Originally published in: Cook, Paul/Mitrović, Jelena/Parra Escartín, Carla/Vaidya, Ashwini/Osenova, Petya/Taslimipoor, Shiva/Ramisch, Carlos (Eds.): Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021). – Stroudsburg: Association for Computational Linguistics, 2021. Pp. 13-22.
DOI: *https://doi.org/10.18653/v1/2021.mwe-1.3*

# Data-driven Identification of Idioms in Song Lyrics

**Miriam Amin**[1], **Peter Fankhauser**[2], **Marc Kupietz**[2], **Roman Schneider**[2]

[1]Leipzig University, Leipzig, Germany
miriam_amin@web.de
[2]Leibniz Institute For The German Language, Mannheim, Germany
{fankhauser|kupietz|schneider}@ids-mannheim.de

## Abstract

The automatic recognition of idioms poses a challenging problem for NLP applications. Whereas native speakers can intuitively handle multiword expressions whose compositional meanings are hard to trace back to individual word semantics, there is still ample scope for improvement regarding computational approaches. We assume that idiomatic constructions can be characterized by gradual intensities of semantic non-compositionality, formal fixedness, and unusual usage context, and introduce a number of measures for these characteristics, comprising count-based and predictive collocation measures together with measures of context (un)similarity. We evaluate our approach on a manually labelled gold standard, derived from a corpus of German pop lyrics. To this end, we apply a Random Forest classifier to analyze the individual contribution of features for automatically detecting idioms, and study the trade-off between recall and precision. Finally, we evaluate the classifier on an independent dataset of idioms extracted from a list of Wikipedia idioms, achieving state-of-the art accuracy.

## 1 Introduction

Traditional accounts of idiomaticity distinguish idiomatic use of language from literal use, claiming that idioms are multiword expressions (MWEs) which do not conform to Frege's principle, i.e. whose meaning as a whole cannot fully be derived from the aggregated meaning of their components (Gibbon, 1982). In other words, the definition refers to non-compositionality and non-transparency – idiomatic MWEs seem semantically opaque; Baldwin and Kim (2010) consider this "lexical idiomaticity" to be one of five sub-types of idiomacity. Classifying idioms is not trivial: With reference to recent findings in discourse analysis and psycholinguistics, Wulff (2008) describes idiomaticity as a non-binary, multifactorial concept for a "continuum ranging from clearly non-idiomatic patterns to core idioms"; Pradhan et al. (2018) support this observation experimentally. At least core idioms are considered to be (mentally) lexicalized: Schneider et al. (2014) describe them as "lexicalized combinations of two or more words" which, though often syntactically diverse, "are exceptional enough to be considered as individual units in the lexicon". This corresponds to Sinclair's idiom principle (Sinclair, 1991), postulating that text is often constructed from ready made phrases. Due to morphological and syntactic variation, the degree of formal fixedness ranges from semi- to fully-fixed. However, idiomaticity should be corpus-based verifiable, as e.g. Gries (2008, p. 22) states that "researchers interested in phraseologisms use frequencies and other more elaborated statistics" to identify "symbolic units and constructions". Some of these statistics may relate to local contexts, because one can reasonably argue that words that are not used literally will probably be somehow surprising in their context.

Against this background, we regard idioms as a subcategory of MWEs that are conspicuous in function, form and distribution – and with fuzzy boundaries to other multiword units like metaphors (Stefanowitsch and Gries, 2007) or proverbs. Our objective is to cover idiom characteristics with an innovative set of quantitative features, taking up some ideas described in the subsequent section, and to apply and evaluate machine-learning classifiers for a presumable idiomatically rich specialized corpus.

## 2 Related work

Idioms are a key concern and pose challenging problems for NLP applications such as information extraction, retrieval, summarization and translation, as well as for lexicographical studies or lan-



*Proceedings of the 17th Workshop on Multiword Expressions*, pages 13–22
Bangkok, Thailand (online), August 6, 2021. ©2021 Association for Computational Linguistics

*Publikationsserver des Leibniz-Instituts für Deutsche Sprache*
*URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-106825*

guage learning; see Constant et al. (2017). Sag et al. (2002) refer to them as "a pain in the neck for NLP"; consequently their machine-supported recognition constitutes an ideal testbed for a variety of methodical approaches and is the subject of shared tasks; see, e.g., Markantonatou et al. (2020).

