Shared Task on the Disambiguation of German Verbal Idioms at KONVENS 2021

Rafael Ehren[1], **Timm Lichte**[2], **Jakub Waszczuk**[1], **Laura Kallmeyer**[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

The processing of multiword expressions (MWEs) has gained a lot of attention in recent years. Not least thanks to the shared tasks on verbal MWE identification organized by the PARSEME network. A phenomenon that because of its inherently rare nature is quite infrequent in datasets such as the PARSEME corpus are literal readings of verbal idioms (VIDs). This makes it difficult to efficiently train systems capable of identifying them. To alleviate this issue for German we built a VID corpus with a higher than usual number of literal readings. Together with the SemEval 2013 Task 5b dataset this formed the corpus for the Shared Task on the Disambiguation of German Verbal Idioms at KONVENS 2021. This paper describes the organization of the competition as well as the results of the participating systems.

1 Introduction

The processing of multiword expressions (MWEs) has gained a lot of attention in recent years. Not least thanks to the interdisciplinary scientific network PARSEME (PARSing and Multiword Expressions), which is concerned with MWEs in parsing and as of yet has organized three different tasks on the identification of verbal MWEs (VMWEs) (Savary et al., 2017; Ramisch et al., 2018, 2020). The term *MWE identification* denotes the automatic annotation of MWEs in a text by a given system. A subtask of this process is the *disambiguation* of idiomatic MWEs and their literal counterparts (if existent), e.g. in *John kicked the bucket* John could have struck a pail with its foot or passed away. Only in the latter case, the string *kicked the bucket* should be annotated as a VMWE, thus it would not be enough for a system to just compare input to a MWE lexicon and apply the annotation to every string that matches, but it has to rely on context[1] or morphosyntactic clues. Obviously, there are several NLP tasks, like semantic parsing or machine translation (MT), where an erroneous disambiguation would affect results negatively.

While the sheer amount of training data available to some applications (like MT) might be enough to solve this task, low(er) resource applications lack the training data to do so. After all, as Savary et al. (2019) showed, literal counterparts of VMWEs are quite a rare phenomenon. In the German dataset of the PARSEME corpus, only 2% of potentially idiomatic expressions[2] (PIEs) have a literal reading. This stark imbalance significantly aggravates the efficient training of classifiers which are able to correctly distinguish VMWEs and their literal counterparts. In order to alleviate this issue for German, we created COLF-VID (Corpus of Literal and Figurative Readings of Verbal Idioms), which has a much lower than usual idiomaticity rate of roughly 78% (Ehren et al., 2020). Together with the SemEval-2013 Task 5b dataset (Korkontzelos et al., 2013) this formed the corpus for the present shared task. The goal of this shared task was to invite other members of the community to use our data as a testbed for their disambiguation methods.

2 Related Work

In order to avoid redundancy, in this section, we will limit ourselves to the description of similar shared tasks, since Ehren et al. (2020) already contains a compilation of related methods and corpora.

The most well-known shared tasks concerning VMWE identification are those organized by the PARSEME network (Savary et al., 2017; Ramisch et al., 2018, 2020). The PARSEME corpora provided for these competitions are highly multilingual

[1]Sometimes beyond the sentence itself as our example illustrates.

[2]A term coined by Haagsma et al. (2020).

(up to 20 languages) and seek to cover all types of VMWEs. Participating systems were ranked according to their performance regarding VMWE types and tokens. While a system had to label all components of a VMWE instance to have it included in the type score, it was enough to label only a few of the components to influence the token score. Furthermore, in addition to a *closed* track, there was an *open* track where participants could draw on resources not provided by the organizers.

It is noticeable that methods relying on syntactic information and parsing performed well during the shared tasks. As in all of NLP, from edition to edition, neural architectures increasingly found their way into the competition. Most participating systems had in common that the token based score usually exceeded the type based score, as well as a highly varying performance across the languages. The same goes for unseen vs. seen VMWE types, with systems achieving (much) worse results for unseen types across the board. This is why the organizers focused on this latter aspect during their latest edition of the shared task (1.2) by ranking their systems according to their performance on unseen types.

