbridge, D., & Falconer, A. (2006). " More of an art than a science": Supporting the creation of playlists and mixes.

Dai, W., Xue, G. R., Yang, Q., & Yu, Y. (2007). Transferring naive bayes classifiers for text classification. In *AAAI* (Vol. 7, pp. 540-545).

Davis, A. (2018). Classical Composer Identification on Interval Features for CS230-Spring 2018. Retrieved from http://cs230.stanford.edu/projects_spring_2018/reports/8289789.pdf

Elaidi, H., Elhaddar, Y., Benabbou, Z., & Abbar, H. (2018). An idea of a clustering algorithm using support vector machines based on binary decision tree. In *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)* (pp. 1-5). IEEE.

Fell, M., & Sporleder, C. (2014). Lyrics-based analysis and classification of music. In Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers (pp. 620-631).

Fukunaga, K., & Hayes, R. R. (1989). Estimation of classifier performance. *IEEE transactions on pattern analysis and machine intelligence*, *11*(10), 1087-1101.

Gama, J., Medas, P., & Rodrigues, P. (2005). Learning decision trees from dynamic data streams. In *Proceedings of the 2005 ACM Symposium on Applied computing* (pp. 573-577).

Gomaa, W. (2022). Lyrics Analysis of the Arab Singer Abdel ElHalim Hafez. *Transactions on Asian and Low-Resource Language Information Processing*.

Haggblade, M., Hong, Y., & Kao, K. (2011). Music genre classification. *Department of Computer Science, Stanford University*.

Homem, N., & Carvalho, J. P. (2011). Authorship identification and author fuzzy "fingerprints". In *2011 Annual Meeting of the North American Fuzzy Information Processing Society* (pp. 1-6). IEEE.

Howard, S., Silla Jr, C. N., & Johnson, C. G. (2011). Automatic lyrics-based music genre classification in a multilingual setting. In *Proceedings of the Thirteenth Brazilian Symposium on Computer Music*.

Hu, X., & Downie, J. S. (2010). When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis. In *ISMIR* (pp. 619-624).

Hu, X., Downie, J. S., & Ehmann, A. F. (2009). Lyric text mining in music mood classification. *American music*, *183*(5,049), 2-209.

Idicula-Thomas, S., Kulkarni, A. J., Kulkarni, B. D., Jayaraman, V. K., & Balaji, P. V. (2006). A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli. *Bioinformatics*, *22*(3), 278-284.

Jiang, M., Liang, Y., Feng, X., Fan, X., Pei, Z., Xue, Y., & Guan, R. (2018). Text classification based on deep belief network and softmax regression. *Neural Computing and Applications*, *29*(1), 61-70.

Jin, D. Y., & Ryoo, W. (2014). Critical interpretation of hybrid K-pop: The global-local paradigm of English mixing in lyrics. *Popular Music and Society*, *37*(2), (pp. 113-131).

Juba, B., & Le, H. S. (2019). Precision-recall versus accuracy and the role of large data sets. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 4039-4048).

Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, *1*(1), 4-20.

Kohavi, R. (1996). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd* (Vol. 96, pp. 202-207).

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150.

Labatut, V., & Cherifi, H. (2012). Accuracy measures for the comparison of classifiers. *arXiv preprint arXiv:1207.3790*.
Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning* (pp. 4-15). Springer, Berlin, Heidelberg.
Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). Thresholding classifiers to maximize F1 score. *arXiv preprint arXiv:1402.1892*.
Logan, B., Kositsky, A., & Moreno, P. (2004). Semantic analysis of song lyrics. In *2004 IEEE International Conference on Multimedia and Expo- ICME* (pp. 827-830).
Mandel, M. I., & Ellis, D. P. (2005). Song-level features and support vector machines for music classification. (pp. 594-599).
Mayer, R., Neumayer, R., & Rauber, A. (2008). Rhyme and Style Features for Musical Genre Classification by Song Lyrics. In *Ismir* (pp. 337-342).
Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006). Spam filtering with naive bayes-which naive bayes?. In *CEAS* (Vol. 17, pp. 28-69).
Narayanan, V., Arora, I., & Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced Naive Bayes model. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 194-201). Springer, Berlin, Heidelberg.
Nitze, I., Schulthess, U., & Asche, H. (2012). Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification. *Proceedings of the 4th GEOBIA, Rio de Janeiro, Brazil*, *79*, 3540.
Pettijohn, T. F., & Sacco Jr, D. F. (2009). The language of lyrics: An analysis of popular Billboard songs across conditions of social and economic threat. *Journal of language and social psychology*, *28*(3), (pp. 297-311).
Risch, J., Ruff, R., & Krestel, R. (2020). Offensive language detection explained. In *Proceedings of the second workshop on trolling, aggression and cyberbullying* (pp. 137-143).
Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46).
Salley, K. (2011). On the interaction of alliteration with rhythm and metre in popular music. *Popular Music*, *30*(3), 409-432.
Schneider, R. (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* (pp. 842-848).
Shih, H. C., Stow, D. A., & Tsai, Y. H. (2019). Guidance on and comparison of machine learning classifiers for Landsat-based land cover and land use mapping. *International Journal of Remote Sensing*, *40*(4), 1248-1274.
Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, *60*(3), 538-556.
Stamatatos, E., & Widmer, G. (2002). Music performer recognition using an ensemble of simple classifiers. In *ECAI* (pp. 335-339).
Tabak, J. (2004). Probability and Statistics: The Science of Uncertainty. Facts on File. *Inc., NY, USA*, 46-50.

Ting, S. L., Ip, W. H., & Tsang, A. H. (2011). Is Naive Bayes a good classifier for document classification. *International Journal of Software Engineering and Its Applications*, *5*(3), 37-46.
Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, *25*(3), 2277-2293.
Yang, D., & Lee, W. S. (2009). Music emotion identification from lyrics. In *2009 11th IEEE International Symposium on Multimedia* (pp. 624-629).
Yang, F. J. (2018). An implementation of naive bayes classifier. In 2018 International conference on computational science and computational intelligence (CSCI) (pp. 301-306). IEEE.

Correspondence

Akshay Mendhakar
University of Warsaw, Faculty of Applied Linguistics, Poland
University of Vienna, Empirical visual aesthetics lab, Austria
a.mendhakar@uw.edu.pl

Mesian Tilmatine
Free University of Berlin, Department for Experimental and Neurocognitive Psychology
Radboud University Nijmegen, Centre for Language Studies, The Netherlands
m.tilmatine@fu-berlin.de

NOTE: The authors are employed under the ELIT network. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860516.