Pop Lyrics Through Time: Challenges in Corpus-Based Modeling of Linguistic and Emotional Dynamics in German Pop Lyrics

Roman Schneider

Leibniz Institute for the German Language
R5 6-13, 68161 Mannheim / Germany
schneider@ids-mannheim.de

Abstract

This paper presents a large-scale diachronic analysis of German pop lyrics based on a linguistically rich, TEI-encoded monitoring corpus. We describe multi-layer annotation and reproducible workflows for deriving higher-level features at scale, including lexical diversity indices, a pronoun-based subjectivity measure, modal particle density, and a length-normalized sentiment intensity score. Particular attention is paid to the development and evaluation of pipelines for two notoriously challenging phenomena: modal particles and sentiment. For modal particles, we build a manually curated gold standard and train sequence models whose performance we relate to inter-annotator agreement. For sentiment, we integrate a lexicon-based resource with a dedicated human annotation experiment to assess reliability and alignment with expert judgments. On this basis, we investigate how structural and affective features co-vary in the corpus and how they change over time, showing, among other trends, declining lexical diversity and sentiment intensity alongside a slight increase in first- and second-person pronouns. Beyond the empirical findings, the paper highlights practical challenges in managing culturally specific corpora, and makes evaluation materials available to support transparent, reusable corpus-based research on popular music and related domains.

Keywords: culture-specific corpora; diachronic analysis; annotation quality

1. Introduction

The empirical study of language increasingly depends on large-scale resources that capture diverse communicative domains. While substantial progress has been made in developing corpora for functional, non-literary registers, popular culture remains comparatively underrepresented. Pop song lyrics – despite their considerable communicative "impact factor" (Kreyer and Mukherjee, 2007) – have traditionally been relegated to literary and cultural studies rather than linguistics, reflecting both disciplinary conventions and the former absence of sustainable reference corpora.

Recent years have seen a shift, as research on lyrics has begun to gain visibility within empirical linguistics (cf. Werner et al., 2025), supported by the emergence of dedicated resources (Schneider, 2026). For German, the release and continuous expansion of the extensively annotated Songkorpus (Schneider, 2020, 2025) has established pop lyrics as a legitimate and sustainable object of quantitative investigation. From a linguistic perspective, sung – or in some cases more spoken, as in rap lyrics are particularly interesting because they constitute a hybrid text type that draws on features of both conceptual orality and literacy. Songs often incorporate colloquial and spoken-language elements while at the same time adhering to genre-specific structural and stylistic conventions. This duality makes lyrics a valuable source for studying contemporary language use, the interaction between spoken and written registers, and broader cultural dynamics.

Building on work that locates texts in a hybrid space between "language of immediacy" and "language of distance" (Koch and Oesterreicher, 2012), we combine structural features with sentiment analysis to provide a multidimensional account that integrates both linguistic and affective dimensions. At the word-form level, features such as modal particle proportion, pronoun proportion, and lexical richness have been shown to be key predictors of a text's position in this hybrid space (Broll and Schneider, 2023). Our approach extends this by adding an empirically validated measure of sentiment intensity, enabling us to examine how markers of proximity and distance interact with affective tone.

On the empirical side, large-scale studies of English lyrics report that lyrics have become simpler and more repetitive in terms of lexical richness, readability, and repetition over the last five decades, with rap retaining relatively high complexity (Parada-Cabaleiro et al., 2024). For German popular music, Hunke et al. (2025) find increasingly negative tendencies in chart songs, though their claims are constrained by multilingual sampling and translation into English. This paper addresses both lexical and affective developments in a substantially larger, German-only corpus and explicitly links them to core features of German grammar and discourse such as modal particles and pronouns.

The selection of linguistic features in this study is guided by the assumption that song lyrics are not only vehicles of thematic content, but also sites of perspective-taking, stance marking, and affective expression. Pronoun usage provides a window into how subjectivity and interpersonal orientation

are linguistically constructed, for instance through shifts between self-reference and addressivity. Modal particles, a characteristic feature of German interactional language, index speaker stance, pragmatic nuance, and degrees of communicative immediacy. Sentiment measures, in turn, offer a coarse-grained approximation of affective framing at scale. Taken together, these features are intended to capture complementary dimensions of how lyrics position speakers, encode attitudes, and shape emotional expression in a genre that is simultaneously written, performed, and socially embedded.

This paper is conceived as a methodological contribution to corpus-based research on creative, non-standard language. Its central aim is to demonstrate robust strategies for extracting linguistic and affective features from song lyrics, a domain that systematically challenges standard NLP assumptions. The substantive analysis of German pop lyrics over time serves as an empirical test case that illustrates the potential and limitations of these methods. While the findings offer insights into linguistic and emotional dynamics, they are intended as exploratory and methodologically grounded rather than as definitive claims about cultural change.

1.1 Related work

In the broader domain of song lyrics studies, Parada-Cabaleiro et al. (2024) analyze 12,000 English-language songs across five genres, focusing on lexical diversity, readability, and repetition. They report that lyrics have become simpler and more repetitive over time, with lexical richness declining and repetition increasing. Rap stands out as retaining higher lexical complexity relative to other genres. These findings raise the question of whether similar trends can be observed in other languages and corpora with different sampling strategies.

Modal particles (MPs) are a quintessentially German word class, used to signal subjectivity, discourse stance, or speaker-listener intimacy. Their insertion or omission can subtly shift force, politeness, or hedging. Because MPs sit at the interface of semantics, pragmatics, and discourse, they resist simple categorization: they overlap with adverbs, discourse particles, or focus markers, and their acceptability is context- and order-sensitive (Diewald, 2007; Blühdorn, 2019; Schoonjans, 2018). Frequent homography and context dependence complicate automated detection (Storø, 2022), and many standard tag sets do not label MPs separately. While individual studies have examined MPs in literary genres (e.g. Hentschel, 2010), large-scale analyses on pop corpora are lacking.

