Supervised Disambiguation of German Verbal Idioms with a BiLSTM Architecture

Rafael Ehren[1], **Timm Lichte**[2], **Laura Kallmeyer**[1], **Jakub Waszczuk**[1]
[1]Heinrich Heine University, Düsseldorf, Germany
[2]University of Tübingen, Tübingen, Germany
{ehren|kallmeyer|waszczuk}@phil.hhu.de
timm.lichte@uni-tuebingen.de

Abstract

Supervised disambiguation of verbal idioms (VID) poses special demands on the quality and quantity of the annotated data used for learning and evaluation. In this paper, we present a new VID corpus for German and perform a series of VID disambiguation experiments on it. Our best classifier, based on a neural architecture, yields an error reduction across VIDs of 57% in terms of accuracy compared to a simple majority baseline.

1 Introduction

Figurative language is not just a momentary product of creativity and associative processes, but a vast number of metaphors, metonyms, etc. have become conventionalized and are part of every speaker's lexicon. Still, in most cases, they can simultaneously be understood in a non-figurative, literal way, however implausible this reading might be. Take, for example, the following sentence:

(1) He is in the bathroom and talks to Huey on the big white telephone.

The verbal phrase *talk to Huey on the big white telephone* can be understood as a figurative euphemism for being physically sick. But it could also be taken literally to describe an act of remote communication with a person called Huey. Despite the ambiguity, a speaker of English will most probably choose the figurative reading in (1), also because of the presence of certain syntactic cues such as the adjective sequence *big white* or the use of *telephone* instead of, for example, *mobile*. Omitting such cues generally makes the reader more hesitant at selecting the figurative meaning. There is thus a strong connection of non-literal meaning and properties pertaining to the form of the expression, which is characterstic for what Baldwin and Kim (2010) call an idiom. Since the figurative expression in (1) consists of a verb and its syntactic arguments, we will furthermore call it a Verbal Idiom (VID) adapting the terminology in Ramisch et al. (2018).

While it is safe to assume that the VID *talk to Huey on the big white telephone* almost never occurs with a literal reading, this does not hold for all idioms. The expression *break the ice* for example can easily convey both a literal (*The trawler broke the ice*) and a non-literal meaning (*The welcome speech broke the ice*) depending on the subject. Although recent work suggests that literal occurrences of VIDs generally are quite rare in comparison to the idiomatic ones (Savary et al., 2019), it remains a qualitatively major problem with the risk of serious errors due to wrong disambiguation.

However, tackling this problem with supervised learning poses special demands on the learning and test data in order to be successful. Most importantly, since the semantic and morphosyntactic properties of VID types (and idioms in general) are very diverse and idiosyncratic, the data must contain a sufficient number of tokens of both the literal and non-literal readings for each VID. In addition, each token should allow access to the context because the context can provide important hints as to the intended reading.

In this paper, we investigate the supervised disambiguation of potential occurrences of German VIDs. For training and evaluation, we have created COLF-VID (Corpus of Literal and Figurative Readings of Verbal Idioms), a German annotated corpus of literal and semantically idiomatic occurrences of 34 preselected VID types. Altogether, we have collected 6985 sentences with candidate occurrences that have been semantically annotated by three annotators with high inter-annotator agreement. The annotations overall show a relatively low idiomaticity rate of 77.55 %, while the idiomaticity

Proceedings of the Second Workshop on Figurative Language Processing, pages 211–220

https://doi.org/10.18653/v1/P17

rates of the single VIDs vary greatly. The derived corpus is made available under the Creative Commons Attribution-ShareAlike 4.0 International license.[1] To the best of our knowledge, it represents the largest available collection of German VIDs annotated on token-level.

Furthermore, we report on disambiguation experiments using COLF-VID in order to establish a first baseline on this corpus. These experiments use a neural architecture with different pretrained word representations as inputs. Compared to a simple majority baseline, the best classifier yields an error reduction across VIDs of 57% in terms of accuracy.

2 Related Work

2.1 VID Resources

In this section, we discuss previous work on the creation of token-level corpora of VID types.

Cook et al. (2007) draw on syntactic properties of multiword expressions to perform token-level classification of certain VID types. To this end they created a dataset of 2984 instances drawn from the BNC (British National Corpus), covering 53 different verb-noun idiomatic combination (VNIC) types (Cook et al., 2008). The annotation tag set includes the labels LITERAL, IDIOMATIC and UNKNOWN which correspond to three of the four labels used for COLF-VID, albeit the conditions for the application of UNKNOWN where a bit different, since the annotators only had access to one sentence per instance. The overall reported unweighted Kappa score, calculated on the dev and test set, is 0.76. Split decisions were discussed among the two judges to receive a final annotation.

