Originally published in: Proceedings of the shared task on the disambiguation of German verbal idioms at KONVENS 2021, Düsseldorf, Germany. – Genf: Zenodo, 2021. 9 pp.
DOI: *https://doi.org/10.5281/zenodo.5769519*

Shallow Context Analysis for German Idiom Detection

Miriam Amin
Leipzig University
Leipzig, Germany
miriam_amin@web.de

Peter Fankhauser, Marc Kupietz, Roman Schneider
Leibniz Institute for the German Language
Mannheim, Germany
{fankhauser|kupietz|schneider}@ids-mannheim.de

Abstract

In order to differentiate between figurative and literal usage of verb-noun combinations for the shared task on the disambiguation of German Verbal Idioms issued for KONVENS 2021, we apply and extend an approach originally developed for detecting idioms in a dataset consisting of random ngram samples. The classification is done by implementing a rather shallow, statistics-based pipeline without intensive preprocessing and examinations on the morphosyntactic and semantic level. We describe the overall approach, the differences between the original dataset and the dataset of the KONVENS task, provide experimental classification results, and analyse the individual contributions of our feature sets.

1 Introduction

Idiomatic expressions are generally considered to be (more or less) fixed word combinations that are semantically non-decomposable, i.e. whose overall meaning cannot readily be deduced from the individual constituent semantics and their syntactic structure. Obviously, this definition assumes that for normal (*literal*) multi-word combinations the Frege principle applies, according to which the meaning of a language construct can be explained rationally and results from the meanings of its elements (Gibbon, 1982). Also associated with semantic opacity are questions regarding the clear distinction between idiomatic and non-idiomatic use of expressions. Wulff (2008) highlights the fact that idiomacity should be considered as a rather fuzzy, non-binary concept and that the analysability of idioms is closely related to their different degrees of non-transparency, formal restriction and variance. In analogy, Cook et al. (2007, p. 44) assume "a continuum from completely semantically transparent, or literal, to entirely opaque, or idiomatic". So between core idioms at the one end and clearly non-idiomatic expressions at the other end of a continuous scale, we have to deal with methodologically difficult to treat cases in between. As a first indication, formal fixedness can be helpful, since idiomatic expressions tend to have just a small number of canonical forms with varying degrees of lexical variability and syntactic flexibility; Sinclair (1991, p. 110) speaks of "semi-preconstructed phrases". Literal usages, in contrast, usually allow more variation.

A second defining characteristic of idioms is their often somewhat unexpected and somehow unusual occurrence. This imprecise wording already implies certain challenges for practical applications. In the following, we expect that idioms are often found in uncommon contexts, and we model this assumption by way of recourse to corpus contexts. In simplified terms: An expression like *auf dem Abstellgleis stehen* (engl. *to stand on the railway siding*, idiomatic for *no longer in demand*), featuring a base noun that comes from the railway sector, is quite likely used idiomatically if we do not see any railroad jargon in the surrounding sentences, and rather literally if it neighbours on sentences containing words like *Zug* (engl. *train*), *Bahnhof* (engl. *station*), *S-Bahn* (engl. *suburban railway*), or *Verzögerung im Betriebsablauf* (engl. *delay in operations*).

Of course, both key criteria – formal fixedness and uncommon usage; many others has been developed with different scientific backgrounds, see e.g. (Fernando, 1996) as starting point – change over time, as idiomacity does in general. Somewhere along the way, multi-word expressions may switch from one (predominant) meaning to another, allow more or less variation, and are used in different contexts. Nevertheless, these two criteria form a promising – because reasonably to implement – conceptual basis for the empirical classification approach described below.

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-119560

Our objective is to disambiguate literal and semantically idiomatic occurrences of German verbal idioms (VIDs). To this end, we do not carry out deep analysis of idiom candidates, which would require intensive preprocessing and examinations on the morphosyntactic and semantic level, but attempt a rather shallow approach. Encouraged by previous good results when identifying idiomatic expressions in a specialized corpus, our classification pipeline models idiom characteristics with statistical measures and, based on that, trains machine-learning classifiers suitable to determine the correct reading (*literal* vs. *figurative*) of candidate expressions. We apply count-based collocation measures for the detection of phraseness (i.e. more or less stable occurrences throughout the dataset) and context features for the detection of uncommon usage.