Although related, the PARSEME competitions differ in a few aspects from our task. The most important one being that they are about identification of VMWEs, thus the disambiguation of PIEs is only a byproduct. In addition, as already discussed in the introduction, the PARSEME datasets are not really suitable for training disambiguation systems anyway.

A competition that – like ours – is exclusively concerned with the disambiguation of PIEs is the SemEval 2013 task on the evaluation of phrasal semantics (Korkontzelos et al., 2013). More precisely, subtask 5b which is to decide on the compositionality of phrases in a given context. The shared task had two different settings: one for *known phrases* and one for *unknown phrases*, i.e. instances of phrases not seen during training. As expected, the results for unknown phrases were much worse and barely beat the majority baseline. Although an English and a German corpus were provided for the competition, only results for English were reported. Thus it appears, no results for German were submitted. Since we decided to incorporate the German corpus into our dataset, we will describe it more thoroughly in Section 3.2.

3 Data

The corpus for the shared task is a merger of two datasets, COLF-VID and the dataset for the SemEval 2013 task 5b. In this section, we will describe these two datasets.

3.1 COLF-VID

COLF-VID is a lexical sample corpus, which means that it only contains instances of a prechosen set of VID types. It was constructed this way because, as we have seen, literal readings of PIEs are quite rare and large amounts of data would be needed to get enough training examples without any pre-selection. In the set of the VID types, we included those we thought would increase the number of literal readings. Every sentence in COLF-VID was extracted from the German newspaper corpus TüPP-D/Z, thus it is a very homogeneous corpus regarding the genre. While the first version of COLF-VID, as described in Ehren et al. (2020), had 6985 sentences, the current version has 40 less, because during the clean-up of the data some duplicates were found and removed. Furthermore, the original corpus only contained one sentence per PIE instance, however, in order to align it with the SemEval dataset, the two surrounding sentences were also included. The data was annotated by three annotators with rather high Cohen's Kappa scores (0.77, 0.8, 0.9), and the labels they applied were LITERAL, FIGURATIVE, UNDECIABLE and BOTH.

3.2 SemEval 2013 5b dataset for German

The SemEval 2013 5b dataset for German is very similar to COLF-VID concerning a lot of aspects: it contains annotations for the different readings of PIEs, it is a lexical sample corpus, and more or less the same label set is used[3]. It was also annotated by three annotators with a very high pairwise agreement of 90% to 95%.

The main differences are the inclusion of nonverbal PIEs (like *steif und fest* ('stubbornly') or zweite Geige ('second fiddle')), the size (2961 instances) and, most importantly, a much higher idiomaticity rate. As can be seen in Table 1, after filtering out the nonverbal PIEs, it has an idiomaticity rate of 93.58, and no instances labeled as IMPOSSIBLE or BOTH were left. Another difference is the origin of the data: the sentences were extracted from the

[3]The label IMPOSSIBLE is used instead of UNDECIABLE if the context is unsufficient to decide on the reading.

deWaC corpus, which is a web corpus that was built by crawling the .de domain, thus it is much more heterogeneous than COLF-VID.

Since it seems, as mentioned above, that no results were submitted for the German dataset during the SemEval 5b shared task, it makes sense to incorporate it into ours, so it can serve its intended purpose.

3.3 Combined datasets

The similarities of the two corpora made it very easy to combine them. All we had to do was to filter out the non-verbal MWEs and align the formats. Since, in contrast to COLF-VID 1.0, the SemEval 2013 5b dataset not only comprises the sentence containing the PIE, but also the two surrounding sentences, we added the same amount of context to COLF-VID.[4] Thus, every instance in the combined dataset consists of three sentences: the sentence with the PIE and the preceding and succeeding sentence. In total, the merged corpus comprises 9901 instances of 67 VID types. When looking at Table 2, it is conspicuous that some types seem to be duplicates, e.g. *mit Feuer spielen* vs. *mit dem Feuer spielen* ('to play with (the) fire'⇒'to put oneself in a dangerous situation'). The reason is that, while in the SemEval dataset only the canonical form was annotated, in COLF-VID instances of *mit Feuer spielen* did not necessarily have to include the determiner. We opted to treat these as two different types, because we wanted to preserve the integrity of the SemEval dataset as much as possible.