The VU Amsterdam Metaphor Corpus (Steen et al., 2010) is currently probably the largest manually annotated corpus of non-literal language and is freely available. It comprises roughly 200,000 English sentences from different genres and provides annotations basically for all non-functional words following a refined version of the Metaphor Identification Procedure (MIP) (Pragglejaz Group, 2007). Regarding only verbs, this yields an impressive overall number of 37962 tokens with 18.7% "metaphor-related" readings (Steen et al., 2010; Herrmann, 2013). Due to its general purpose and the lack of lexical filtering, however, this is hardly comparable with COLF-VID.

The IDIX (IDioms In Context) corpus created by Sporleder et al. (2010) can be seen as the English counterpart of COLF-VID. It is an add-on to the BNC XML Edition and contains 5836 annotated instances of 78 pre-selected VIDs mainly of the form V+NP and V+PP. As for our corpus, expressions were favoured that presumably had a high literality rate. The employed tag set was more or less identical with ours. Quite remarkably, and in stark contrast to COLF-VID and other comparable corpora, the literal occurrences in the IDIX corpus represent the majority class with 49.4% (vs. 45.4% instances being tagged as NON-LITERAL). They report a Kappa score of 0.87 which was evaluated using 1,136 instances that were annotated independently by two annotators.

Fritzinger et al. (2010) conduct a survey on a German dataset similar to ours. They extracted 9700 instances of 77 potentially idiomatic preposition-noun-verb triples from two different corpora. Two annotators independently classified the candidates according to whether they were used literally or idiomatically in a given context. The tag set also included an AMBIGUOUS label, but, as was the case with Cook et al. (2008), only single sentences were available as context to determine the correct reading. An agreement rate of 97.9% was computed on the basis of 6,690 instances. The biggest difference to our and other presented corpora is the very high idiomaticity rate of 96.12%. However, this dataset does not seem to be publicly available.

Horbach et al. (2016) are concerned with German infitive-verb compounds such as *sitzen lassen* ('let sit'⇒'leave someone'), i.e. verb groups with an idiomatic reading that consist of an inflected head verb and an infinitive modifier. In order to conduct experiments on automatic detection and disambigution of these kinds of VIDs they created a corpus of 6000 instances of 6 different infinitive-verb compounds which were annotated by two experts with the label set LITERAL, IDIOMATIC and ? (for undecidable). In contrast to Cook et al. (2008) and Fritzinger et al. (2010), a context of one sentence to the left and one sentence to the right of the candidate was taken into account. The annotation process proved to be especially challenging since some of the examined compounds had several literal and figurative meanings. Nevertheless, they achieved high agreement values of ($0.6 < \kappa < 0.8$) or ($\kappa > 0.8$) for most expressions with a mean idiomaticity rate of 65.5%.[2]

[1] https://github.com/rafehr/COLF-VID

[2] Kappa scores and idiomaticity rates were reported independently for each expression.

2.2 VID Disambiguation

Even though literal occurrences of VIDs seem to be a rare phenomenon (Savary et al., 2019), it is still desirable to account for them, i.e. to disambiguate between idiomatic and literal reading. It may be a quantitatively minor problem, but qualitatively it continues to be a major challenge for NLP, for instance for machine translation systems.

VIDs exhibit a variety of properties exploitable for determining the correct reading of a candidate expression. On the morphosyntactic level a lot of VIDs are less flexible than their literal counterparts, e.g. the idiomatic *kick the bucket* is not readily passivizable. On the semantic level VIDs often disrupt the cohesion of a sentence, because of their non-compositionality, or they violate selectional preferences, for example in the sentence *The city shows its teeth*.

Examples for a morphosyntactic approach are the works of Cook et al. (2007) and Fazly et al. (2009). They show that it is possible to leverage automatically acquired knowledge about the syntactic behaviour of VNICs, i.e. their syntactic fixedness, to perform token-level disambiguation.

Katz and Giesbrecht (2006) draw on semantic properties by using dense word vectors to identify literal and idiomatic occurrences of the German VID *ins Wasser fallen* (idiomatically 'to be cancelled', literally 'to fall into the water'). They assumed that the contexts of the literal and idiomatic use of this expression differ which in turn is represented by their distributional vectors. Test instances are then compared to these vectors in order to classify them.

Li and Sporleder (2009) and Ehren (2017) both used cohesion-based graphs for the disambiguation task, the assumption being that semantically idiomatic expressions disrupt the cohesion of the context they appear in. The former used Normalized Google Distance, while the latter used the cosine between word embeddings to capture the semantic similarity of words. To classify the test instances in an unsupervised way, graphs were built based on the two mentioned metrics and if the mean value rose after the removal of the instance, it was classified as idiomatic.

Shutova et al. (2010) and Haagsma and Bjerva (2016) employ the knowledge that metaphors tend to violate selectional preferences to detect them in running text.

Building on these insights from previous work, in this paper, we will use a BiLSTM architecture based on different types of word embeddings that is intended to capture the semantic properties of the VID itself, together with the context and the morphosyntactic flexibility of the specific VID instance.

