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 $\mathrm{KLD}_{\mathrm{disp}}$ 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 $\mathrm{KLD}_{\mathrm{freq}}$ value of 0.81 and a difference in $\mathrm{KLD}_{\mathrm{disp}}$ values of -0.001, respectively. The very high frequency of the word compared to the reference corpus is reflected in a very high $\mathrm{KLD}_{\mathrm{freq}}$ result. At the same time, the vanishingly small difference in $\mathrm{KLD}_{\mathrm{disp}}$ 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).

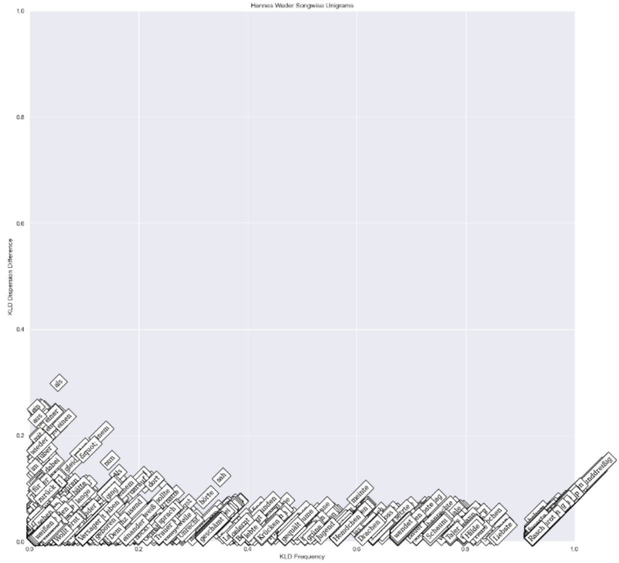

Figure 1: KLD Keyword analysis for Hannes Wader.

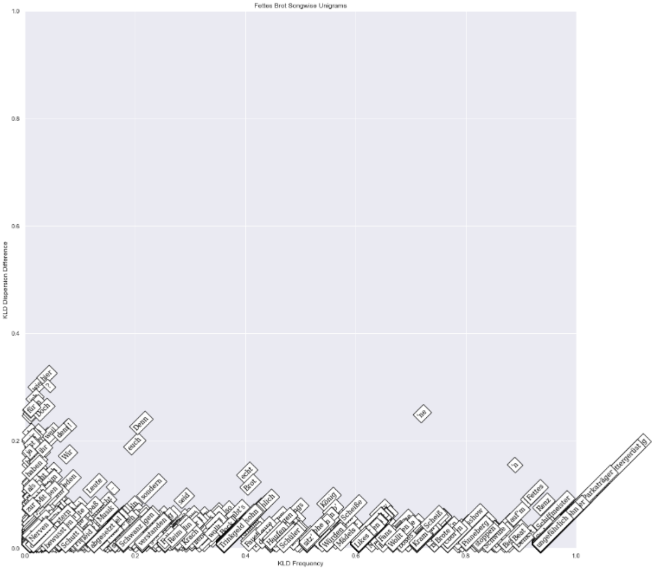

Figure 2: KLD Keyword analysis for Fettes Brot.

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.