Fazly and Stevenson (2006) propose measures that quantify the degree of lexical and syntactic fixedness. Verma and Vuppuluri (2015) rely on lexical features in order to identify MWEs whose meanings differ from their components' meanings. Sporleder and Li (2009) include the collocational contexts of idiomatic MWEs into their computation; they model semantic relatedness with the help of lexical chains and cohesion graphs, and, based on this, compare supervised with unsupervised approaches for token-based idiom classification. Katz and Giesbrecht (2006) use latent semantic analysis in order to verify whether context word vector-similarity between idiomatic MWEs and its constituents helps with the calculation. Muzny and Zettlemoyer (2013) achieve a precision level of 65% for the distinction between idiomatic and literal wiktionary phrases, using lexical and graph-based features in order to quantify the assumption that literal phrases are more likely to have closely related words in their definition clause than idiomatic phrases. Salton et al. (2016) investigate whether Sentential Distributed Semantics of idiomatic verb-noun (VN) combinations show significant differences from non-idiomatic usage, and therefore train Sent2Vec models for sentence-level contexts. Using the same dataset, Peng et al. (2018) compute local context differences between word vector matrices on the basis of Frobenius norm. Senaldi et al. (2019) train vector-based models on a gold standard of VN constructions that has been annotated regarding idiomaticity on a 1-7 Likert scale. Hashempour and Villavicencio (2020) use contextualized word embeddings in order to distinguish between literal and idiomatic senses of MWEs that are treated as individual tokens in training and testing, producing average F1-scores of more than 70%.

We take up the idea of evaluating different context representations, expand corresponding measures with syntagmatic and other statistical features, and analyze how they complement each other to characterize idioms. Furthermore, we broaden the scope by extending the dataset beyond VN combinations, including all kinds of MWEs without morphosyntactic restrictions.

## 3 Dataset and features

The aim of this study is to evaluate quantitative features of MWEs with regard to their suitability of detecting idiomatic MWEs in a given text corpus. Contemporary pop song lyrics – a yet sparsely examined register – seem intrinsically promising for two reasons: Firstly, lyrics combine qualities of spoken and written language (Werner, 2012) with wordplay creativity (Kreyer, 2012) and can thus be expected to constitute a valuable source of both well-known and innovative idiomatic constructions. Secondly, on account of their formal structure, catchy and often idiomatic phrases tend to be repeated in choruses, so that there should be good prospects for empirical evidence. We use the freely available Corpus of German Song Lyrics (Schneider, 2020), covering a period of five decades and a broad range of artists, in order to ensure that our findings can be reproduced and compared by future studies. The general approach should also be applicable to languages other than German.

Although the corpus comes with XML-coded multi-layer annotations, we mainly work on the raw data and do not rely on linguistic preprocessing like parsing or lemmatization. To avoid reference to lexica or pre-defined syntactic template lists (like V-NP constructions), we include any ngram, spanning a minimum of two word tokens and a maximum of six word tokens within sentence boundaries. This yields a dataset of more than six million ngrams. From these we randomly select a sample of 10,000 ngrams.

This dataset is manually annotated by a native speaker in order to serve as a gold standard. To cope with the abovementioned fact that idiomatic status cannot always be described as either clearly idiomatic or clearly literal, we allow for three categories and mark idiom candidates as either literal, idiomatic, or partly idiomatic, where the latter comprises ngrams with both idiomatic and non-idiomatic content, which are excluded for our analysis, see Table 4 in Section 4, for exact numbers.

As a starting point for our evaluation, each dataset entry is automatically annotated with a number of features. We distinguish between three main groups of features to characterize idioms, for a detailed break down see Table 5.

Syntagmatic features (SY) measure collocation strength between all word pairs within an idiom candidate. Context features (CO) measure semantic similarity between the words within an idiom can-

didate and the words in its left/right context. Finally, other features (O) represent a variety of counts to assess the amount of evidence available, such as number of words in an idiom candidate.

SY_C1 and SY_C2 comprise a number of count-based collocation measures between a word and its neighbours within a window of +/- 5[1] (Evert, 2008). SY_C1 are based on the counts in DeReKo (Kupietz et al., 2010), whereas SY_C2 are based on the counts in the pop lyrics corpus. These count-based measures all aim at identifying MWEs that occur more often than randomly expected. We expect that idioms, like other MWEs, are characterized by high SY_C.

SY_W comprises a number of predictive collocation measures. These are all calculated by aggregating the output activations in a three layer neural network using the structured skipgram variant (Ling et al., 2015) of word2vec (Mikolov et al., 2013), again with a window size of +/- 5[2]. As shown by Levy and Goldberg (2014), these output activations approximate the shifted pointwise mutual information[3]. These predictive measures generalize from actually used collocations by means of dimensionality reduction in the hidden layer and thus can also predict unseen but meaningful collocations. However, due to generalization they are typically biased towards the dominant, usually literal usage. Thus, we expect that idioms, unlike other MWEs, are characterized by low SY_W.