Sentiment analysis of song lyrics has begun to attract attention. Hunke et al. (2025) compute topic and sentiment models for German chart songs and report increasingly negative tendencies over time, though their sample is relatively small and multilingual. Beyond bag-of-words scoring, discourse-aware models such as Discourse-LSTM (Kraus and Feuerriegel, 2019) incorporate rhetorical structure to mitigate position biases. From the broader NLP literature, classic overviews (Pang and Lee, 2008; Liu, 2015) emphasize pitfalls that are especially relevant for lyrics: irony, context dependence, lexicon domain mismatch, polysemy, and short-text limitations. At the same time, contemporary sentiment research increasingly employs neural and transformer-based models that integrate context and negation, though their behavior on poetic and lyrics data has not yet been systematically studied.

Across all phenomena under investigation, the Tool Misuse perspective (Sluyter-Gäthje and Trilcke, 2022) is relevant. It reminds us that the “errors” of NLP tools on literary texts may reflect systematic stylistic deviation, while such tools remain essential for analyzing large corpora.

1.2 Research questions and contribution

We address three research questions that bridge methodological and substantive concerns:

- **RQ1:** How reliably can automated methods detect notoriously difficult-to-capture linguistic and affective phenomena – specifically modal particles and sentiment intensity – as assessed against manually curated gold standards or experiments involving human participants?
- **RQ2:** To what extent do lexical-syntactic and emotional phenomena co-vary? In particular, how are lexical diversity, modal particle usage, pronoun use, and sentiment intensity associated, as evaluated using correlation and regression analyses?
- **RQ3:** Can temporal patterns or trends be identified across the dataset, indicating systematic change over time in these structural and affective features?

Taken together, these questions are designed to first establish the reliability of the analytical approach (RQ1), and then to explore its application to diachronic linguistic and affective patterns (RQ2, RQ3).

The study’s contributions are fourfold:

1. We use Songkorpus, a large, multi-layer annotated corpus of German song lyrics, to conduct a diachronic analysis spanning more than five decades.
2. We develop and evaluate an automated approach to modal particle identification in lyrics, grounded in a manually annotated gold standard.

3. We integrate sentiment analysis with a dedicated human annotation experiment to derive and validate a measure of sentiment intensity at the song level.
4. We model the interrelations and temporal development of lexical diversity, modal particles, pronouns, and sentiment intensity, thereby linking structural features of German with affective expression in popular music.

In Section 2, we address RQ1 by detailing the corpus, feature extraction, and validation of the most error-prone automatic annotations, before turning in Section 3 to RQ2 and RQ3, where we analyze interrelations among the features and their diachronic development.

2. Data and methods

2.1 Corpus

Songkorpus (Schneider, 2025) is currently the largest scientifically curated monitoring collection of German-language song lyrics and constitutes a unique public resource for linguistic and cultural research. With more than 15,000 lyrics and approximately five million tokens, it covers over six decades of music history and continues to be updated. Archives are revised annually (e.g. through the addition of current chart songs), and recent developments include an expanded hip-hop ("Deutschrap") archive reflecting the official German hip-hop charts.

The corpus is systematically stratified into multiple archives. Artist-specific archives collect the complete works of established performers and emerging artists. Thematic archives are organized along historical lines – such as songs from the former GDR – or by genre, including a dedicated archive of early-1980s "Neue Deutsche Welle". A monitoring archive covers all German-language chart songs since the mid-1950s; for this study, however, we focus on data from 1970 onward, as this reflects the more data-intensive period, enabling robust temporal stratification and meaningful diachronic analysis. This combination allows the data set to represent both mainstream and subcultural repertoires across a wide range of themes and time periods.

Intellectual property considerations play a central role in the compilation and maintenance of the Songkorpus. As song lyrics are typically protected by copyright, full-text data cannot be freely redistributed. The corpus therefore follows a controlled access model: while the texts are stored and processed for research purposes, public distribution is restricted to derived data (e.g. annotations, frequency information, or aggregated features). For artist-specific archives, this framework is complemented by individual agreements with rights holders, which allow more extensive use under clearly defined conditions.

Each text is encoded in TEI format and enriched with bibliographic metadata (e.g., author, release year, genre, source); the central importance of metadata for linguistic data is discussed in Trippel (2025). Crucially, lyrics are annotated at multiple linguistic levels, including lemmata, part-of-speech tags (STTS), named entities, neologisms, syntactic constituents, verse structures, and, in some cases, rhyme schemes. The corpus is accessible via an online portal with tailored search forms, visualizations, and downloadable datasets (bag-of-words, n-grams, word vectors).

To manage the corpus as a continuously updated monitoring resource, we rely on an automated processing pipeline that standardizes ingestion, annotation, and export of derived variables (cf. Schneider, in preparation). Crucially, this pipeline also computes the linguistic and affective features that we investigate in the present study.

Code-switching between German, English, and other languages is a frequent feature of contemporary song lyrics and is explicitly represented in the Songkorpus. The corpus uses TEI-based language attributes to mark segments at a fine-grained level. This allows non-German material (e.g. English insertions or multilingual passages) to be systematically identified and filtered. For the present analyses, only segments annotated as German are included to ensure methodological consistency.

2.2 Feature set and scope

We focus on four feature families that jointly capture aspects of lexical richness, interpersonal alignment, discourse marking, and affective density:

- **Lexical diversity** (MATTR, MTLD) as indicators of lexical richness and, indirectly, of how planned and elaborated (more distant) vs. formulaic and repetitive (more immediate) the language is.
- **Pronoun index** (PRON) to capture subjectivity and address patterns through first-/second- vs. third-person forms.
- **Modal particle ratio** (PTKM) as a core marker of German discourse stance, subjectivity, and speaker–listener intimacy.
- **Sentiment intensity** (SI) as a length-normalized measure of affective density, irrespective of polarity.

This selection does not exhaust the space of relevant features. We do not explicitly model syntactic complexity, rhythmic structure, or topic/lexical field distributions, nor do we include discourse-structural sentiment models or multi-dimensional emotion categories. We therefore understand our account as *multidimensional* but not fully comprehensive: it targets key lexical-syntactic and affective indicators while leaving other dimensions to future work.