2 Related Work

The fact that idioms are conspicuous in function, form and distribution makes them one of the most tricky parts of language, both for machine-driven processing and regarding cognitive aspects of language comprehension. Natural Language Processing (NLP) tasks such as Information Retrieval, Automatic Text Summarization, or Machine Translation need to disambiguate (potentially) idiomatic expressions (Constant et al., 2017), and since idioms are a central part of everyday language, also language learning must impart suitable skills in order to give learners "the ability to speak a fluent and appropriate version of a language" (Grant and Bauer, 2004). This is supported by the findings of Burchardt et al. (2006), that between 15% and 83% – depending on verb frequency classes – of all verb occurrences in German newspapers are actually used figuratively.

It is therefore not surprising that idiom disambiguation has long been subject of applied research. We provide a brief overview of some state-of-the-art approaches in (Amin et al., 2021), and mention related work here only insofar as it is connected methodologically with our shared task classification pipeline.

The interdisciplinary pan-European PARSEME (PARSing and Multi-word Expressions) network has developed resources and tools for a variety of languages, including German (Savary et al., 2018). Recently, activities has been transferred to the ACL Special Interest Group on the Lexicon (SIGLEX) (Markantonatou et al., 2020). The dataset for the KONVENS 2021 shared task at hand has used PARSEME's criteria for VIDs as annotation guidelines.

Corpus-derived collocation strength between a word token and its neighbours as an indicator for formal fixedness of idiom candidates relies on well-established frequency-based association measures and information-theoretic measures, see e.g. (Evert, 2008) and (Proisl, 2019). Fazly and Stevenson (2006) examine the performance of such measures in an experimental setup using BNC verb-noun pairs.

The idea of measuring local word contexts for token-based idiom classification by means of (differences between) word embeddings matrices has e.g. been proposed by Peng et al. (2018) and already evaluated in Peng and Feldman (2016) or Fazly et al. (2009). Sporleder and Li (2009) point out that idioms behave similarly to spelling errors in the sense that they often do not fit their context; they thus include the collocational contexts of multi-word expressions in their classification model and compile a dedicated corpus of (English) idioms in context (Sporleder et al., 2010). More recently, related measures for the detection of non-literal meaning has been used in Köper and Schulte im Walde (2017) or Kurfalı and Östling (2020); Socolof et al. (2021) use word embeddings to compute a so-called "measure of conventionality" using the BERT language model.

3 Datasets

We now introduce special requirements of the given dataset, and compare it to starting points and goals of our earlier work.

Our original classification approach has been designed for a dataset derived from a highly idiosyncratic text collection – the Corpus of German Pop Song Lyrics (Schneider, 2020) – and displays its strength by detecting multi-word idioms and similar content. This also refers to proverbs, sayings and metaphors – cf. e.g. (Stefanowitsch and Gries, 2007) or (Burger, 2015) – like *wie die Made im Speck* (engl. *like a maggot in bacon*, literally referring to a luxurious lifestyle) or *Ei des Kolumbus* (engl. *egg of Columbus*, semantically completely unmotivated way of describing a both brilliant and easy idea) – within large datasets of randomly extracted word ngrams. So its original intention was not to distinguish between clearly idiomatic and clearly non-idiomatic use of given multi-word ex-

pressions, but to seek for previously unknown idiomatic content, including the already mentioned intermediate and related forms, within an unfiltered corpus.