The data was split according to a 70/15/15 ratio. Since the numbers of examples per VID type varies strongly (see Table 2), we had to ensure that the same ratio was applied to each type and not to the dataset as a whole in order to prevent an imbalance of types in the split dataset. Furthermore, to challenge the ability of the systems to generalize, we added instances of 3 unseen VID types to the dev and the test set, respectively (270 to test, 268 to dev). This resulted in a train set with 6902, a dev set with 1488 and a test set with 1511 instances.

Even though the Semeval dataset's idiomaticity rate is rather high, the idiomaticity rate of the combined dataset is with 82.39% still much lower than usual.

[4] An obvious shortcoming of COLF-VID we aimed to address anyway.

	Lit.	Idiom.	Und.	Both	I%
COLF-VID	1511	5386	33	10	77.61
SemEval 5b data	190	2771	0	0	93.58
Total	1701	8157	33	10	82.39

Table 1: Total dataset statistics

4 Task

At this point, it is important to highlight the differences to the identification task as it is performed in the context of the PARSEME shared tasks. In contrast to identification, during the pure disambiguation task we assume that another process has already pre-identified the PIEs. Figure 1 shows an example from the dataset for the VID type *ins Wasser fallen* ('to get canceled'⇒'to fall into the water') [5].

```
T890202.28.4077 \t in wasser fallen
\t figuratively \t Der Streit ums
Hormonfleisch zwischen USA und EG
provozierte den Polizeieinsatz . Aber
nicht nur der Steakverkauf , auch die
Aktionen gegen den Hormonstand , auf
die sich Gruppen der Bauernopposition
schon vorbereitet hatten , <b>fielen</b>
<b>ins</b> <b>Wasser</b> . Die
Fleischexporteure der USA wollten
ihrerseits die " Gruene Woche " zur "
Aufklaerung " nutzen .
```

Figure 1: A sample from the corpus.

We can see that the PIE components of *ins Wasser fallen* are alraedy marked with the <b> tag, so a system does not have to make this decision. It only needs to decide on the reading of that expression, or, to be more precise, on whether the expression is interpreted literally or figuratively. In theory, one could of course use the dataset for the identification task as well, but this could prove problematic because of the way it was built. First a set of VID types was compiled, then sentences that contained examples for those types were extracted from a corpus. But these sentences of course could comprise instances of VID types not present in the set of pre-chosen types. We want our system to learn to generalize and it would be confusing if some VID instances were labeled while others are not.

[5] Translation of the text in the sample: *The dispute over hormone meat between the US and the EEC provoked the police action. But not only the steak sale, but also the actions against the hormone stand, for which groups of the farmer's opposition had already prepared, were canceled. The US meat exporters, for their part, were eager to use the "Green Week" for "clarification".*