3 The Creation of the Corpus

3.1 The Data

As mentioned above, literal occurrences of VIDs usually seem to occur quite rarely. The German dataset of the PARSEME 1.1 corpus (Ramisch et al., 2018) consists of 8996 sentences with 1341 instances of VIDs. These 1341 instances have an idiomaticity rate of 98%, i.e. the whole dataset only includes a handful of literal occurrences. Training and evaluating a classifier with such an imbalance of classes would prove rather difficult. Thus, it is not feasible to gather a sufficient amount of data by selecting sentences at random – at least if human resources are limited – and it is not possible to build a huge dataset so that the natural occurrence rate will give us enough literal readings. In order to alleviate the data sparsity, we hand-picked a number of VID types with presumably high numbers of literal occurrences. Afterwards we extracted sentences (along with their contexts) from the German newspaper corpus TüPP-D/Z[3] that contained the lexical components of our VID types as lemmas. We then manually filtered out coincidental occurrences with an undesired coarse syntactic structure (Savary et al., 2019), leaving us with only valid candidates for our corpus. Table 1 shows the 34 different types. One thing that immediately stands out is the fact that most of the pre-chosen VID types (26 to be exact) consist of a prepositional phrase (PP) and a verb. The rest consists of verb-noun combinations with the noun in direct object position. Another salient property of this dataset is the high variance with respect to the number of candidates per type. For the VID *an Glanz verlieren* ('loose sheen'⇒'loose attractivity'), we only found 5 instances, while *auf dem Tisch liegen* ('lay on the table'⇒'be topic') is represented by 951 candidates.

3.2 The Annotation Labels

Besides the labels LITERAL, IDIOMATIC we also use the labels UNDECIDABLE and BOTH in cases

[3] http://hdl.handle.net/11858/00-1778-0000-0007-5E99-D

VID type	Lit.	Idiom.	Und.	Both	I%
am Boden liegen	35	11	0	1	23.4
an Glanz verlieren	0	15	1	0	93.75
an Land ziehen	25	235	0	0	90.38
am Pranger stehen	0	5	0	0	100.0
den Atem anhalten	10	30	0	0	75.0
auf dem Abstellgleis stehen	15	11	0	0	42.31
auf den Arm nehmen	39	50	0	0	42.31
auf der Ersatzbank sitzen	16	5	0	0	23.81
auf der Straße stehen	93	156	1	0	62.4
auf der Strecke bleiben	4	616	1	0	99.19
auf dem Tisch liegen	262	678	10	1	71.29
auf den Zug aufspringen	5	121	0	0	96.03
eine Brücke bauen	109	238	1	0	68.39
die Fäden ziehen	9	164	0	0	94.8
im Blut haben	29	7	0	0	19.44
in den Keller gehen	34	91	0	0	72.8
in der Luft hängen	28	256	0	0	90.14
im Regen stehen	69	302	4	4	79.68
ins Rennen gehen	11	51	0	0	82.26
in eine Sackgasse geraten	2	99	0	0	98.02
im Schatten stehen	7	52	0	1	86.67
in Schieflage geraten	3	40	1	0	90.91
ins Wasser fallen	67	186	0	0	73.52
Luft holen	100	66	4	0	38.82
mit dem Feuer spielen	9	74	2	0	87.06
einen Nerv treffen	1	284	0	0	99.65
die Notbremse ziehen	51	275	0	0	84.36
eine Rechnung begleichen	89	162	0	0	64.54
von Bord gehen	45	48	0	0	51.61
vor der Tür stehen	189	409	1	1	68.17
ein Zelt aufschlagen	53	41	6	0	41.0
über Bord gehen	62	52	1	0	45.22
über Bord werfen	54	389	0	0	87.81
über die Bühne gehen	2	198	0	0	99.0
Total	1527	5417	33	8	77.55

Table 1: Statistics of COLF-VID

where an expression can be seen as LITERAL and IDIOMATIC at the same time for different reasons.

As to UNDECIDABLE, the disambiguation of an expressions is not possible due to the lack of context. For instance, this is notoriously difficult for metonymic expressions whose literal meaning describes a bodily action that typically co-occurs with the idiomatic meaning. An example of that is the German expression *sich die Haare raufen* ('to scuffle one's hair'⇒'to be worried/upset'): A person that is upset can often be seen scuffling their hair.[4]

By contrast, the label BOTH applies to cases where the literal and idiomatic readings seem to be both intended, as illustrated in (2):

(2) Wer möchte, könnte ihm den Kopf waschen,
Who wants, could him the head wash,
ihm mal auf den Zahn fühlen oder ihn gar
him once on the tooth feel or him even
auf den Arm nehmen [...].
on the arm take [...].

This sentence originates from an article depicting proposals on how to proceed with the statue of a certain historic personality and it contains the VIDs *jmdm. den Kopf waschen* ('wash someone's head'⇒'scold someone'), *jmdm. auf den Zahn fühlen* ('feel someone's tooth'⇒'interrogate someone') and *jmdn. auf den Arm nehmen* ('take someone on your arm'⇒'taunt someone'[5]). The author of the sentence suggests to tear the statue down and to perform the aforementioned actions in an effort to demystify the person represented by the statue. The wordplay used here relies on the fact that all the VIDs relate to bodily actions and could be performed on a statue. Thus, both readings, literal and idiomatic, are active at the same time.