Tables 1 and 2 exemplify the interplay between count-based and predictive collocations. Among the top 10 count-based collocates of 'Kuh' (cow), there are 6 collocates (in bold) stemming from idiomatic use, for example, 'die Kuh vom Eis kriegen' literally for 'getting the cow from the ice' meaning 'working out a situation'. In contrast, the predictive collocates all pertain to the literal meaning of cow as a domestic animal; e.g., 'Eis' does not occur among the top 400 predictive collocates.

The count-based and predictive collocates of 'Versuch' ('attempt'), on the other hand, show no such difference. Both refer to the literal meaning of 'Versuch'. However, also here we can observe a bias of the predictive collocates towards 'failed attempts'.

| Kuh | German | English |
|---|---|---|
| Count | Kalles **heilige blöde Blinde Bunte** lila Rosmarie **dumme** Yvonne **Eis** | Kalle's **holy silly blind colorful** purple Rosemary **stupid** Yvonne **ice** |
| Pred | ausgebüxte geschlachtete entlaufene geklonte trächtige geschlachteten weidende verwesende Kalles tote | escaped slaughtered runaway cloned pregnant slaughtered grazing decaying Kalle's dead |

Table 1: Count-based and predictive collocates for Kuh (cow)

| Versuch | German | English |
|---|---|---|
| Count | unternommen gescheitert Beim zweiten gescheiterten wert dritten gestartet unternehmen scheiterte | made failed in second failed worth third started make failed |
| Pred | untauglicher vergeblicher missglückter unternommene krampfhaften fehlgeschlagener (…) | unsuitable futile failed made convulsive failed failed desperate unsuitable desperate |

Table 2: Count-based and predictive collocates for Versuch (attempt)

SY_R comprises non-parametric variants for some collocation measures by means of their ranks to account for the different scales of SY_C1 and SY_W. This includes SY_C1_R, SY_W_R1, SY_W_R2, and the rank difference SY_R_D.

As depicted in Equation 1, for all syntagmatic collocation measures *col*, we take the average over all pairs of words $w_i, w_j$ in an idiom candidate of size $|w|$. Null-values, occurring when there exists no pair with measures from DeReKo, are transformed

[1] All measures with *autofocus* (AF) select those neighbours in the window which maximize the measure.

[2] DeReKoVecs (Fankhauser and Kupietz, 2019, http://corpora.ids-mannheim.de/openlab/derekovecs, accessed 2021-04-23)) has been trained on DeReKo.

[3] $SPMI(w, w_i) = log(\frac{p(w,w_i)}{p(w)p(w_i)}) - log(k)$, with $k$ the number of negative samples used during training, and $p(w)$, $p(w_i)$, $p(w, w_i)$ the individual and joint relative frequencies of a word $w$ and its neighbour $w_i$


Figure 1: Local context of ngrams

to min (or max) values appropriate for each feature.

$$\sum_{i \neq j} col(w_i, w_j)/|w|(|w|-1) \quad (1)$$

The context features CO_VEC and CO_VEC_LEX aim at identifying idioms based on the heuristics that they occur within unusual thematic contexts. Idiomatic ngrams such as 'Perlen vor die Säue werfen' ('cast pearls before swine') are often found in local contexts that are thematically rather untypical for non-idiomatic uses of the individual ngram words. The expression can be expected in a theatre review or a political speech, but rather not in texts explicitly dealing with jewellery or livestock. To this end, CO_VEC uses cosine similarity between word vectors, which identifies paradigmatically related words occurring in similar usage contexts, comprising (near) synonyms, but also hyponyms, meronyms, etc.

Continuing with the above example, among the most similar words for 'Perle' are words like 'Kostbarkeit' ('preciousness'), 'Schatztruhe' ('treasure chest'), 'Liebeserklärung' ('declaration of love') or 'Brosche' ('brooch'). Close to 'Säue', we find 'Rindvieh' ('cattle'), 'Schafe' ('sheep'), 'Köter' ('pooch'), 'Hufe' ('hooves') or 'Schlachtbank' ('slaughterhouse'). Assuming that these words appear less likely in the local contexts of our example idiom than in the typical contexts of its constituents, low value for CO_VEC may indicate idiomatic use.

More specifically, CO_VEC is calculated as the mean cosine similarity between all pairs of words $w_i$ in the idiom candidate of size $|w|$ and words $c_j$ in the left/right context of size $|c|$ (in the present case we include five context words to the left and right[4]; see Figure 1 and Equation 2). CO_VEC_LEX is calculated like CO_VEC, but only takes lexical words into account, i.e. nouns, verbs, adverbs and adjectives. If the idiom candidate appears at several places within the corpus, an average is calculated.

$$\sum_{i,j} sim(w_i, c_j)/|w||c| \quad (2)$$

The last group O comprises O_GRAM, the number of words in an idiom candidate, O_NSTOPW[5], the number of non stopwords, and O_DEREKO, the number of words for which a word embedding is available.