VID type	Lit.	Idiom.	Und.	Both	I%	VID type	Lit.	Idiom.	Und.	Both	I%
am Boden liegen	35	11	0	1	23.40	auf die Nase fallen	7	69	0	0	90.79
an Glanz verlieren	0	14	0	0	100.00	Korb bekommen	12	82	0	0	87.23
an Land ziehen	25	234	0	0	90.35	Auge zudrücken	8	89	0	0	91.75
am Pranger stehen	0	5	0	0	100.00	Dampf ablassen	5	103	0	0	95.37
Atem anhalten	10	30	0	0	75.00	die Stiefel lecken	2	10	0	0	83.33
auf Abstellgleis stehen	15	11	0	0	42.31	einen Korb geben	7	81	0	0	92.05
auf Arm nehmen	39	50	0	0	56.18	gute Karten haben	5	92	0	0	94.85
auf Ersatzbank sitzen	16	5	0	0	23.81	Handtuch werfen	6	99	0	0	94.29
auf Straße stehen	92	156	1	0	62.65	Hose anhaben	2	11	0	0	84.62
auf Strecke bleiben	4	610	1	0	99.19	im gleichen Boot sitzen	0	94	0	0	100.00
auf Tisch liegen	254	677	10	1	71.87	in den Sand setzen	8	87	0	0	91.58
auf Zug aufspringen	5	186	0	0	97.38	in den Schatten stellen	3	92	0	0	96.84
Brücke bauen	108	237	1	0	68.50	keinen Bock haben	0	91	0	0	100.00
Fäden ziehen	36	226	0	0	86.26	Korb kriegen	0	6	0	0	100.00
in Blut haben	29	7	0	0	19.44	mit dem Feuer spielen	3	76	0	0	96.20
in Keller gehen	33	89	0	0	72.95	rote Zahlen schreiben	0	104	0	0	100.00
in Luft hängen	28	256	0	0	90.14	über den Tisch ziehen	2	91	0	0	97.85
in Regen stehen	69	301	4	4	79.63	Braten riechen	6	84	0	0	93.33
in Rennen gehen	11	50	0	0	81.97	die Daumen drücken	0	95	0	0	100.00
in Sackgasse geraten	2	98	0	0	98.00	gegen den Strom schwimmen	0	80	0	0	100.00
in Schatten stehen	7	52	0	1	86.67	Geld zum Fenster hinauswerfen	1	25	0	0	96.15
in Schieflage geraten	3	39	1	0	90.70	Löffel abgeben	1	85	0	0	98.84
in Wasser fallen	66	183	0	0	73.49	heilige Kuh schlachten	1	83	0	0	98.81
Luft holen	99	66	4	0	39.05	Hut nehmen	6	69	0	0	92.00
Nerv treffen	1	282	0	0	99.65	im Geld schwimmen	0	99	0	0	100.00
Notbremse ziehen	57	367	0	1	86.35	ins Gras beißen	3	78	0	0	96.30
Rechnung begleichen	88	160	0	0	64.52	Öl ins Feuer gießen	0	99	0	0	100.00
von Bord gehen	45	48	0	0	51.61	schlechte Karten haben	4	96	0	0	96.00
vor Tür stehen	189	407	1	1	68.06	Rücken stärken	10	81	0	0	89.01
Zelt aufschlagen	52	40	7	1	40.00	Vogel abschießen	11	80	0	0	87.91
über Bord gehen	61	51	1	0	45.13	unter Strom stehen	23	65	0	0	73.86
über Bord werfen	54	389	0	0	87.81	mit Feuer spielen	9	73	2	0	86.90
über Bühne gehen	2	198	0	0	99.00	Frucht tragen	20	70	0	0	77.78
auf dem Schlauch stehen	1	83	0	0	98.81						

Table 2: Dataset statistics

Although, in theory, there are four different labels to predict, the skewness of their distribution makes it very unlikely that a system would factor in the labels UNDECIDABLE and BOTH. So effectively it is a binary task with the two classes LITERAL and IDIOMATIC.

5 Evaluation

The systems were ranked according to their F1-score for the literal class. As already mentioned, to test the generalizing capabilities of the systems, we added the instances of three VID types to the dev and test set (respectively) that were not present in the train set. The *F1-unseen* score reflects the performance of the systems with respect to those unseen types, but they affected the ranking only implicitly as the *F1-all* score was used to decide the ranking of the systems.

6 Shared Task Organization

The shared task was organized on CodaLab[6], an open-source web-based platform that is widely used for machine learning competitions. Since CodaLab ran low on storage space during our shared task, we hosted the data separately on GitHub[7]. CodaLab allows for two different submission modes: either participants submit their systems or only their system outputs, where in both cases the evaluation is conducted automatically. We opted for the latter submission mechanism. A modified version of our evaluation script can be found in the GitHub repository. The training phase went from May 15 to June 23, and the evaluation phase, during which participants could submit results for up to three systems, went from June 23 to June 30.