3.3 The Annotation Guidelines

The annotation guidelines basically consisted of definitions of the applicable labels, coupled with examples. A condensed version of the definitions is given below:

- **LITERAL**: In the context of this annotation task we equate literality with compositionality. We understand compositionality as the property that the semantics of an expression is determined by the most basic meanings of its components without any form of figuration involved.

- **IDIOMATIC**: According to Baldwin and Kim (2010)[6] there are different forms of idiomaticity: lexical, syntactic, semantic, pragmatic and statistical. In the context of this annotation task, "idiomatic" is used synonymously with "semantically idiomatic", i.e. the property of an expression that it is not possible to fully derive its meaning by only considering the semantics of its components. Thus we understand semantic idiomaticity as a lack of compositionality.

- **UNDECIDABLE**: This label is for cases in which it is not possible to decide whether the target expression is literal or idiomatic.

- **BOTH**: While the label UNDECIDABLE means that there is only one possible reading, but it's not feasible to decide which, the label BOTH denotes the phenomenon of the two readings being activated at the same time.

[4] *Pull out one's hair* would be the English equivalent, but very seldomly, if not for huge emotional distress, people actually pull out their hair when upset.

[5] The literal meaning of *jmdn. auf den Arm nehmen* would be 'pick someone up'. A translation to English that keeps reference to a corresponding bodily action would be *to pull someone's leg*.

[6] The annotators were required to read Baldwin and Kim (2010) prior to the annotation.

The annotation task then consisted of applying one of the labels to each candidate.

4 Annotation Results

The annotation was performed by three trained linguists on the whole dataset. The annotation results are summarized in Table 1. Columns 2 to 5 contain the counts of the majority decisions for the different labels, while column 6 contains the idiomaticity rate of a VID type. Figure 1 shows an example for an instance of the VID type *die Notbremse ziehen* ('pull the emergency breaks'⇒'quickly terminate a process')[7] in the column format of the corpus. The

```
# global.columns = ID FORM LEMMA
POS ANNO_1 ANNO_2 ANNO_3 MAJORITY_ANNO
# article_id = T890825.128
# text = Bundesbahn will die
Notbremse ziehen
# context_judgement_1 = 0
# context_judgement_2 = 0
1 Bundesbahn Bundesbahn NN * * * *
2 will wollen VMFIN * * * *
3 die die ART * * * *
4 Notbremse Notbremse NN 2 2 2 2
5 ziehen ziehen VVINF 2 2 2 2
```

Figure 1: A sample idiomatic instance in COLF-VID

last four columns contain the annotations: columns 5 to 7 are the annotations of the three different annotators, the last column contains the majority annotation. Since all the annotators agreed that the reading of this instance is idiomatic (2 stands for the tag IDIOMATIC), this is an example for a clear-cut decision. In the rare cases where there was a split decision and every annotator chose a different label, the label UNDECIDABLE was employed.

What immediately stands out is that the overall idiomaticity rate is not nearly as high as the 98% reported for the German PARSEME dataset mentioned in Section 3.1 It ranges from 19.44% (*im Blut haben* 'be in one's blood') to 99.65%[8] (*den Nerv treffen*) and is 77.55% in total. But one has to keep in mind that these two datasets are hardly comparable regarding their statistics, since COLF-VID was created with the intention to maximize the number of literal occurrences by only choosing VID types with a presumably high literality count. Even though there are some VID types with an unexpectedly high idiomaticity rate (*auf der Strecke bleiben*, *in eine Sackgasse geraten* or *über die Bühne gehen* to name a few), the large majority of the chosen VID types is indeed represented with a relatively low idiomaticity rate.

Only 0.59 of the instances received the labels UNDECIDABLE or BOTH (see Figure 2), but this is hardly surprising. We nevertheless wanted to include these tags for the sake of completeness and linguistic interest.

For the three annotators we calculated the following Cohen's Kappa scores on the basis of the whole dataset:

- annotator 1 – annotator 2: 0.9
- annotator 2 – annotator 3: 0.8
- annotator 1 – annotator 3: 0.77

Thus, the agreement is high for all three annotators, which is expected given the nature of the task and the equally high agreement scores reported for comparable corpora (Cook et al. (2008), Sporleder et al. (2010), Fritzinger et al. (2010)).

Another feature of COLF-VID is the context judgement provided by two of the annotators. These judgements can be seen in Figure 1 in the last two lines (starting with a hash tag) before the beginning of the sentence. They indicate whether the annotators needed more than one sentence to determine the reading of an instance. The two zeros denote that this was not the case for this candidate expression ("1" would indicate the opposite). Even if the sentence is rather short with only five words, the fact that the pulling of an emergency break requires an animate agent if used literally was enough information for both annotators to make their decisions. The context judgement feature provides the possibility of excluding candidates where none of the annotators was able to determine the reading only from a single sentence. As a result, instances where one sentence is not sufficient to make an informed decision would be prevented from entering a given system (e.g. a classifier which aims to disambiguate the candidates).