Due to its special media and conceptual conditions, and its diversity in terms of topics and vocabulary (Schneider et al., 2021), the underlying lyrics' corpus forms a valuable source for both already known and innovative idiomatic constructions. But still, the generated datasets contain substantially less idiomatic than non-idiomatic content, as usual in authentic language. In particular, the original training set comprises a very high number of ngrams without any even potential idiomaticity. As a rough indication: for every 10,000 ngrams (bi-, tri-, tetra-, penta- and hexagrams) there are only some hundred idiomatic expressions in our manually annotated gold standard. Besides close-to-standard language, they also comprise colloquial spoken language (*umme Ecke bringen* instead of *um die Ecke bringen*; idiomatic for engl. *to murder*). Quite often, the unfiltered ngrams contain innovative wordplays and variations of established idiomatic phrases (*dahin, wo der Flavour wächst* instead of *dahin, wo der Pfeffer wächst* or *red mir eine Frikadelle ans Ohr!* instead of *das Ohr abkauen*), both idiomatic and non-idiomatic content (*Achterbahnfahrt der Gefühle keiner der*; engl. *roller coaster of feelings none of the*), or incomplete idioms (*ich werf die Flinte nicht [ins Korn]*; engl. *I do not throw the rifle [into the grain]*). Our approach has been proven to function well (F1-Score of 61.9% for a cutoff of 0.3) as a recognition procedure for idiomatic multiword expressions (MWE).

The shared task dataset, in contrast, operates on pre-selected data. Syntactically, it restricts itself to verbal idioms. It also focuses on figurative usages and leaves out other idiomatic or related phenomena mentioned above. As a merger of the COLF-VID (Ehren et al., 2020) and the German SemEval-2013 task 5b (Korkontzelos et al., 2013) datasets, the shared task collection features literal and semantically idiomatic occurrences of well-known German VID types. In other words: it covers exclusively MWEs that are used literally and idiomatically, or have at least the potential to be used idiomatically in another context. The training dataset is only rudimentarily balanced in terms of figurative and literal occurrences (5705 *figuratively* vs. 1172 *literally*, with additional 6 *both* and 19 *undecidable*).

We nevertheless assume the shared task data design to remove a potential limitation of our approach, namely the identification of idioms where the idiomatic use constitutes the overwhelmingly dominant use – or cases where in language reality we only observe idiomatic use at all. A typical example is the word *Hucke* (originally the burden that you carry on your back) that nowadays is only found in idiomatic expressions like *jemandem die Hucke voll lügen* (engl. *lie to someone badly*), *die Hucke voll kriegen* (engl. *get beaten up*), or *sich die Hucke voll saufen* (engl. *get drunk*). For such idioms, our attempt to identify uncommon usage with context analyzes would probably not produce meaningful results. Our expectation is that – since such cases are not part of the dataset – context features can play a more significant role in our shared task pipeline.

The 1511 given unclassified instances of German VIDs are enriched with altogether three sentences each. Apart from the sentence containing the idiom candidate, the previous and following sentences are available as well. The actual parts of the multiword expression – which may have discontinuous structures – are marked with and tags. For the feature computation, we distinguish between these explicitly tagged words (hereinafter being referred to as 'core ngrams') and everything between the first and the last ('full ngrams'). So, for example, in the sentence *Frahm nahm den Jungen auf den Arm*, the core ngram would be *nahm auf Arm*, and the full ngram would be *nahm den Jungen auf den Arm*. This distinction is reflected in the feature set, exhaustively listed in Table 2.

4 Features

In the following, we explain how we classify the given dataset, and which linguistic and extra-linguistic features we use.

As in our approach in (Amin et al., 2021), we try to distinguish between figurative and literal use by means of shallow features calculated from the multiword expressions and their context. We employ three feature sets (for a detailed breakdown see Table 2): Syntagmatic features (SY) measure collocation strength between all word pairs within a multiword expression. Context features (CO) measure semantic similarity between the words within a multiword expression and the words in its left/right context. Finally, other features (O) represent a variety of counts to assess the amount of evidence avail-

able, such as number of words in a MWE.