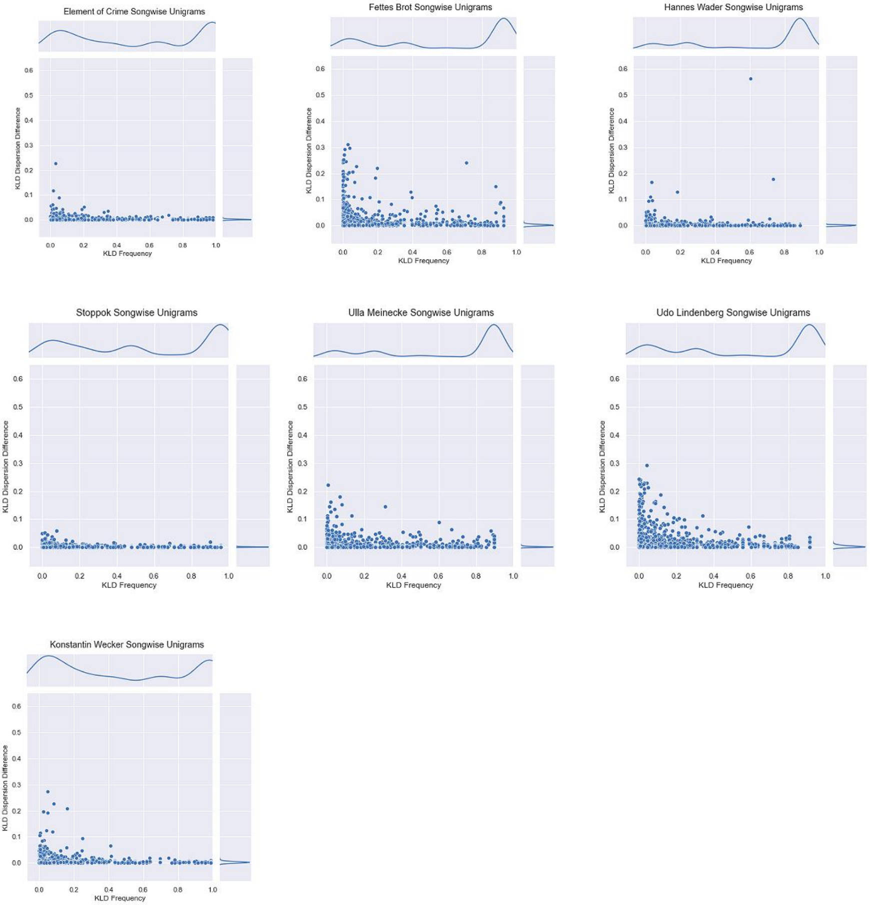

Figure 3: KLD-Keyword analysis for all artists in the corpus (dots represent words).