2.3 Linguistic features

2.3.1 Lexical diversity

Lexical diversity is a key criterion in distinguishing between spoken and written language, or more generally between language of immediacy and language of distance (Malvern et al., 2004; Koch and Oesterreicher, 2012). Simple type-token ratios (TTR) are strongly correlated with text length and thus limited in interpretive value. A Pearson correlation across the corpus reveals a moderate positive relationship between token count and year of publication ($r \approx 0.40$), indicating that song texts tend to become longer over time; the number of texts per year also increases slightly. Both trends motivate the use of more robust metrics.

We consider three established measures:

- **Standardized TTR (STTR)**, which divides texts into successive windows of fixed size (here: 100 tokens) and computes the mean TTR across segments (Scott, 2004).
- **Moving-Average TTR (MATTR)**, which applies a sliding window over the text and averages TTR across overlapping windows, thereby avoiding partial final segments (Covington and McFall, 2010).
- **Measure of Textual Lexical Diversity (MTLD)**, which estimates diversity through sequential passes (forward and backward) and is designed to be comparatively robust to length (McCarthy and Jarvis, 2010).

All three measures are sensitive to very short texts, a general limitation emphasized by Bestgen (2024). In song lyrics, this is particularly relevant because many songs would be relatively short without repeated refrains and hooks, which at the same time introduce strong local repetition. After initial exploration of STTR, MATTR, and MTLD, we retain the latter two: MATTR is particularly attractive for this genre because its sliding-window procedure is sensitive to the recurring, stylistically driven repetitions typical of musical texts. MTLD, by contrast, assesses lexical diversity sequentially across the entire text and captures how repetitions influence lexical variation at a more global level. Using both measures thus allows us to compare a locally sensitive, window-based approach with a sequential, text-wide measure, and to evaluate which metric more adequately handles the specific properties of pop lyrics, also in interaction with the other features.

In this study, we compute both measures for all texts, report their correlation, and explore their differential behavior. In the subsequent regression models, we focus on MATTR, which yields more stable and interpretable associations with other features in this corpus, and we explicitly revisit the divergence between MATTR and MTLD in the discussion.

2.3.2 Pronoun index

Personal and possessive pronouns are included as features because they index interpersonal relations and subjectivity. First- and second-person pronouns (German: *ich, du, mein, dein*, etc.) are comparatively infrequent in conventional written corpora, where third-person forms (German: *er, sie, es, sein, ihr,* etc.) predominate, but are expected to occur more frequently in song lyrics, likely reflecting their orientation toward intimacy, immediacy, and emotional expression. The lyrical subject ('I') and addressee ('you') are frequently positioned in relation to, or in contrast with, a third person ('he/she'). We operationalize this by measuring the polarity between first- and second-person pronouns versus third-person pronouns using automated part-of-speech tagging combined with curated wordform lists. As with the lexical diversity measures, we compute a single score (PRON) for each text.

2.3.3 Modal particles

Modal particles pose a particular challenge in corpus-based research due to their polyfunctionality and homonymy with other word classes. Identification based solely on word forms is error-prone, as many candidate forms also occur as adverbs, connectors, or discourse markers. While MPs in German typically appear in the Mittelfeld position, their placement is flexible and influenced by discourse and syntax, and their subtle pragmatic functions cannot be captured by syntax alone.

To illustrate the linguistic behavior of MPs, consider a few examples. **„**Aber**“** can convey emphasis or surprise: *„Das ist aber schön geworden!“* ["That has really turned out nicely!"] signals positive astonishment, while *„Du bist aber spät dran“* ["You are really late"] indicates mild criticism. **„**Doch**“** often marks known or expected information, e.g., *„Das ist doch nicht so schlimm“* ["That's really not so bad"], or introduces contrast, as in *„Er wollte zuerst nicht, doch er kam trotzdem“* ["He didn't want to at first, but he came anyway"]. **„**Bloß**“** can relativize or warn: *„Mach bloß keinen Blödsinn!“* ["Just don't do anything stupid!"] highlights caution, whereas *„Er hat bloß Pech gehabt“* ["He just had bad luck"] downplays the situation. Other particles include **„**ja**“** (*„Du weißt das ja auch“* ["You know that too"]) for self-evident information, **„**mal**“** (*„Ich schau mal, ob ich Zeit habe“* ["I'll just see if I have time"]) for tentative or polite action, and **„**vielleicht**“** (*„Ich war vielleicht überrascht“* ["I was maybe surprised"]) for uncertainty or hedging. These examples show how modal particles operate in subtle, context-dependent ways, reinforcing the need for context-sensitive modeling.

We target a core inventory of 14 high-frequency MPs: *aber, auch, bloß, denn, doch, eben, etwa, halt, ja, mal, nur, schon, vielleicht*, and *wohl*. Without making any definitional claim, this set is widely recognized as constituting a stable MP core. We adopt a two-step approach (Section 2.5): (i) create a manually annotated gold standard for these forms in context, then (ii) train and evaluate a sequence model for automatic MP detection. For the corpus-level analyses, we compute PTKM, the ratio of automatically detected MPs to total tokens per text.

2.4 Sentiment and emotional modeling

2.4.1 Automated sentiment computation

Most classical sentiment approaches in NLP rely on lexicons and rule-based heuristics; more recent work increasingly uses neural and transformer-based models that integrate contextual information. For German lyrics, however, large genre-specific labeled datasets for supervised training are scarce, and interpretability is a concern. We therefore adopt a lexicon-based approach as a transparent and easily interpretable baseline.

Specifically, we use SentiMerge (Emerson and Declerck, 2014), developed to integrate and harmonize multiple German sentiment resources. SentiMerge assigns sentiment scores to lemmata based on several lexica and applies weighting schemes that account for both reliability and frequency. For example, the adjective *abscheulich* 'abhorrent' receives a strongly negative score of approximately -0.9 with a weight of 9.7, while the noun *Abgott* 'idol' is assigned a positive score of +0.9 with a weight of 6.7.