The basic idea behind syntagmatic features (SY) is to assess fixedness of idiomatic constructions by means of collocation strength. We employ a variety of count based collocation measures (SY_C, (Evert, 2008)), and predictive collocation measures (SY_W). The predictive collocation measures 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), with a window size of +/- 5[1]. However, due to the different dataset the syntagmatic features need to be adapted: First, because there is no underlying corpus of the complete texts for the shared task available, the collocation features SY_C2 based on the counts of our lyrics corpus cannot be employed, leaving us with collocation features based on the counts of DeReKo, as a background corpus. Second, because the marked MWEs can be both, figurative and literal, we take into account one word to the left and to the right of the multiword expression as additional context.

The basic idea of the context features (CO) is to assess uncommon usage of idiomatic contstruction. We extend the context features (CO) used in (Amin et al., 2021) by some new cosine similarity measures described in Table 2. As described above, we perform the context calculations twice for each instance: The first time the local context of the full idiom candidate is considered, the second time the local context of the core idiom candidate only. In order to discriminate between lexical words and function words, as it has to be done for the CO_VEC_LEX features, the context sentences are preprocessed. Basic word segmentation is conducted with the KorAP-Tokenizer (Kupietz, 2021), word classes are annotated with TreeTagger (Schmid, 1995). CO_VEC_LEX features take only nouns, verbs, adverbs, and adjectives within an idiom candidate and its context into account. Incidentally, the CO_VEC values each relate to the context of a single instance. For those cases in which an idiom occurs several times in the dataset, one could still think about calculating averages of all occurences; after all, the 1511 dataset instances can be traced back to only 64 unique VID types.

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

5 Results

To evaluate our feature set we have trained a Random Forest classifier on the training set of the shared task. Figure 1 shows Recall, Precision, F1-score for the class *literal*, and Balanced Accuracy on the development set. The best F1-Score (0.5) is achieved at a cutoff of 0.7, the best balanced accuracy (0.71) at a cutoff of 0.8. Our submission on the test set used cutoff 0.8, achieving a similar F1-Score of about 0.45.

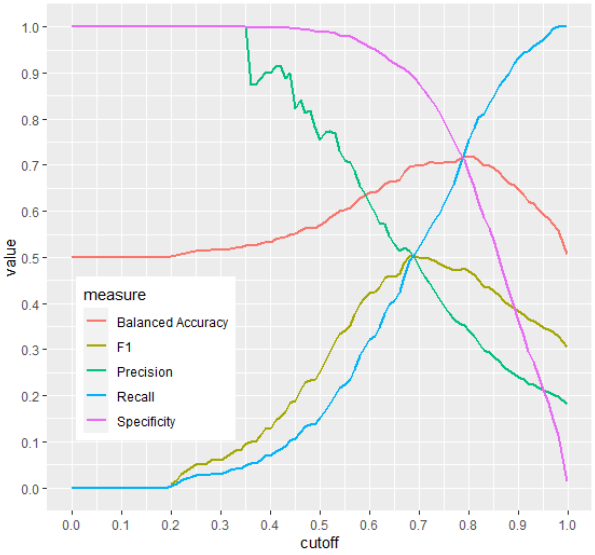

Figure 1: Trade-off curves for Random Forest cut-off

The development set contains three types with 270 instances which are not available in the training set. As shown in Figure 2, the best performance on these unseen types is achieved at a cutoff of 0.65, with an F1-Score of 0.5, and a balanced accuracy of 0.7. With the submission cutoff of 0.8 though, the F1-Score is only 0.32 (0.3 on the testset).

For a more comprehensive assessment of the generalizability of our approach to unseen types, we have also performed a leave-one-type-out crossvalidation over all 63 types in the combined training and development set, using for each type all corresponding instances as test set, and all other instances as training set. This achieves an overall F1-Score of 0.39, with Recall 0.66 and Precision 0.28, at cutoff 0.8. Table 1 lists the top and bottom three types by F1-Score, together with the number of underlying instances (Insts.) and the percentage of literal instances (Literal%). Clearly the best F1-Scores are accomplished for VID types that are mainly used literally. Indeed, there exists a very strong positive correlation of 0.9 (Spearman and Pearson) between