In summary, the syntagmatic features (SY) analyze idiom candidates for frequent (SY_C), but unusual (SY_W) collocations along the syntagmatic axis to assess their phraseness and non transparency. The context features (CO) analyze their surrounding context for unsimilar words along the paradigmatic axis as a complementary measure of non transparency. Both feature sets utilize the observation that word embeddings are typically biased towards the dominant/transparent meaning.

## 4 Methods and results

To evaluate our feature set we have trained a Random Forest classifier [6]. Unless stated explicitly otherwise, all results have been obtained using 5-fold cross validation. To avoid overlap between training and test sets, we have removed all duplicates after lower-casing and stopword removal, leaving a dataset with 542 idioms and 8697 non-idioms.

Because this dataset is highly unbalanced, we have systematically varied the Random Forest's cutoff hyperparameter (default 0.5). As shown in Figure 2, a cutoff of 0.3 achieves the best F1-Score of 61.9%, balancing recall and precision around 62%. The best balanced accuracy of 83% is achieved at a much smaller cutoff of about 0.05. This may be a more appropriate cutoff for explorative idiom detection, where sensitivity (recall) is more important than precision.

To assess the contribution of the individual feature sets, we compare classification performance between using all features, each feature set individually, and subsets of features obtained by excluding individual feature sets.

Table 3 summarizes the results [7]: All individual feature sets except O contribute to classification performance. The biggest contribution comes from the collocation features based on DeReKo counts (SY_C1), followed by the collocation features based on the (much smaller) pop lyrics corpus (SY_C2) and the predictive collocation features SY_W.

[4]Similar measures, applied to context words within sentence boundaries, has been used in Köper and Schulte im Walde (2017) or Kurfalı and Östling (2020) for the detection of non-literal meaning.

[5]SY_C1 and S_W features are calculated on the idiom candidate after stopword removal.

[6]Support Vector Machines yield similar accuracies and scores.

[7]Standard deviation of Balanced Accuracy, measured over 10 5 x cross validations with different seeds is around 0.5 for all feature combinations.

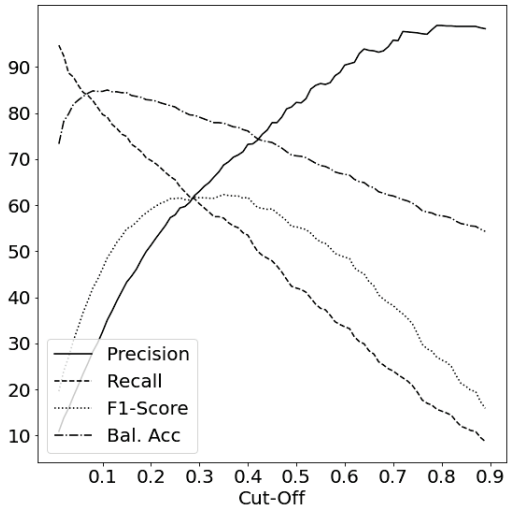


Figure 2: Trade-off curves for Random Forest cut-off

| Feature set | Precision | Recall | F1-Score | Bal. Acc. |
|---|---|---|---|---|
| All features | 62.7 | 59.9 | 61.3 | 78.9 |
| SY_C1 | 44.2 | 38.7 | 41.2 | 67.8 |
| SY_C2 | 32.9 | 30.6 | 31.7 | 63.4 |
| SY_W | 39.2 | 24.9 | 30.3 | 61.3 |
| SY_R | 31.2 | 28.0 | 29.5 | 62.1 |
| CO | 11.8 | 7.4 | 9.1 | 52.0 |
| O | 0.0 | 0.0 | 0.0 | 50.0 |
| w/o SY_C1_R | 55.8 | 48.9 | 52.1 | 73.2 |
| w/o SY_C2 | 60.3 | 53.3 | 56.5 | 75.6 |
| w/o SY_W_R | 61.0 | 58.7 | 59.8 | 78.2 |
| w/o SY_R | 63.0 | 60.9 | 61.9 | 79.3 |
| w/o CO | 59.9 | 60.3 | 60.1 | 78.9 |
| w/o O | 61.0 | 55.9 | 58.3 | 76.8 |

Table 3: Performance of different feature sets in a Random Forest with cutoff=0.3. SY_C1: Count-based collocation measures based on DeReKo. SY_C2: Count-based collocation measures based on pop lyric corpus. SY_W: Predictive collocation measures. SY_R: Rank-based collocation measures. SY_C1_R: SY_C1+SY_R, SY_W_R: SY_W+SY_R. CO: Context features. O: Other

The bottom half of the table analyzes how much performance is lost when excluding a feature set. The relative order is largely consistent with the upper half. In particular, also from this perspective, count-based collocations SY_C1 (including their rank variants) turn out to be most important, i.e., they lead to the largest loss in performance.