To apply SentiMerge consistently, we harmonize POS tags between the STTS tagset used in Songkorpus and SentiMerge's categories: all nouns are mapped to "N", full verbs in all forms to "V", and other categories (e.g. onomatopoetic forms, discourse markers) to the closest available types. Sentiment is computed at the level of the entire song text, which we treat as a coherent unit, consistent with the other features. While verses or stanzas may vary in tone, the present study focuses on overall song-level affect rather than within-song dynamics.

Simple sentiment sums depend on text length: longer songs naturally accumulate more sentiment-bearing words, regardless of orientation. Preliminary analyses show that the songs with the highest total positive sentiment also rank highest in total negative sentiment, and total sentiment is negatively correlated with token count. We therefore experiment with several normalizations, including division by total tokens and by sentiment-bearing tokens. Polarity measures (ratio of positive to negative sums) turn out to be relatively stable over time, showing only a slight shift toward negativity and providing limited diachronic discrimination.

For our main analyses, we focus on Sentiment Intensity (SI), defined as the sum of the weighted absolute values of all sentiment scores in a song divided by token count. SI captures affective density (how strongly emotional language is mobilized) without privileging positive or negative orientation. This aligns with our primary interest in *how much* emotional language is used, rather than in assigning songs an overall positive or negative label or distinguishing between specific evaluative stances.

2.4.2 Why not transformer-based sentiment?

Transformer-based models offer powerful context-sensitive sentiment and emotion classification and can handle negation and long-distance dependencies more effectively than lexicons. However, there are several reasons why we do not use them here:

1. **Domain mismatch and training data.** Existing German BERT-based sentiment models are primarily trained on user-generated content such as tweets, online posts, and product or movie reviews, rather than on lyrics or poetry (Bello et al, 2023; Guhr et al., 2020). Across languages, work that targets song lyric relies on transfer learning from other domains: for instance, emotion models for pop lyrics pre-train on large generic or social-media corpora and are then fine-tuned on comparatively small lyric datasets (Dahary et al, 2025). To our knowledge, there are currently no widely used German transformer-based sentiment models that are trained directly on lyrics, and we are not aware of systematic evaluations of such models. Where literary texts are analyzed, authors typically highlight data scarcity and therefore rely on cross-domain transfer or on lexicon-based methods adapted to the target domain (Öhman, 2021; Fehle et al., 2021).
2. **Interpretability.** Our aim is to relate sentiment to lexical-syntactic features. Lexicon-based scores make this mapping explicit; neural models are less transparent.
3. **Cross-temporal and cross-linguistic comparability.** Lexicon-based approaches can more easily be harmonized across decades and languages in future work, whereas neural models may introduce additional diachronic and domain-specific biases.
4. **Expected gains relative to task difficulty.** As shown in Section 2.5.2, even human annotators exhibit only moderate agreement at the aggregate level, with most disagreements involving adjacent categories rather than polarity reversals. This suggests that

sentiment in song lyrics is often inherently ambiguous or underspecified. In such settings, more complex models may yield only limited improvements over simpler approaches.

We therefore treat SentiMerge as a robust, interpretable baseline for song-level affect. At the same time, the gold-standard annotations created in this study provide a useful testbed for future work: transformer-based classifiers or prompting-based large language models could be evaluated systematically against human judgments to assess whether they yield measurable improvements or different patterns of disagreement. Future work, as discussed in the conclusion, should test whether such approaches produce different diachronic patterns or stronger associations with structural features.

2.5 Evaluation of automatic annotation (RQ1)

RQ1 concerns the quality and reliability of the automatic methods used to derive our most error-prone key features, namely modal particles and sentiment intensity. We therefore conduct two dedicated evaluations. Other features (lexical diversity measures and pronoun counts) are based on relatively robust, well-studied procedures (tokenization, lemmatization, POS tagging, dictionary lookup) and are therefore used descriptively without additional task-specific validation in this study.

2.5.1 Modal particle gold standard and CRF models

We construct a gold standard based on a stratified sample of 1,400 instances of the 14 target word forms (100 per form), drawn from Songkorpus. Each example is presented with surrounding context (typically one sentence to the left and right), and four native speakers independently annotate whether the candidate functions as an MP or not, following detailed guidelines with positive and negative examples.

Inter-annotator agreement is substantial to almost perfect (Landis and Koch, 1977): pairwise Fleiss' κ values range from 0.64 to 0.83, with most above 0.75. Certain particles present systematic challenges; *mal*, for example, is highly polyfunctional (emphasis, softening, temporal adverbial uses), and the lack of prosodic cues in written data increases variability. Disagreements are resolved by majority vote; ties are adjudicated by a trained annotator, yielding a curated gold standard.

Building on this resource, we train Conditional Random Field (CRF) models using CRF++ (Kudo, 2005) for token-level MP classification (MP vs. non-MP). Features include token form, lemma, and STTS tags within variable context windows. We compare three configurations:

- Reduced model (narrow context ±2, basic lemma and POS features).
- Original CRF++ model (same context, richer feature conjunctions).
- Extended model (broader context ±3).

Contrary to expectations, the reduced model performs best, correctly identifying 202 of 286 MPs in the test set and achieving high precision and recall for particles such as *etwa*, *halt*, and *wohl* (>90%). Performance for more polyfunctional items (*ja, vielleicht, schon*) is weaker, mirroring human disagreement patterns. A simple wordform baseline that classifies all occurrences of the 14 forms as MPs would perform reasonably for some particles but poorly overall, underscoring the benefit of context-sensitive modeling.

The close alignment between human variability and model performance suggests that the limitations of automated MP identification partly reflect ambiguity inherent in the data. Overall, the reduced CRF model provides sufficient quality for corpus-level modeling, with known weaknesses explicitly acknowledged.

2.5.2 Sentiment annotation experiment

To assess the fundamental quality of SentiMerge-derived sentiment scores, we conduct an annotation experiment that compares system scores with human judgments and tests intra-rater stability (Abercrombie et al., 2023).

We select 20 songs: five that SentiMerge classifies as very positive, five as very negative, five with moderately positive scores, and five with moderately negative scores (based on corpus-level means). Seven raters classify each song twice, separated by several days, using four ordered categories: very negative, rather negative, rather positive, very positive. This yields 280 decisions (20 songs × 7 raters × 2 sessions).