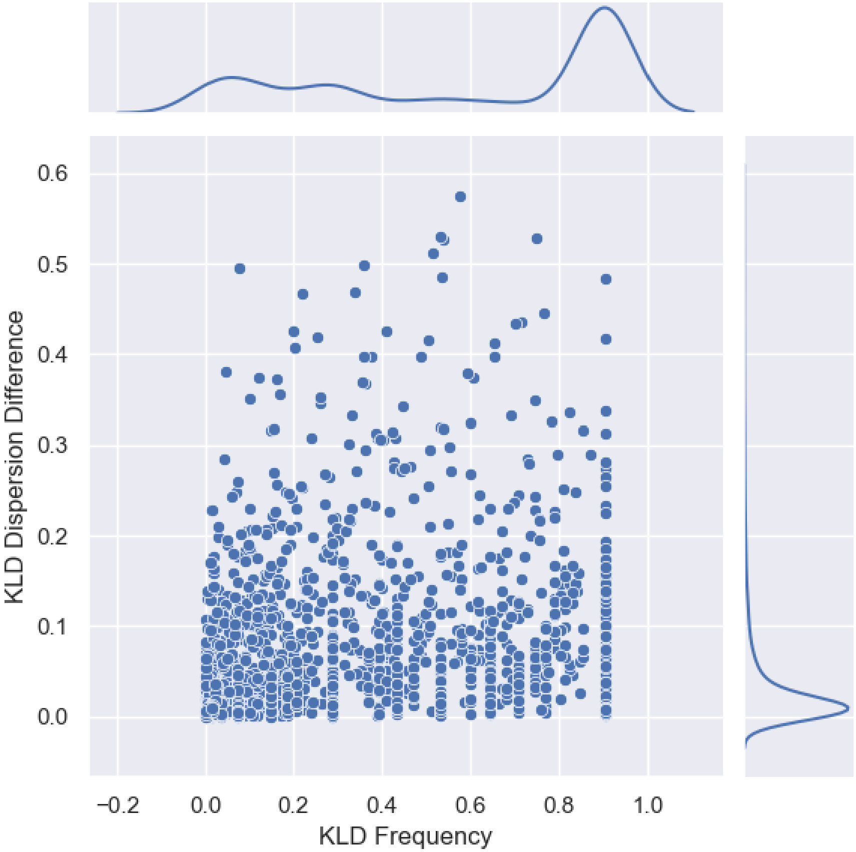

Figure 4: KLD-Keyword analysis replication of Gries' (2021) results including marginal plots for the Clinton-Trump Corpus (Brown 2016). Minor differences due to a different method of tokenization.

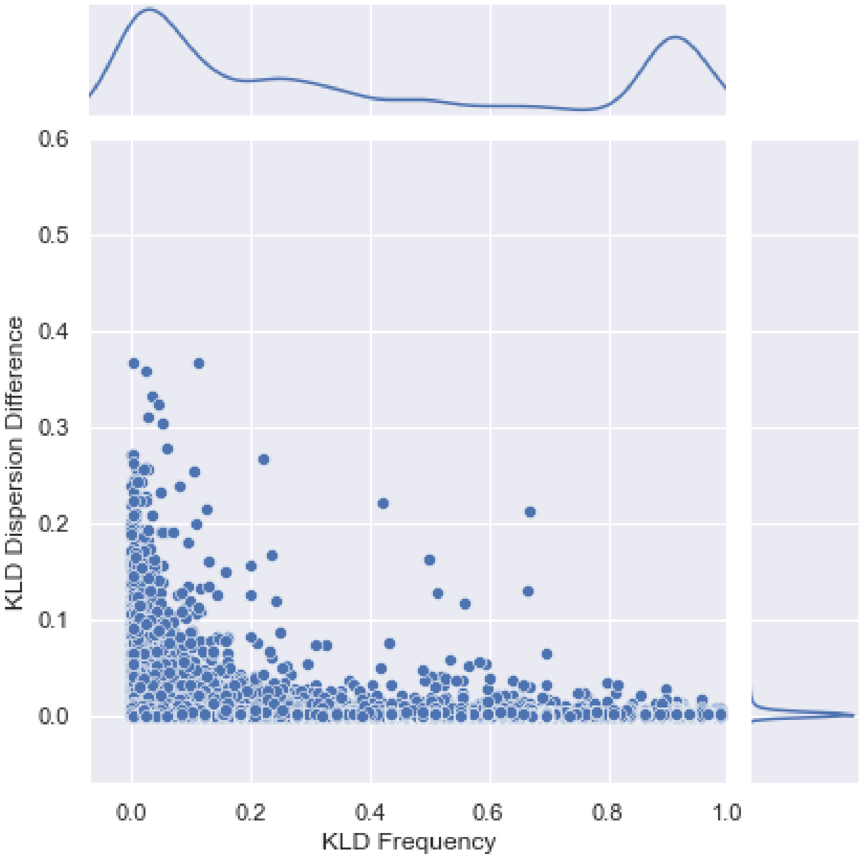

Figure 5: KLD POS Key-trigram analyses for all artists in the corpus (dots represent trigrams).

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.