Interestingly, omitting the other features (O) also decreases performance, even though they do not contribute individually. This may be due to the fact that they do not model intrinsic characteristics of idioms, but just the number of word pairs available for estimating SY and CO feature sets, i.e., essentially the amount of evidence available. Thus they are only useful in combination with other feature sets.

For SY_R the effect is the other way around. SY_R has a remarkable F1-Score of 29.5% when taken alone, but the overall performance increases, when the classifier is trained on all feature sets but SY_R. The lack of loss in performance may be due to the fact SY_R is highly correlated with SY_C1 and SY_W by construction, and thus does not add information. The slight increase seems to be a random effect.

Table 4 details the classification performance for the best feature set (w/o SY_R). Interestingly enough, when inspecting the false positives, we find that our approach identifies full idioms overlooked by the manual dataset annotation, such as ‘in meine Fußstapfen treten’ (‘follow in my footsteps’) or ‘hinter Gitterstäben’ (lit. ‘behind thick bars’, meaning: ‘in prison’). We also see partly idiomatic MWEs like ‘süßes Gift’ (‘sweet poison’), as well as supposedly incomplete idioms like ‘nur ein leeres [Versprechen?]’ (‘only an empty [promise?]’). The automatic classification even detects previously hidden teenage slang idioms such as ‘Optik schieben’ (lit‘to push optics’, approximately: ‘to be under the influence of hallucinogenic drugs’). Besides, related phenomena like metaphors (‘fahren in Richtung Gold’, literal: ‘drive towards gold’) and allegories (‘das ganze Leben ist ein Quiz’, literal: ‘all of life is a quiz’) are labelled. Indeed, approximately 8% of the false positives show idiomatic or figurative use.

In order to better understand the interplay between features, Table 5 analyzes the contributions of the individual features for the classification task. *MDA* gives the random forest’s estimate of the mean decrease in accuracy per feature, *IGain*

| | | prediction outcome | | |
|---|---|---|---|---|
| | | idiom | no idiom | **total** |
| **actual value** | idiom | 327 | 215 | 542 |
| | no idiom | 191 | 8506 | 8697 |
| | **total** | 518 | 8721 | |

Table 4: Confusion Matrix for prediction with the best feature set

the information gain (*1000), *TTest* the degree of significance by a Welch two sample t-test for confidence levels 0.95 (*), 0.99 (**), and 0.999 (***), and $\Delta$ the sign of the difference between the mean of a feature for idioms vs. non-idioms.

The context features CO_VEC and CO_VEC_LEX have the highest *MDA* followed by the other features O and the count-based collocation features estimated from the pop lyrics corpus SY_C2. All collocation (and rank) features estimated from DeReKo are in a similar range. Note however, that *MDA* tends to be shared among correlated features.

*IGain* assesses the individual (univariate) contribution of the features for classification. The two estimates of the overall frequency of an idiom candidate O_C2_N and O_C2_SGT have the highest *IGain*, closely followed by the count-based collocation features SY_C2 and SY_C1. The predictive collocation features SY_W and context features CO have slightly smaller *IGain*. This largely corroborates the results of the analysis of feature sets above.

With the exception of CO_VEC and two of the predictive collocation features, the difference between the means of all features in idioms vs. non-idioms is highly significant.

To better understand the contribution of the individual features, it is helpful to look at the difference $\Delta$ between their means: Compared to all non-idioms, words within idioms have a lower cosine similarity CO_VEC (but still higher CO_VEC_LEX) to their left and right neighbours, i.e., indeed they occur in unusual contexts. On the other hand, they have a higher count-based and predictive collocation strength among each other (SY_C1, SY_C2, SY_W) with some exceptions (SY_C1_LL,SY_W_CON,SY_W_NSUMAF). Consequently, they also have a smaller rank for these measures (SY_C1_R, SY_W_R1, SY_W_R2), although we would expect larger ranks.

However, non-idioms comprise random ngrams that do not occur more often than expected as well as frequent MWEs with high collocation strength. Thus it is instructive to constrain the comparsion as follows: $\Delta'$ gives the sign of the difference between the mean for idioms and all those non-idioms with SY_C1_LD larger than the mean of SY_C1_LD of all non-idioms, i.e., only the non-idiomatic but still frequent MWEs. Incidentally, all these differences are highly significant (at least 0.99), with the exception of CO_VEC. In this comparison, the context features CO and both, the count-based and predictive collocation features estimated from DeReKo (SY_C1 and SY_W, except SY_C1_MI,) are smaller, and accordingly the corresponding rank features are larger for idioms. In particular, the rank difference SY_R_D between count-based and predictive collocation is larger, i.e., co-occuring words in an idiom tend to be less represented by the predictive collocations which are biased towards the dominant meaning.