We encode the categories on an ordinal scale that treats differences as distances: one point for adjacent categories, two points when one category lies in between, and three for extremes, positive versus negative. This allows us to quantify both categorical and graded disagreements.

Three complementary evaluations are conducted:

1. **Inter-rater reliability.** Krippendorff's α across raters is around 0.30 for both sessions, indicating modest aggregate agreement. Average pairwise weighted Cohen's kappa κ is substantially higher (≈0.71 – 0.72), suggesting that most rater pairs are consistent and that the low α reflects outliers.
2. **System-human alignment.** Weighted κ between SentiMerge and individual raters ranges from ≈0.52 to 0.76 (moderate to

substantial agreement), with accuracy scores between 0.5 and 0.7 and Mean Absolute Error (MAE) between 0.5 and 0.8. Disagreements mainly involve neighboring categories rather than opposites, indicating that the system rarely mistakes strongly positive for strongly negative songs or vice versa.

3. **Intra-rater reliability.** Weighted kappa between each rater's two sessions is very high (0.92 – 1.00), with most values ≥0.96, indicating that individual judgments are highly stable over time.

Taken together, these results confirm that SentiMerge-based scores operate within the same reliability range as human judgments for the selected subset. They provide a justified basis for using sentiment intensity as a corpus-level measure, while acknowledging that subtle nuances, ironic uses, or complex metaphors remain challenging.

3. Empirical Results

Having established the robustness and limitations of the feature extraction procedures, the now following section applies these methods to the diachronic analysis of German pop lyrics.

Given that Section 2 offers a largely positive answer to RQ1 for modal particles and sentiment intensity, we now turn to RQ2 and RQ3. First, we examine how the four feature families co-vary (correlations and regression models), then we model their temporal development.

3.1 Associations between structural and affective features (RQ2)

We compute Spearman's ρ correlations among lexical diversity (MATTR, MTLD), sentiment intensity (SI), modal particle ratio (PTKM), and pronoun index (PRON). The resulting matrix shows the following patterns:

- **Lexical diversity.** MATTR and MTLD are only weakly correlated (≈0.07), suggesting that they capture different aspects of diversity in this corpus. This may reflect the fact that MATTR is sensitive to local repetition (e.g. choruses), whereas MTLD is driven by global variation.
- **Sentiment intensity.** SI correlates strongly and positively with MATTR (≈0.70), suggesting that lexically more diverse songs more densely mobilize sentiment-bearing words. SI also shows a moderate positive association with PTKM (≈0.29), implying that MP-rich texts tend to be more affectively loaded.
- **Modal particles.** MATTR shows a moderate positive association with PTKM (≈0.40), indicating that texts with richer local vocabularies also tend to exhibit higher modal particle density.
- **Pronouns.** PRON shows only weak correlations with all other variables, indicating that pronoun usage is largely independent of lexical diversity, text length, sentiment intensity, and MP usage.

Figure 1 visualizes all associations, highlighting weak, moderate, and strong correlations.

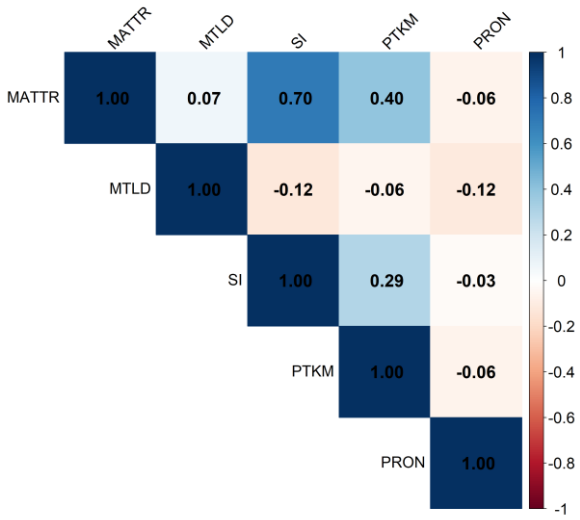

Figure 1: Correlation Heatmap

The correlation patterns motivate a series of multiple regression models that treat each feature in turn as a dependent variable. We report here the main patterns rather than every coefficient.

- **Predicting PTKM from MATTR, SI, and PRON.** The model is statistically significant but explains only about 10% of the variance ($R^2 \approx 0.10$). MATTR is the dominant positive predictor; SI is statistically significant but substantively negligible; PRON does not contribute meaningfully. This suggests that MP density is associated with lexical richness but may influenced by additional, unmodeled factors such as genre or artist style.
- **Predicting SI from PTKM, MATTR, and PRON.** This model has substantially greater explanatory power ($R^2 \approx 0.53$). Both MATTR and PTKM are positive and strong predictors, and the overall model accounts for over half of the observed variation in SI, which is notable for corpus-linguistic and social-science data. PRON is also a statistically significant positive predictor, but its effect size is small relative to MATTR and PTKM. These results indicate that songs with higher lexical diversity and higher MP density are more emotionally intense.
- **Predicting MATTR from PTKM, SI, and PRON.** The model explains roughly 54% of variance ($R^2 \approx 0.54$). PTKM is the strongest

positive predictor, pointing to a close coupling between MP use and lexical richness. SI is statistically significant but substantively small once scaling is considered. PRON shows a small negative association with MATTR: texts with higher first-/second-person prominence tend to be slightly less lexically diverse.

- **Predicting PRON from PTKM, SI, and MATTR.** This model has negligible explanatory power ($R^2 \approx 0.005$). Some predictors are statistically significant due to the large N, but their effect sizes are trivial. Pronoun variation appears driven by factors outside the present feature set (e.g. narrative perspective, genre conventions, artist-specific style).

Taken together, these results indicate a robust triangle of associations linking lexical diversity (MATTR), modal particles (PTKM), and sentiment intensity (SI): lexically richer texts tend to use more MPs and exhibit higher affective density, and MP-rich texts tend to be more emotionally intense. Pronouns play a marginal role in this structural-affective nexus.