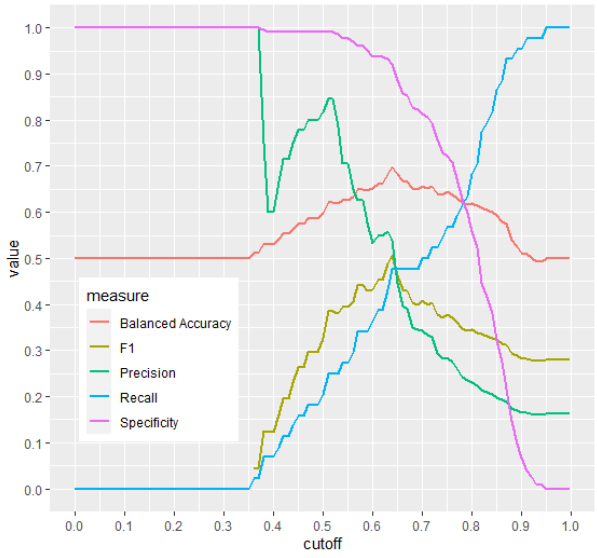

Figure 2: Trade-off curves for unseen types

the percentage of literal instances and F1-Score. Thus, apparently our approach typically misclassifies VIDs with dominant idiomatic use, resulting in low Precision and thereby low F1-Score.

Type	Insts.	Literal%	F1-Score
in Blut haben	20	80.0	88.9
Luft holen	79	73.4	83.1
über Bord gehen	26	61.5	75.0
auf Zug aufspringen	165	1.8	6.1
über Bühne gehen	174	1.1	3.5
auf Strecke bleiben	534	0.7	2.6

Table 1: Top and bottom 3 F1-Scores by type

Table 3 analyses the contribution of the individual feature sets to Precision, Recall, F1-Score, and Balanced Accuracy.[2] Clearly, the context features CO contribute most, i.e., they achieve the best performance individually, and leaving them out (w/o CO), results in the worst performance.

The count based collocation measures (SY_C), however, are not very far behind. As described above these are calculated on the ngram extended by one word to the left and one word to the right, otherwise count based collocation measures could hardly differentiate between figurative and literal use. All other feature sets (SY_W, SY_R, O) contribute hardly, in fact Balanced Accuracy for O is basically at the level of a random baseline (50%). The small contribution of the predictive collocation measures SY_W also indicates that the figurative expressions in this dataset often constitute dominant usage.

Table 2 gives a detailed breakdown of the contribution of the individual features to the classification task. *MDA* gives the random forest's estimate of the mean decrease in accuracy per feature, *IGain* the information gain[3], *TTest* the degree of significance by a Welch two sample t-test for confidence levels 0.95 (*), 0.99 (**), and 0.999 (***), and Δ the sign of the difference between the mean of a feature for figurative ngrams vs. literal ngrams.

Consistent with the analysis in Table 3 the context features CO contribute most, both w.r.t. MDA and IGain. Of those, CO_VEC_LEX_2 has by far the largest contribution, followed by its maximum variant CO_VEC_LEX_MAX_2. The average and maximum variants of CO_VEC also are significantly different according to the TTest between figurative and literal ngrams, whereas the minimum variants are in general not significantly different, and contribute rather little to the classification. With the exception of the (insignificant) CO_VEC_MIN_1 feature, the difference in means between figurative and literal ngram features is negative, i.e., figurative ngrams tend to be less similar to words in their broader context than literal ngrams.

The count-based collocation measures SY_C also have a fairly high MDA, and notably SY_C_LDAF has a fairly high information gain. Almost all SY_C features differ significantly for figurative vs. literal ngrams, and the difference is positive, i.e., figurative ngrams tend to have higher collocation strength than literal ngrams.

The predictive collocation measures SY_W on the other hand have very low information gain, and do not differ significantly between figurative and literal usage. This is in contrast to the the song lyrics dataset, where a corpus of complete texts can be evaluated statistically. On top of this comes the fact that lyrics contain more innovative idioms, which do not constitute the dominant use of words, so that predictive collocation measures exhibit the largest contribution for idiom detection.