In summary, idioms, like non-idiomatic MWEs, are characterized by high collocation strength in comparison to randomly selected ngrams. However, in comparison with non-idiomatic but frequent MWEs, they are characterized by occurring in unusual contexts (low CO_VEC), and by low predictive collocation strength SY_W; or, put more bluntly, idiomatic MWEs occur frequently but are unusual.

To demonstrate the transferability of our approach, we have applied it to a dataset of German idioms extracted from German Wikipedia[8]. After removing duplicates (72) with our gold standard [9], and all idioms that consist of less than 2 words after stopword removal, this set comprises 760 idioms.

As training set for this out-of-domain scenario, we use a sample of 80% of non-idioms and all idioms of our base data set. The test set consists of the remaining 20% of the non-idioms and the Wikipedia idioms. We train the classifier on the feature ensemble SY_C1 + SY_W + SY_R + O (without the feature O_DEREKO). This is because the feature sets SY_C2 and CO are calculated based on

[8] https://de.wikipedia.org/wiki/Liste_deutscher_Redewendungen, accessed February, 22, 2021.

[9] All these duplicates have been independently annotated correctly as idioms.

| Feature | MDA | IGain | TTest | Δ | Δ′ | Description |
|---|---|---|---|---|---|---|
| SY_C1_LD | 9.8 | 30.4 | *** | + | - | logdice (Rychlý, 2008) |
| SY_C1_LDAF | 11.7 | 34.3 | *** | + | - | logdice with autofocus |
| SY_C1_LL | 13.7 | 43.4 | *** | - | - | loglikelihood |
| SY_C1_MI | 19.5 | 48.5 | *** | + | + | (pointwise) mutual information, MI |
| SY_C1_MI3 | 11.7 | 34.8 | *** | + | - | MI³ (Daille, 1994) |
| SY_C2_LD | 20.3 | 19.4 | *** | + | + | logdice in pop lyrics corpus |
| SY_C2_LL | 12.1 | 51.8 | *** | + | + | loglikelihood in pop lyrics corpus |
| SY_C2_MI | 13.6 | 52.8 | *** | + | + | (pointwise) mutual information, MI in pop lyrics corpus |
| SY_C2_MI3 | 11.8 | 51.2 | *** | + | + | MI³ in pop lyrics corpus |
| SY_C2_G | 23.5 | 12.4 | *** | + | + | lexical gravity in pop lyrics corpus (Daudaravičius and Marcinkevičienė, 2004; Gries and Mukherjee, 2010) |
| SY_C2_N | 10.7 | 49.6 | *** | + | + | number of occurrences in pop lyrics corpus |
| SY_C2_SGT | 19.0 | 55.2 | *** | + | + | Simple Good-Turing estimate of probability in pop lyrics corpus |
| SY_W_AVG | 12.7 | 19.0 | * | + | - | average of output activations with autofocus |
| SY_W_CON | 13.9 | 20.5 | *** | - | - | conorm of column normalized output activations with autofocus |
| SY_W_MAX | 10.2 | 11.6 | *** | + | - | max of output activations |
| SY_W_NSUM | 10.6 | 16.7 | | + | - | sum of output activations normalized by total sum over all columns |
| SY_W_NSUMAF | 20.2 | 30.1 | | - | - | sum of output activations normalized by total sum over all selected columns with autofocus |
| SY_C1_R | 16.9 | 53.0 | *** | - | + | rank by SY_C_LD |
| SY_W_R1 | 14.3 | 23.0 | *** | - | + | rank by SY_W_CON |
| SY_W_R2 | 13.9 | 20.5 | *** | - | + | rank by SY_W_NSUM |
| SY_R_D | 18.9 | 55.0 | *** | + | + | rank difference: SY_W_R1-SY_C1_R |
| CO_VEC | 24.3 | 14.4 | | - | - | avg. cosine similarity between words in ngram and words in +/-5 context in pop lyrics corpus |
| CO_VEC_LEX | 20.8 | 13.9 | * | + | - | like CO_WIN5_VEC but only on lexical words |
| O_GRAM | 17.2 | 13.5 | *** | - | - | number of ngram words |
| O_DEREKO | 15.1 | 12.3 | | - | - | number of ngram words available in DeReKo |
| O_NSTOPW | 29.6 | 14.7 | *** | - | - | number of non stop words in ngram |