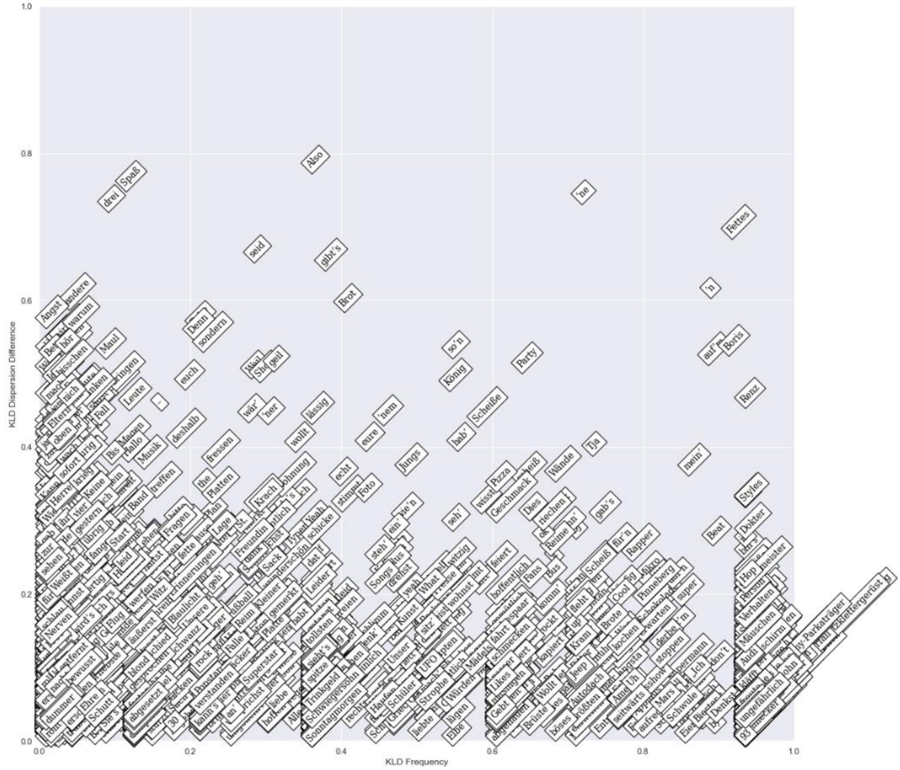

Figure 6: Albumwise KLD Keyword analysis for Fettes Brot.

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.

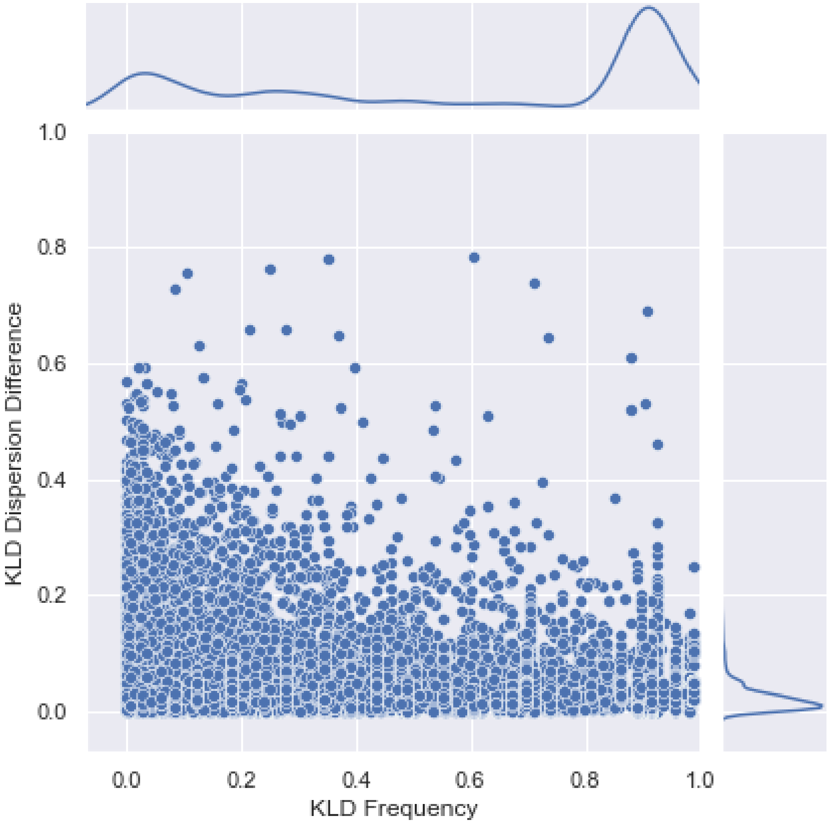

Figure 7: Albumwise KLD Keyword analysis for all artists plotted on top of each other.

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