5 VID Disambiguation Experiments

5.1 Setup

The Task The goal of the presented experiments is to train a classifier capable of distinguishing the different readings of a candidate expression. It is important to emphasize that this task is different

[7] Translation: "The federal railway wants to pull the emergency breaks". The combination of *federal railway* and *pull emergency breaks* is very frequent in COLF-VID for obvious reasons.

[8] *Am Pranger stehen* 'stand in the pillory' has an idiomaticity rate of 100%, but its 5 candidates might not be that representative.

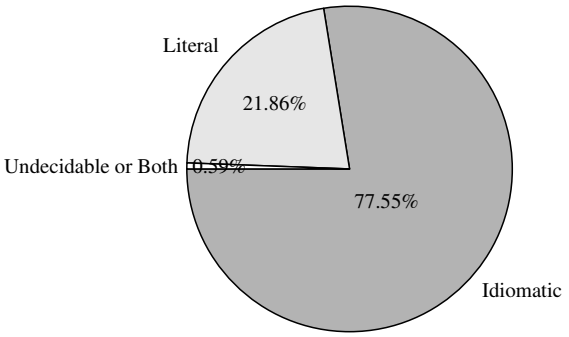

Figure 2: Distribution of annotation labels in COLF-VID

from *identification*, where all the VID occurrences are to be identified in a sentence, e.g. by applying a sequential model to label every token as a VID component or not. The reason for this is that COLF – for now – is a lexical sample corpus, which means it consists of a pre-selected set of target expressions annotated with respect to their contexts. In other words, the sentences could contain non-annotated instances of VID types that weren't part of the pre-selected set, which in turn could confuse the system during training and skew the evaluation results (we will further address this issue in section 6.)

Thus, we modeled the task assuming another process had pre-identified the candidate expressions, which is the usual approach when it comes to the disambiguation of VIDs (Constant et al., 2017). The classifier then only has to decide which label to apply given a certain instance and its context. This means that, although all components of a VID instance received a label during annotation[9] (cf. Figure 1), during classification we conflated all labels of a VID instance into one label for the whole expression. This is possible, since we did not allow for components of an instance to have different labels. For example, the verb cannot be literal while the noun is idiomatic.

Word Representations During the experiments we employed word representations that were pre-trained on other, considerably larger corpora with three different models: Word2vec (Skip-gram) (Mikolov et al., 2013), fastText (CBOW) (Bojanowski et al., 2016) and ELMo (Embeddings from Language Models) (Peters et al., 2018). We trained the Word2vec embeddings ourselves[10] on a variant of the German web corpus DECOW16 (Schäfer and Bildhauer, 2012) which consists of 11 billion tokens and shuffled sentences. The resulting vectors have 100 dimensions. As for the other models we reverted to already existing resources. The fastText embeddings were trained on Common Crawl and Wikipedia with a dimensionality of 300[11]. The German ELMo model was trained on a special Wikipedia corpus that also included the comments besides the articles (May, 2019)[12]. The underlying bidirectional language model provided us with 3 different word representations of size 1024 for each input token. These were averaged to give us one embedding per token.

Architecture There are different properties on the morphosyntactic and semantic level we can leverage during the disambiguation process. E.g. some VIDs do not possess the same lexical or morphological flexibility as their literal counterparts. The VID *kick the bucket*, for instance, does not allow for *bucket* to be replaced by a synonym like *pail* or for it to be in plural form, hence both would be strong indicators for literality. On the semantic level the surrounding context can of course give clues about the correct readings. An observation made during annotation was that, over and over again, the violation of selectional preferences gave a strong indication on how to annotate a candidate. For example in a sentence like *Berlin holds its breath*, *Berlin* is no animate subject which immediately gives away the non-literal nature of the sentence. This is why we settled for a classifier architecture that is best suited for taking the context into account. Figure 3 shows a graph of our architecture.

For an input sentence s of length n with words $w_1, ..., w_n$ we associate every word w_i with its corresponding pretrained word embedding which gives us our input sequence of vectors $x_{1:n}$:

$$x_i = e(w_i)$$

In the case we use Word2vec embeddings, a sequence $w_{1:n}$ consists of lemmas, while for fastText it consists of tokens, because the former model was trained on lemmas and the latter on n-grams.

After the embedding assignment the sequence $x_{1:n}$ is fed into a bidirectional recurrent neural net

[9] In order to allow for a different kind of task at a later point.

[10] We used the word2vec implementation of the python package gensim (Řehůřek and Sojka, 2010).