Table 5: Features

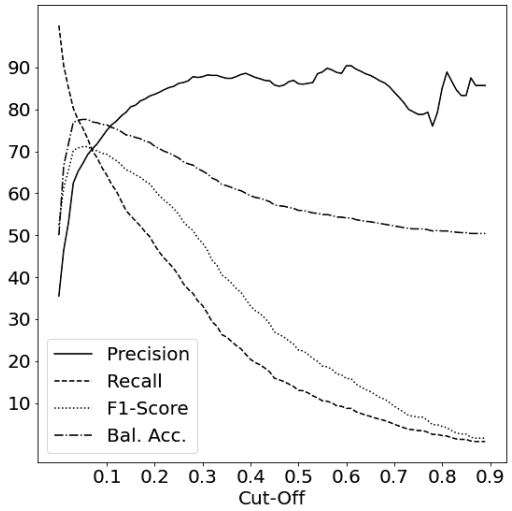


Figure 3: Trade-off curves for Random Forest cut-off on the Wikipedia dataset

| | | **prediction outcome** | | |
|---|---|---|---|---|
| | | idiom | no idiom | **total** |
| **actual value** | idiom | 528 | 232 | 760 |
| | no idiom | 247 | 1135 | 1382 |
| | **total** | 775 | 1367 | |

Table 6: Confusion Matrix for prediction on idioms from Wikipedia with cut-off=0.05

the ngram context within the pop lyrics corpus and are consequently not available for out-of-domain data. Figure 3 shows the trade-off curves of the predictions on the Wikipedia dataset for a range of cut-off thresholds.

The obtained results are rather convincing. With a cutoff threshold of 0.05, the classifier achieves an F1-Score of 71.0% and a recall of 80.3%, which means that the classifier is able to detect the majority of the unknown Wikipedia idioms. While not directly comparable due to different datasets and classification tasks, these results are in the same ballpark as e.g. Hashempour and Villavicencio (2020) who report F1-Scores of 70%.

Table 6 gives the confusion matrix of the prediction on the unknown idioms.

## 5 Conclusions

The aim of this study was to model well-studied idiom characteristics with quantitative features and to evaluate them on suitable datasets. Our evaluations show that count-based collocation measures indeed characterize idioms' frequent usage and stable occurrence, i.e. phraseness. The predictive collocation measures and the context features on the other hand are able to model uncommon usage, that is, non transparency.

By applying our model, trained on an annotated dataset that was sampled from a pop lyrics corpus, to an out-of-domain dataset of idioms crawled from Wikipedia, we demonstrated the generalizability of our approach.

The introduced features do not require sophisticated or knowledge intensive preprocessing, and need only minimal context. Even, when no context is available, as for the out-of-domain dataset, we achieve state-of-the art classification performance.

However, the feature set also has limitations. For idioms that consist of only one content word, possibly with some stopwords, the collocation measures do not produce very meaningful results. In this case we need to entirely rely on the context features. In a similar vein, count based collocation strength obviously does not apply to novel idioms. Moreover, when idiomatic use constitutes the overwhelmingly dominant use, such as 'kenne meine Pappenheimer' (literal: 'know my Pappenheimers', roughly: 'know the weak people (in my team)'), neither CO nor SY_W features can contribute.

But in sum, all evaluation results – and the detailed analysis of how the count-based and predictive features complement each other for discriminating between idioms and non idioms – shed an additional empirical light on the linguistically intricate and multifaceted phenomenon of idiomaticity. Waiving limitations on morphosyntactic templates (like, e.g., VN constructions), our approach should work well for any potentially idiomatic MWEs.

For future work, we intend to apply the approach to bigger datasets; attractive candidates might be the corpora of the PARSEME (PARsing and Multiword Expressions) network Savary et al. (2018) or the COLF-VID dataset of verbal idioms Ehren et al. (2020). We will also experiment with additional features, in particular to better capture fixedness of idiomaticity and cope with non transparent compound idiomatic words.

All data and source code is publicly available

under a Creative Commons license at http://songkorpus.de/data/.

## References


Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*, pages 267–292. CRC Press, Boca Raton.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Béatrice Daille. 1994. *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*. Ph.D. thesis, Paris 7.

Vidas Daudaravičius and Rūta Marcinkevičienė. 2004. Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics*, 9(2):321–348.

Rafael Ehren, Timm Lichte, Laura Kallmeyer, and Jakub Waszczuk. 2020. Supervised disambiguation of German verbal idioms with a BiLSTM architecture. In *Proceedings of the Second Workshop on Figurative Language Processing*, Online. Association for Computational Linguistics.