Regarding lexical diversity measures, the divergence between MATTR and MTLD implies that not all diversity metrics are equally informative for lyrics with strong repetition and chorus structures. Our key RQ2 findings are therefore conditional on MATTR's behavior as a locally sensitive measure; MTLD appears less aligned with sentiment intensity in this genre, and alternative diversity metrics may yield partly different patterns. We treat this as a limitation and an avenue for future research rather than as evidence that one measure is universally superior.

3.2 Temporal trends (RQ3)

To address RQ3, we fit separate linear models with publication year as the predictor and each feature as the dependent variable. These models are intentionally simple; they are not intended as causal models but as first-pass summaries of diachronic trends.

1. **Lexical diversity (MATTR ~ YEAR).** We find a significant negative effect of year on MATTR ($\beta = -0.003857$, $p < 2e\text{-}16$, $R^2 \approx 0.21$). Lexical diversity decreases steadily over time, with modern songs using relatively simpler vocabularies. In relative terms, this corresponds to an approximate decrease of 1 – 2% per year, though precise percentage estimates depend on scaling and should be interpreted with caution. This pattern aligns with earlier findings for English lyrics (Parada-Cabaleiro et al., 2024), suggesting a cross-linguistic trend towards simplification.
2. **Modal particle density (PTKM ~ YEAR).** PTKM shows a small but highly significant negative slope ($\beta = -0.0001675$, $p < 2e{-}16$), with year explaining about 7 – 8% of the variance ($R^2 \approx 0.075$). MPs occur slightly less often in more recent songs, consistent with a shift towards more direct or pragmatically "lighter" language. Given the low mean and modest R^2, this trend should be interpreted as a weak but reliable tendency rather than a dramatic change.
3. **Sentiment intensity (SI ~ YEAR).** Sentiment intensity declines significantly over time ($\beta = -6.092e{+}10$, $p < 2e\text{-}16$, $R^2 = 0.14$). Relative-change estimates suggest a stronger proportional decrease than for MATTR or PTKM, though these percentages are sensitive to the absolute scale of SI. Substantively, the model indicates that songs have become somewhat less affectively dense, even if polarity remains relatively stable.
4. **Pronoun index (PRON ~ YEAR).** The model indicates a small increase in first-/second-person pronoun prominence over time ($\beta = 0.040649$, $p = 3.13e\text{-}7$, $R^2 = 0.0039$), but with very low explanatory power. While statistically robust, the effect is tiny and by itself insufficient to support strong claims about increased personalization. We therefore interpret it cautiously, as a weak trend that complements the structural changes observed in MATTR, PTKM, and SI.

Overall, the diachronic analyses suggest that, within the examined feature space, contemporary songs exhibit measurable temporal change: lexical diversity, modal particle density, and sentiment intensity all decline, while direct address (PRON) slightly increases (Figure 2). Together, these trends point towards a gradual simplification and mild "flattening" of emotional intensity, coupled with modest increases in subjectivity or listener orientation. However, given the modest R^2 values and the possibility of genre, artist, and topic effects, these interpretations should be seen as indicative rather than definitive.

A descriptive breakdown of personal pronouns in the corpus shows a clear dominance of first- and second-person forms. The most frequent item by a large margin is *ich* (255,260 instances), followed by *du* (126,909), while other pronouns such as *wir* (52,525), *es* (46,574), and *sie* (33,579) occur considerably less often. This distribution suggests that song lyrics in the corpus are strongly oriented toward self-reference and direct address, with a particular emphasis on the speaker's perspective. However, since these figures are aggregated across the entire time span, they do not by themselves indicate which pronouns drive the observed diachronic increase in the PRON index. A more fine-grained temporal analysis of individual pronouns is therefore left for future work.

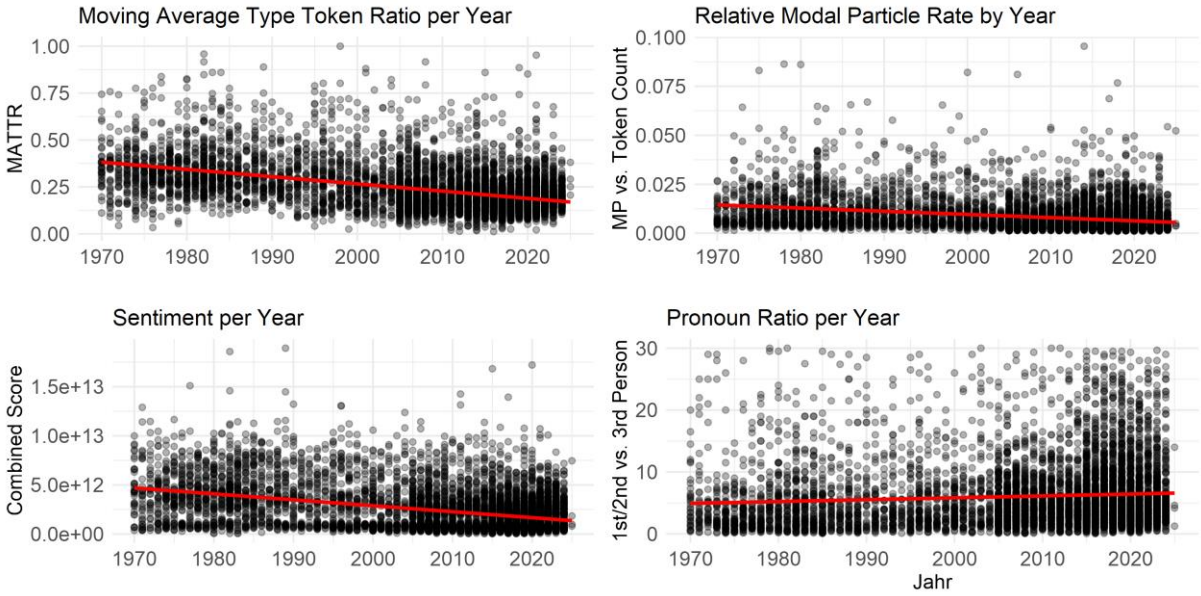

Figure 2: Linear Model Scatterplots of Temporal Trends in Sentiment and Linguistic Features

4. Conclusion and Outlook

This study has presented a multidimensional, empirically grounded analysis of linguistic and emotional patterns in Ger man pop lyrics across five decades, based on a large, longitudinally maintained and richly annotated corpus. Using Songkorpus, it has been sho wn how features such as lexical diversity, pronoun usage, modal particle density, and sentiment intensity can be operationalized and combined to explore how structural and affective dimensions co-vary over time. The resulting patterns point to potentially meaningful developments, including shifts in affective expression. However, the extent to which such patterns can be interpreted as evidence of broader cultural tendencies – such as forms of emotional flattening – remains necessarily limited. In particular, the interaction between linguistic form, genre conventions, and modeling assumptions makes it difficult to disentangle cultural change from representational and methodological effects. The analysis therefore provides a structured empirical perspective on these dynamics rather than definitive claims about their cultural significance.