[11] https://fasttext.cc/docs/en/crawl-vectors.html

[12] https://github.com/t-systems-on-site-services-gmbh/german-elmo-model

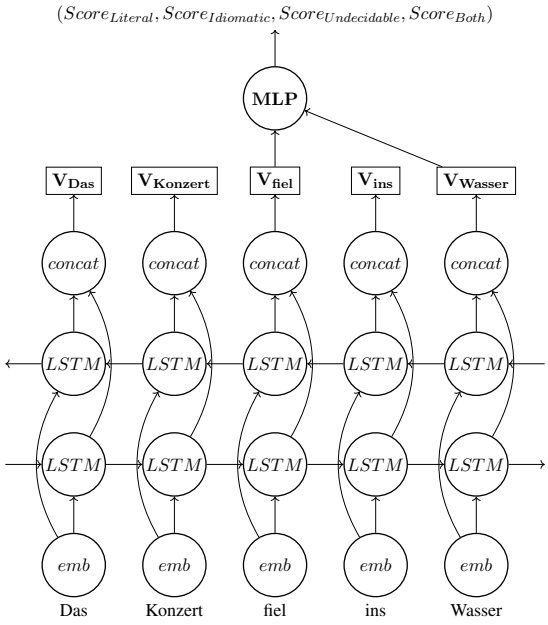

Figure 3: Architecture of the neural model

with LSTM (Hochreiter and Schmidhuber, 1997) units (BiLSTM) in order to receive contextualized representations v_i of each input element w_i:

$$v_i = LSTM_{\theta_F}(x_{1:n}, i) \circ LSTM_{\theta_B}(x_{1:n}, i)$$

The contextualized representation v_i is the concatenation (denoted by $\circ$) of the outputs computed by the forward ($LSTM_{\theta_F}$) and backward ($LSTM_{\theta_B}$) LSTM. Hence, v_i ideally contains information about all the preceding and succeeding items.

We then take two of those vectors, namely those for the verb and noun of the potential VID[13], concatenate them and feed the result into a multi-layer perceptron (MLP) to obtain the final scores:

$$SCORE(v_i \circ v_j) = MLP(v_i \circ v_j)$$

where v_i and v_j are the contextualized representations of the verb and the noun of the potential VID, respectively. We did not include prepositions into the input for the final scoring, because some expressions in COLF come without a lexicalized preposition (even though most do).

Till now, we only considered Word2vec and fastText embeddings as inputs. However, for ELMo things are a bit different on the input level. While Word2vec and fastText are functions that map each word to exactly one embedding, ELMo assigns different embeddings to the same word, depending on its context:

$$x_i = ELMo(w_{1:n}, i)$$

This means, we introduce context already at the very beginning, which we assume is a great advantage for the system, since the components of the candidates receive different vectors depending on their context. E.g. during the classification process with Word2vec or fastText embeddings, the word *ice* in the sentences *The weight of the ship broke the ice* and *With a joke he broke the ice* would receive the same vector, while ELMo should assign them different representations.

Training and Hyperparameters We split the COLF-VID dataset into train (70%), validation (15%) and test (15%) data. During the split we had to consider the high variance of the number of instances per VID type as to make sure that every split mirrors the distribution of types in the original data. E.g. *am Boden liegen* (48 instances) and *auf dem Tisch liegen* (951 instances) are represented with the same ratio in all three data sets.

The objective of the training was to minimize the cross entropy loss and for optimization we used the gradient descent variant Adam with a learning rate of 0.01. As for the labels we chose the majority annotation. We trained the models for 15 (Word2vec, fastText) respectively 18 (ELMo) epochs with a batch size of 30. The input size of our models was dependent on the dimensionality of the pretrained embeddings which had 100 (Word2vec), 300 (fastText) and 1024 (ELMo) dimensions. The forward and backward LSTMs were one-layered and the size of the hidden state was 100 for all three models, despite the considerable difference in input sizes which could have warranted testing larger hidden states for larger embeddings. But we refrained from doing so to keep the numbers of parameters in the MLP constant and thereby the model computationally less expensive. Hence, the MLP itself had an input size of 400 for all models, coupled with a hidden layer of size 100 and an output layer of size 4. The implementations of the three models are available on GitHub.[14]

5.2 Results

In this section we will present the results of our experiments on the disambiguation of German VIDs in context (see Table 2). We report precision, recall

[13] Remember, we assume for this task that another process already has identified the candidate expressions.

[14] https://github.com/rafehr/colf-bilstm-classifier

Validation set:

	class idiomatic			class literal			weighted macro average			
Model	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Acc**
Majority baseline	75.39	100.00	85.97	0	0	0	56.78	75.32	64.75	75.32
Word2vec+LSTM+MLP	90.60	90.25	90.42	70.47	72.76	71.60	85.30	85.59	85.44	85.59
fastText+LSTM+MLP	91.77	92.85	92.31	77.41	75.20	76.29	87.86	88.14	87.99	88.14
ELMo+LSTM+MLP	90.70	96.36	93.44	85.71	70.73	77.51	89.05	89.71	89.14	89.71