Stefan Evert. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, chapter 58, pages 1212–1248. Mouton de Gruyter, Berlin, Germany.

Peter Fankhauser and Marc Kupietz. 2019. Analyzing domain specific word embeddings for a large corpus of contemporary German. In *International Corpus Linguistics Conference, Cardiff, Wales, UK, July 22-26, 2019*, Mannheim. Leibniz-Institut für Deutsche Sprache (IDS).

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 337–344.

Dafydd Gibbon. 1982. Violations of frege's principle and their significance for contrastive semantics. *Papers and Studies in Contrastive Linguistics*, 14:5–24.

Stefan Gries. 2008. Phraseology and linguistic theory: A brief survey. In Sylviane Granger and Fanny Meunier, editors, *Phraseology: An interdisciplinary perspective*, pages 3–25. Amsterdam: John Benjamins.

Stefan Gries and Joybrato Mukherjee. 2010. Lexical gravity across varieties of english: An ice-based study of n-grams in asian englishes. *International Journal of Corpus Linguistics*, 15:520–548.

Reyhaneh Hashempour and Aline Villavicencio. 2020. Leveraging contextual embeddings and idiom principle for detecting idiomaticity in potentially idiomatic expressions. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 72–80. Association for Computational Linguistics.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multiword expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia. Association for Computational Linguistics.

Rolf Kreyer. 2012. "Love is like a stove – it burns you when it's hot": A corpus-linguistic view on the (non-)creative use of love-related metaphors in pop songs. *Language and Computers*, pages 103–115.

Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. 2010. The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In *Proceedings of the Seventh International Conference On Language Resources And Evaluation (LREC'10)*, page 1848–1854, Valletta / Paris. European Language Resources Association (ELRA).

Murathan Kurfalı and Robert Östling. 2020. Disambiguation of potentially idiomatic expressions with contextual embeddings. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 85–94, online. Association for Computational Linguistics.

Maximilian Köper and Sabine Schulte im Walde. 2017. Applying multi-sense embeddings for German verbs to determine semantic relatedness and to detect non-literal language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 535–542.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*.

Wang Ling, Chris Dyer, Alan W. Black, and Isabel Trancoso. 2015. Two/too simple adaptations of Word2Vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1299–1304, Denver, Colorado. Association for Computational Linguistics.

Stella Markantonatou, John McCrae, Jelena Mitrović, Carole Tiberius, Carlos Ramisch, Ashwini Vaidya, Petya Osenova, and Agata Savary, editors. 2020. *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Grace Muzny and Luke Zettlemoyer. 2013. Automatic idiom identification in Wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421, Seattle, Washington, USA. Association for Computational Linguistics.

Jing Peng, Katsiaryna Aharodnik, and Anna Feldman. 2018. A distributional semantics model for idiom detection - the case of english and russian. *ICAART*, pages 675–682.

Manali Pradhan, Jing Peng, Anna Feldman, and Bianca Wright. 2018. Idioms: Humans or machines, it's all about context. In *Computational Linguistics and Intelligent Text Processing - 18th International Conference, CICLing 2017, Revised Selected Papers*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 291–304. Springer.

Pavel Rychlý. 2008. A lexicographer-friendly association score. *Proceedings of the 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN*, pages 6–9.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Giancarlo Salton, John Kelleher, and Robert Ross. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 194–204.

Agata Savary, Marie Candito, Verginica Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten Van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke Der, Behrang Qasemi Zadeh, Carlos Ramisch, and Veronika Vincze. 2018. Parseme multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 87–147.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the mwe gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Roman Schneider. 2020. A corpus linguistic perspective on contemporary german pop lyrics with the multi-layer annotated "songkorpus". In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 842–848. European Language Resources Association.

Marco S. G. Senaldi, Yuri Bizzoni, and A. Lenci. 2019. What do neural networks actually learn, when they learn to identify idioms? In *Proceedings of the Society for Computation in Linguistics (SCiL)*, volume 2, pages 310–313.

John Sinclair. 1991. *Corpus, concordance, collocation*. University Press, Oxford.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762.

Anatol Stefanowitsch and Stefan Th. Gries, editors. 2007. *Corpus-Based Approaches to Metaphor and Metonymy*. De Gruyter Mouton.

Rakesh Verma and Vasanthi Vuppuluri. 2015. A new approach for idiom identification using meanings and the web. In *Proceedings of Recent Advances in Natural Language Processing*, pages 681–687, Hissar, Bulgaria.

Valentin Werner. 2012. Love is all around: A corpus-based study of pop lyrics. *Corpora*, 7:19–50.

Stefanie Wulff. 2008. *Rethinking Idiomaticity: A Usage-based Approach*. Studies in Corpus and Discourse. Continuum, London, New York.