Since the analytical payoff of this study lies in making core dimensions of lyrical communication measurable at corpus scale, the methods, models, and evaluation protocols are made openly available on the Songkorpus website[1], establishing a transparent foundation for replication and enabling extensions across languages, genres, and communicative contexts.

From a methodological perspective, the study demonstrates that the automatic identification of linguistically complex features such as modal particles and sentiment intensity can achieve a level of reliability sufficient for exploratory large-scale analysis when supported by targeted validation. The construction of a dedicated gold standard for modal particles, together with a focused sentiment annotation experiment, indicates that model performance approaches the range of human agreement observed for this type of data. At the same time, these findings underline the importance of cautious interpretation: even where annotation quality is comparatively high, the inferential step from measured linguistic features to broader claims about affect or cultural change remains non-trivial. Accordingly, the primary contribution of this study lies in demonstrating how such features can be modeled and evaluated in a challenging domain, thereby enabling more robust future investigations of creative, non-standard language.

Substantively, the analysis identifies robust associations between lexical richness, modal particle density, and sentiment intensity: lexically more diverse songs tend to use more modal particles and exhibit higher affective density. Diachronically, lexical diversity and modal particle use decline, while sentiment intensity also decreases and first- and second-person pronoun prominence increases slightly. These trends suggest a gradual shift towards simpler, somewhat less emotionally charged, and modestly more personalized language in German pop lyrics. They resonate with earlier findings for

[1] https://songkorpus.de/data/

English lyrics and invite further cross-linguistic comparison.

Several caveats and directions for future work follow from our findings, especially with regard to how feature design and validation can be integrated into corpus-processing pipelines to make large, evolving corpora more robustly analyzable over time:

- Our measure of "emotional dynamics" is based on global song-level sentiment intensity and does not capture within-song shifts, mixed polarities, or specific emotion categories. Segment-level modeling and multi-dimensional emotion classification would provide a richer picture of affective structure.
- Lexical diversity behaves differently across measures (MATTR vs. MTLD), especially in a genre characterized by strong repetition. Future work should systematically compare diversity metrics in lyrics and other poetic texts to clarify what aspects of variation they capture.
- The present models are correlational and largely linear; they do not establish causal direction. More advanced approaches – such as genre-stratified models, hierarchical or mixed-effects models, or causal modeling frameworks – could better disentangle the roles of genre, artist, and time.
- Our sentiment analysis relies on a lexicon-based baseline. Transformer-based sentiment and emotion models, fine-tuned on lyrics or related domains, could complement or challenge our results. Evaluating such models on large lyric corpora, ideally in combination with extended human annotation, is a promising avenue for future work.
- Finally, broadening the feature set to include syntactic complexity, rhythmic patterns, lexical fields, and discourse structure would further enhance the account of how linguistic form and affective content jointly shape the expressive texture of popular music.

Despite these desiderata and challenges, the present study demonstrates how large-scale, linguistically annotated corpora can reveal subtle relationships between language structure, affective expression, and cultural change. It provides a transparent methodological framework and open materials that are readily extensible to other languages and communicative contexts, while also illustrating how carefully validated automatic annotations enable empirically grounded analyses of culturally specific text types.

From the perspective of large-scale corpus management, the study directly engages with key challenges central to the CMLC agenda. The Songkorpus comprises texts that are "written to be sung" and therefore systematically diverge from the assumptions underlying most corpus-based modeling approaches, such as stable orthography, consistent segmentation, and standardized grammatical structure. This poses particular difficulties for diachronic analysis, as observed variation may reflect both linguistic change and evolving conventions of textual representation.

At the same time, the corpus highlights issues of access and sustainability: as a collection of copyrighted song lyrics, its availability is necessarily restricted, with implications for reproducibility and the transferability of analytical workflows. The approach adopted here addresses these constraints by focusing on derived representations (e.g., aggregated linguistic and emotional features) and by implementing preprocessing and modeling strategies that are robust to non-standard variation. In this way, the study demonstrates how creative, performance-oriented language can be systematically integrated into large-scale corpus analyses.

5. Bibliographical References

Abercrombie, G., Rieser, V., and Hovy, D. (2023). Natural Language Processing with Intra-Annotator Agreement. *ArXiv*, Preprint 2301.10684. https://arxiv.org/pdf/2301.10684

Bello, A., Ng, S.-C., and Leung, M.-F. (2023). A BERT Framework to Sentiment Analysis of Tweets. *Sensors*, 23(1), 506. https://doi.org/10.3390/s23010506

Bestgen, Y. (2024). Measuring Lexical Diversity in Texts: The Twofold Length Problem. *Language Learning*, 74: 638–671. https://doi.org/10.1111/lang.12630

Blühdorn, H. (2019). Modalpartikeln und Akzent im Deutschen. *Linguistische Berichte*, 259. Hamburg: Buske. 275–317. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-91746

Broll, S. and Schneider, R. (2023). Empirische Verortung konzeptioneller Nähe/Mündlichkeit inner- und außerhalb schriftsprachlicher Korpora. *Journal for Language Technology and Computational Linguistics*, 36(1). 113–150. https://doi.org/10.21248/jlcl.36.2023.240