[2]These numbers have been obtained by 5 times repeated 5-fold crossvalidation on the combined training and development set.

[3]Information Gain estimates how much information a feature contributes to the classification, or more technically, how much the entropy of the class distribution (figurative vs. literal) is reduced by splitting the instances on a particular feature. IGain here is measured in nats and scaled by *1000 for better readability.

Feature	MDA	IGain	TTest	Δ	Description
CO_VEC_1	29.7	2.9	*	-	avg. cosine similarity between words in full ngram and words in +/-5 context
CO_VEC_2	27.4	4.4	***	-	avg. cosine similarity between words in core ngram and words in +/-5 context
CO_VEC_MAX_1	27.0	5.4	***	-	max. cosine similarity between words in full ngram and words in +/-5 context
CO_VEC_MAX_2	23.6	12.5	***	-	max. cosine similarity between words in core ngram and words in +/-5 context
CO_VEC_MIN_1	20.2	0.0		+	min. cosine similarity between words in full ngram and words in +/-5 context
CO_VEC_MIN_2	19.6	0.0		-	min. cosine similarity between words in core ngram and words in +/-5 context
CO_VEC_LEX_1	28.5	23.2	***	-	like CO_VEC_1 but only on lexical words
CO_VEC_LEX_2	42.8	44.0	***	-	like CO_VEC_2 but only on lexical words
CO_VEC_LEX_MAX_1	28.7	22.6	***	-	like CO_VEC_MAX_1 but only on lexical words
CO_VEC_LEX_MAX_2	39.3	41.3	***	-	like CO_VEC_MAX_2 but only on lexical words
CO_VEC_LEX_MIN_1	22.1	0.0		-	like CO_VEC_MIN_1 but only on lexical words
CO_VEC_LEX_MIN_2	21.9	2.4	***	-	like CO_VEC_MIN_2 but only on lexical words
O_DEREKO_1	17.2	0.0		-	number of words of full ngram in DeReKo
O_DEREKO_2	19.3	12.9	***	+	number of words of core ngram in DeReKo
O_GRAM_1	19.6	0.0		-	number of words in full ngram
O_GRAM_2	19.0	12.9	***	+	number of words in core ngram
O_NSTOPW	16.6	0.0		-	number of non stop words in full ngram
SY_C_C_L	26.0	3.4	***	+	raw count of left neighbour
SY_C_C_R	29.8	6.3	***	+	raw count of right neighbour
SY_C_DICE	22.4	13.1	**	+	dice
SY_C_LD	22.7	12.4	***	+	logdice
SY_C_LDAF	23.3	24.6	***	+	logdice with autofocus
SY_C_LL	26.0	5.8	***	+	loglikelihood
SY_C_MI	25.9	8.2	***	+	(pointwise) mutual information, MI
SY_C_MI_L	27.1	10.7	***	+	MI with left neighbour
SY_C_MI_R	29.6	0.0	*	+	MI with right neighbour
SY_C_MI2	24.9	12.0	***	+	MI^2
SY_C_MI3	26.3	12.2	***	+	MI^3
SY_C_NMI	26.9	9.6	***	+	normalized MI
SY_C_R	26.0	5.1	***	-	rank by SY_C_LD
SY_W_AVG	21.0	0.0		-	average of output activations with autofocus
SY_W_CON	22.1	0.0		+	conorm of column normalized output activations with autofocus
SY_W_MAX	22.4	0.0		-	max of output activations
SY_W_NSUM	16.7	0.0		+	sum of output activations normalized by total sum over all columns
SY_W_NSUM_AF	21.4	0.0	*	-	sum of output activations divided by sum over all selected columns with autofocus
SY_W_R1	20.3	2.5		+	rank by SY_W_CON
SY_W_R2	21.5	0.0		-	rank by SY_W_NSUM2
SY_R_D	20.4	6.7	**	+	rank difference: SY_W_R1-SY_C_R