Test set:

	class idiomatic			class literal			weighted macro average			
Model	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Acc**
Majority baseline	76.95	100.00	86.98	0	0	0	59.22	76.95	66.93	76.95
Word2vec+LSTM+MLP	90.40	87.38	88.86	61.05	69.66	65.07	83.17	82.76	82.88	82.76
fastText+LSTM+MLP	91.23	93.94	92.56	77.42	71.79	74.50	87.45	88.29	87.83	88.29
ELMo+LSTM+MLP	93.70	93.94	93.82	78.24	79.91	79.07	89.54	90.10	89.82	90.10

Table 2: Evaluation results

Test set:

			ELMo+LSTM+MLP		
VID	**#**	I%	**Pre**	**Rec**	**F1**
am Boden liegen	8	23.4	77.50	87.50	81.94
an Glanz verlieren	3	93.75	100.00	100.00	100.00
an Land ziehen	39	90.38	100.00	100.00	100.00
am Pranger stehen	1	100.0	100.00	100.00	100.00
den Atem anhalten	6	75.0	88.89	83.33	83.81
auf dem Abstellgleis stehen	4	42.31	56.25	75.00	64.29
auf den Arm nehmen	14	42.31	93.88	92.86	92.89
auf der Ersatzbank sitzen	4	23.81	50.00	50.00	50.00
auf der Straße stehen	38	62.4	87.30	86.84	87.00
auf der Strecke bleiben	94	99.19	97.88	98.94	98.41
auf dem Tisch liegen	143	71.29	89.13	90.21	89.44
auf den Zug aufspringen	19	96.03	100.00	94.74	97.30
eine Brücke bauen	53	68.39	92.45	92.45	92.45
die Fäden ziehen	26	94.8	90.86	88.46	89.49
im Blut haben	6	19.44	100.00	100.00	100.00
in den Keller gehen	19	72.8	95.18	94.74	94.68
in der Luft hängen	43	90.14	89.24	88.37	88.74
im Regen stehen	57	79.68	82.42	85.96	84.09
ins Rennen gehen	10	82.26	64.00	80.00	71.11
in eine Sackgasse geraten	16	98.02	100.00	100.00	100.00
im Schatten stehen	9	86.67	100.00	100.00	100.00
in Schieflage geraten	7	90.91	100.00	100.00	100.00
ins Wasser fallen	38	73.52	92.98	89.47	90.15
Luft holen	26	38.82	83.85	76.92	75.11
mit dem Feuer spielen	13	87.06	85.21	92.31	88.62
einen Nerv treffen	43	99.65	100.00	100.00	100.00
die Notbremse ziehen	49	84.36	92.09	89.80	90.51
eine Rechnung begleichen	38	64.54	78.95	78.95	78.95
von Bord gehen	14	51.61	82.86	71.43	70.24
vor der Tür stehen	90	68.17	83.20	82.22	82.54
ein Zelt aufschlagen	15	41.0	76.67	66.67	65.08
über Bord gehen	18	45.22	84.03	88.89	86.30
über Bord werfen	67	87.81	98.66	98.51	98.54
über die Bühne gehen	20	99.0	90.25	95.00	92.56

Table 3: Evaluation results (weighted macro) per VID on the test set.

and F1-score for the two classes with the most instances – IDIOMATIC and LITERAL – as well as the weighted macro-average for all classes combined. Since there was such a low number of instances with the labels UNDECIDABLE and BOTH for the system to train on (only 28 in the train set), it did not do well on those classes which it always misclassified. In order to account for this stark imbalance in classes, we settled for the weighted macro average instead of the normal macro average and did not include detailed (precision/recall/F1) scores for the two low-number classes.

Overall Results As a baseline we chose a simple majority classifier which already represents a non-trivial hurdle, because of the high idiomaticity rate of COLF-VID. Still, with respect to the F1-score, our system clears it with all three different input types and shows some considerable improvements. Furthermore, as was our hypothesis, the fastText embeddings were an enhancement over Word2vec, which in turn were bested by ELMo. Table 2 shows the increased performance across both classes for the validation and the test set. The highest F1-score on the validation (89.14) and the test (89.82) set were achieved when using ELMo embeddings.

We suspect the superiority of fastText and ELMo over Word2vec lies in the fact that the two former models incorporate subword information. This should allow the classifier to detect morphosyntactic features that give clues on the correct reading of an expression, e.g. when it encounters a form of inflection unusual for a VID which tends to be morphosyntactically fixed. This is something our Word2vec model cannot accomplish, since it was trained on lemmas. Also, it would have been surprising if ELMo's ability to handle polysemy would not have been an advantage in a disambiguation task. This way context is already introduced at the input level.

One apparent weakness of our system is its weaker performance on the LITERAL class in comparison to the IDIOMATIC class – hardly a surprise when considering the unbalanced distribution of labels. Still, a maximum F1-score of 79.07 for LITERAL shows that our efforts to keep the idiomaticity rate of COLF-VID low bear some fruit.