Covington, M. and McFall, J. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17, 94–100. https://doi.org/10.1080/09296171003643098

Dahary, S., Edana, A., Apartsin, A., and Aperstein, Y. (2025). From Joy to Fear: A Benchmark of Emotion Estimation in Pop Song Lyrics. *ArXiv* Preprint 2509.05617. https://arxiv.org/pdf/2509.05617

Diewald, G. (2007): Abtönungspartikel. In Hoffmann, L. (Ed.), *Handbuch der deutschen Wortarten*. Berlin, New York: de Gruyter. 117–142. https://doi.org/10.1515/9783110217087.117

Emerson, G. and Declerck. T. (2014). SentiMerge: Combining Sentiment Lexicons in a Bayesian Framework. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, 30–38, Dublin, Ireland: Association for Computational Linguistics and Dublin City University. https://doi.org/10.3115/v1/W14-5805

Fehle, J., Schmidt, T., and Wolff, C. (2021). Lexicon-based Sentiment Analysis in German: Systematic Evaluation of Resources and Preprocessing Techniques. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, Düsseldorf. https://doi.org/10.5283/EPUB.50833

Guhr, O., Schumann, A.-K., Bahrmann, F., and Böhme, H. J. (2020). Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems. In N. Calzolari et al. (Eds.), *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, 1627–1632. Marseille: European Language Resources Association (ELRA). https://aclanthology.org/2020.lrec-1.202/

Hentschel, E. (2010). Partikelprofile literarischer Texte. In T. Harden, E. Hentschel (Eds.), *40 Jahre Partikelforschung*. Tübingen: Stauffenburg, 97–118.

Hunke, T., Huber, F., and Steffens, J. (2025). The Evolution of Song Lyrics: An NLP-Based Analysis of Popular Music in Germany from 1954 to 2022. *Music & Science*, 8. https://doi.org/10.1177/20592043251331155

Koch, P. and 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, and G. Wolf (Eds.), *Communicative spaces: Variation, contact, and change*, 441–473. Frankfurt: Lang. https://doi.org/10.15496/publikation-20415

Kraus, M., Feuerriegel, S. (2019). Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. E*xpert Systems with Applications*, 118, 65-79, https://doi.org/10.1016/j.eswa.2018.10.002

Kreyer, R. and 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

Landis, J. R. and Koch, G. G. (1977): The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174. https://doi.org/10.2307/2529310

Liu, B. (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge, University Press. https://doi.org/10.1017/CBO9781139084789

Malvern, D., Richards, B., Chipere, N., and Durán, P. (2024). *Lexical Diversity and Language Development. Quantification and Assessment*. London : Palgrave Macmillan. https://doi.org/10.1057/9780230511804

Mccarthy, P. and 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. https://doi.org/10.3758/BRM.42.2.381

Öhman, E. (2021). The Validity of Lexicon-based Sentiment Analysis in Interdisciplinary Research. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, 7–12, NIT Silchar, India. NLP Association of India (NLPAI). https://aclanthology.org/2021.nlp4dh-1.2/

Pang, B., Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. Now Foundations and Trends. https://doi.org/10.1561/1500000011

Parada-Cabaleiro, E., Mayerl, M., Brandl, S., Skowron, M., Schedl, M., Lex, E., and Zangerle, E. (2024). Song lyrics have become simpler and more repetitive over the last five decades. *Scientifc Reports*, 14 (5531). https://doi.org/10.1038/s41598-024-55742-x

Schneider, R. (in preparation). Beyond Standard: Intelligent Modeling of Creative Language Using the German Song Corpus. In L. Herzberg, C. Mair, and A. Witt (Eds.), *Corpus linguistics 2040: Which data, which methods, which models?* Digital Linguistics, 6, Berlin/Boston: De Gruyter.

Schneider, R. (2026). Linguistic resources for the study of pop culture. In Valentin Werner, Cecilia Cutler, and Andrew Moody (Eds.), *Handbook of Language and Pop Culture*. Berlin, Boston: De Gruyter Mouton.

Schneider, R. (2020). A Corpus Linguistic Perspective on Contemporary German Pop Lyrics with the Multi-Layer Annotated “Songkorpus”. In N. Calzolari et al. (Eds.), *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, 835–841. Marseille: European Language Resources Association (ELRA). https://aclanthology.org/2020.lrec-1.105/

Schoonjans, S. (2018). Modalpartikeln als multimodale Konstruktionen. Berlin/Boston: deGruyter. https://doi.org/10.1515/9783110566260

Scott, M. (2004). WordSmith Tools Version 4.0. Oxford: Oxford University Press.

Storø, S. R. (2022). Die Annotation der Modalpartikeln im GeWiss-Korpus. Eine syntaktische und semantisch-pragmatische Analyse der PTKMA-Annotation. *Deutsche Sprache. Zeitschriften für Theorie, Praxis und Dokumentation*, 22 (2), 124–149. https://doi.org/10.1515/ds-2022-0014

Sluyter-Gäthje, H. and Trilcke, P. (2022). Poesie als Fehler. Ein 'Tool Misuse'-Experiment zur Prozessierung von Lyrik. In *Proceedings 8. Tagung des Verbands Digital Humanities im deutschsprachigen Raum (DHd)*, Potsdam. https://doi.org/10.5281/zenodo.6328201

Trippel, T. (2025). Metadata for research data. In P. Bański, U. Heid, and L. Herzberg (Eds.), *Harmonizing language data: Standards for linguistic resources*. Digital Linguistics, 4, Berlin/Boston: De Gruyter 251–279. https://doi.org/10.1515/9783112208212-011

Werner, V., Hiramoto, M., and Flanagan, P. (2025). Language and pop culture. Setting the agenda, *Journal of Language and Pop Culture*, 1(1). 1–17. https://doi.org/10.1075/jlpop.24034.wer

6. Language Resource References

Kudo, T. (2005). *CRF++: Yet Another CRF Toolkit.* http://taku910.github.io/crfpp/

Schneider, R. (2025). *Songkorpus - Linguistic Corpus of German Song Lyrics*. Release 7.0. https://songkorpus.de