Table 2: Features used for the classification pipeline

Feature set	Precision	Recall	F1-Score	Bal. Acc.
all	33.9	73.9	46.4	72.0
CO	28.7	58.6	38.5	64.2
SY_C	33.6	43.6	38.0	62.9
SY_W	30.4	35.2	32.6	59.2
SY_R	29.4	35.5	32.1	58.9
O	29.8	2.1	4.0	50.5
w/o CO	34.3	56.7	42.7	67.1
w/o SY_C	31.1	65.3	42.2	67.7
w/o SY_W	33.7	72.7	46.1	71.5
w/o SY_R	34.0	73.7	46.5	72.0
w/o O	33.0	72.6	45.4	71.0

Table 3: Performance of different feature sets in a Random Forest with cutoff=0.8. CO: Context Features, SY_C: Count-based collocation measures. SY_W: Predictive collocation measures. SY_R: Rank-based collocation measures. O: Other.

6 Conclusions

In this paper we have applied shallow context analysis based on corpus based features for distinguishing between literal and figurative use of verb-noun constructions. The achieved classification performance of 0.45 F1-Score and 0.7 Balanced Accuracy is worse than the previously achieved performance on a dataset consisting of idioms sampled from the song lyrics corpus with 0.6 F-Score, and 0.79 Balanced Accuracy.

This difference can be explained: While two of the introduced feature sets – the context features CO, which detect semantically dissimilar words in the context of an MWE, and the count-based collocation measures SY_C, which assess fixedness of an MWE – do contribute to classification on the shared task dataset, the predictive collocation measures SY_W do not, unlike on the song lyrics dataset. This indicates that the figurative just like the literal MWEs in the current idiom collection often constitute dominant usage, which leads to relatively high predictive collocation strength between the words constituting a MWE, irrespective of them being figurative or literal. Indeed, as the analysis on unseen types shows, our approach typically misclassifies VIDs with highly dominant figurative use as literal.

So, besides the initial classification objective of this study, another lesson learned here confirms the known issue that the quality of corpus-based empirical evidence should always be judged by the composition and stratification of the actually included language data. We will keep on examining the performance of our feature sets for various corpus types.

References

Miriam Amin, Peter Fankhauser, Marc Kupietz, and Roman Schneider. 2021. Data-driven identification of idioms in song lyrics. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021), Special Interest Group on the Lexicon (SIGLEX) of the Association for Computational Linguistics (ACL)*.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Harald Burger. 2015. *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Schmidt, Berlin.

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.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 41–48, Prague, Czech Republic. Association for Computational Linguistics.

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*, pages 211–220, 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, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

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.

Chitra Fernando. 1996. *Idioms and Idiomaticity*. University Press, Oxford.

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

Lynn Grant and Laurie Bauer. 2004. Criteria for redefining idioms: Are we barking up the wrong tree? *Applied Linguistics*, 25:38–61.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. SemEval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, Atlanta, Georgia, USA. Association for Computational Linguistics.

Marc Kupietz. 2021. KorAP/KorAP-Tokenizer v2.1.0.

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.

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.

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.

Jing Peng and Anna Feldman. 2016. Experiments in idiom recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2752–2761, Osaka, Japan. The COLING 2016 Organizing Committee.

Thomas Proisl. 2019. *The cooccurrence of linguistic structures*. doctoralthesis, FAU University Press.

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.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

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.

Roman Schneider, Sandra Hansen, and Christian Lang. 2021. Das Vokabular von Songtexten im gesellschaftlichen Kontext – ein diachron-empirischer Beitrag. In *Sprache in Politik und Gesellschaft: Perspektiven und Zugänge. Jahrbuch 2021 des Instituts für Deutsche Sprache*. De Gruyter.

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

Michaela Socolof, Jackie Chi Kit Cheung, Michael Wagner, and Timothy J. O'Donnell. 2021. Characterizing idioms: Conventionality and contingency.

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.

Caroline Sporleder, Linlin Li, Philip Gorinski, and Xaver Koch. 2010. Idioms in context: The IDIX corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*

(LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

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

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