VID-specific Evaluation Table 3 shows a more fine-grained evaluation of the best performing system by listing the results per VID on the test set. The classifier achieves its best results (100.00 F1-score) for *an Glanz verlieren*, *an Land ziehen*, *am Pranger stehen*, *im Blut haben*, *in eine Sackgasse geraten*, *im Schatten stehen*, *in Schieflage geraten* and *einen Nerv treffen*. That was to be expected, since all these VIDs have a high rate of idiomatic or literal readings – a fact the classifier very likely learnt during training, thus assigning a higher probability to the majority label. Nonetheless, even for those VID types it does not seem to mindlessly apply one label all the time. E.g. for *an Land ziehen* and *im Blut haben*, it correctly classifies the relatively few instances of their respective minority class.

Still, arguably the most interesting VID types with respect to the disambiguation task are those with a (relatively speaking) more balanced distribution of classes, like *auf der Straße stehen*, *auf dem Tisch liegen*, *eine Brücke bauen*, *in den Keller gehen*, *im Regen stehen ins Wasser fallen*, *Luft holen*, *eine Rechnung begleichen*, *von Bord gehen*, *vor der Tür stehen*, *ein Zelt aufschlagen* or *über Bord gehen*, all of which have idiomaticity rates between 38.82% and 79.68%. For all but four of those expressions, the system achieves F1-scores between 82.54 and 94.45. For *ein Zelt aufschlagen* (65.08), *von Bord gehen* (70.24), *Luft holen* (75.11) and *eine Rechnung begleichen* (78.95), the F1-scores are below 80. It would be interesting to investigate whether the difference in performance for the various VID types correlates with the inter-annotator agreement (IAA). We leave this question to future work.

6 Conclusion/Future Work

In this paper we presented COLF-VID, a new corpus with annotated instances of German VIDs and their literal counterparts. Furthermore, we experimented with VID disambiguation on the new corpus and showed that significant improvements can be gained from applying a neural architecture in comparison with a simple majority baseline. The experiments additionally demonstrated the effects of the different word representations on the resulting performance.

For the future we plan on extending the annotation of COLF-VID with those VIDs that were not in the set of pre-chosen expressions and consequently were not annotated. This would allow to use the corpus as a basis for an identification task and not just disambiguation. Concerning the disambiguation task itself, a cornucopia of different approaches – be it supervised or unsupervised – can be imagined. We plan on conducting a survey of different approaches in an attempt to reveal which architectures, context sizes and features are best suited for the task. Last but not least, cross-linguistic experiments with comparable corpora (e.g. IDIX) could be interesting in order to explore language-specific properties of VIDs.

Acknowledgements

We would like to thank Julia Fischer and Kevin Pochwyt for their annotations. We also would like to thank the anonymous reviewers for their helpful reviews and suggestions. This work has been supported by the Deutsche Forschungsgemeinschaft (DFG) within the CRC 991 "The Structure of Representations in Language, Cognition, and Science".

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing*, 2:267–292.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

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.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22.

Rafael Ehren. 2017. Literal or idiomatic? identifying the reading of single occurrences of German multiword expressions using word embeddings. In *Proceedings of the EACL Student Research Workshop*, pages 103–112, Valencia, Spain.

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

Fabienne Fritzinger, Marion Weller, and Ulrich Heid. 2010. A survey of idiomatic preposition-noun-verb triples on token level. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2908–2914, Valletta, Malta. European Language Resources Association (ELRA).

Hessel Haagsma and Johannes Bjerva. 2016. Detecting novel metaphor using selectional preference information. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 10–17.

Julia B. Herrmann. 2013. *Metaphor in academic discourse: Linguistic forms, conceptual structures, communicative functions and cognitive representations*. Phd thesis, Vrije Universiteit Amsterdam, Amsterdam.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Andrea Horbach, Andrea Hensler, Sabine Krome, Jakob Prange, Werner Scholze-Stubenrecht, Diana Steffen, Stefan Thater, Christian Wellner, and Manfred Pinkal. 2016. A corpus of literal and idiomatic uses of German infinitive-verb compounds. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 836–841, Portorož, Slovenia. European Language Resources Association (ELRA).

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. ACL.

Linlin Li and Caroline Sporleder. 2009. A cohesion graph based approach for unsupervised recognition of literal and non-literal use of multiword expressions. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, pages 75–83, Suntec, Singapore.

Philip May. 2019. German ELMo Model.

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*, pages 3111–3119.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Praglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Verginica Barbu Mititelu, Voula Giouli, Ivelina Stoyanova, Nathan Schneider, and Timm Lichte. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of Verbal Multiword Expressions. In *Proceedings of LAW-MWE-CxG 2018*, Santa Fe, USA.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta, and Voula Giouli. 2019. Literal occurrences of multiword expressions: Rare birds that cause a stir. *The Prague Bulletin of Mathematical Linguistics*, 112(1):5–54.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 486–493, Istanbul, Turkey. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1497.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1002–1010. Association for Computational Linguistics.

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).

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification*. Number 14 in Converging Evidence in Language and Communication Research. John Benjamins, Amsterdam, The Netherlands.