7 System Results

Five teams participated in the shared task and they submitted a total of 13 system prediction files. The results can be seen in Table 3. Three of those teams submitted a system description paper. The highest ranking system (*FranziskaPannach*) employed XLM-RoBERTa and a semi-automatic approach to extend the training data (Pannach and Dönicke, 2021). It was the only deep learning ar-

[6] https://competitions.codalab.org/competitions/31715

[7] https://github.com/rafehr/vid-disambiguation-sharedtask

	User	F1-all	F1-unseen
1.	FranziskaPannach	76.19	73.81
2.	JeanWayne	58.78	41.98
3.	PeterFankhauser	45.08	29.79
4.	rusaya	30.84	25.00
5.	alisentas	28.95	00.00

Table 3: Shared task results

chitecture that entered the competition. The team in second place (*JeanWayne*) used a decision tree-based classifier relying on features based on the notions of similarity and concreteness (Charbonnier and Wartena, 2021). The third placed team (*PeterFankhauser*) applied a shallow, statistics-based pipeline that was previously used to detect idioms in another corpus (Amin et al., 2021). The last two teams did not submit a system description paper.

When looking at the results, it is salient that all systems lost ground on unseen VID types. This was to be expected, since systems cannot rely on what types they memorized during training, but have to be able to generalize well. The winning system was the only one whose performance on the unseen types (73.81) came close to the performance on all types (76.19). Actually, the margin is surprisingly small and another testament to the strength of BERT-based architectures. To compare it to another system that leveraged information from a language model, we trained the BiLSTM-MLP architecture with ELMo embeddings from Ehren et al. (2020) on the shared task data. It achieved an F1-all score of 71.46 and an F1-unseen score of 52.05 on the test set, so obviously its generalizing capabilities are much weaker than those of the XLM-RoBERTa system. The main reason might be that the system of *FranziskaPannach* was fine-tuned for the task, while we only took the embeddings from ELMo to feed them into our architecture. Also, BiLSTMs have the reputation to have trouble with very long sequences (see Luong et al. (2015)) which might be a disadvantage, since every input in the shared task usually consisted of three full sentences. An attention based model as XLM-RoBERTa might be better suited for such long inputs.

8 Conclusion

In this paper, we described the Shared Task on the Disambiguation of German Verbal Idioms at KONVENS 2021. Five systems participated in the competition, three of which submitted a system description paper. All models performed much worse on the unseen VID types with the exception of the winning system that only lost a few points compared to the F1-all score.

Future work includes an exploration of how much context is really needed for the successful disambiguation of PIEs, i.e. we will concern ourselves with the question at which point a sequence might be too long. Another very important point is to examine data augmentation strategies, since even with nearly 10000 sentences the shared task corpus is quite small measured by today's machine learning standards. Finally, it would be desirable to develop methods that can handle VID identification and disambiguation jointly.

Acknowledgments

We would like to thank Julia Fischer and Kevin Pochwyt for their annotations of COLF-VID. We also would like to thank Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto and Chris Biemann for providing us with the German SemEval 5b dataset.

References

Miriam Amin, Peter Fankhauser, Marc Kupietz, and Roman Schneider. 2021. Shallow Context Analysis for German Idiom Detection. In *Proc. of the Shared Task on the Disambiguation of German Verbal Idioms at KONVENS 2021*.

Jean Charbonnier and Christian Wartena. 2021. Verbal Idioms: Concrete Nouns in Abstract Contexts. In *Proc. of the Shared Task on the Disambiguation of German Verbal Idioms at KONVENS 2021*.

Rafael Ehren, Timm Lichte, Laura Kallmeyer, and Jakub Waszczuk. 2020. Supervised disambiguation of German verbal idioms with a BiLSTM architecture. In *Proc. of Second Workshop on Figurative Language Processing*, pages 211–220.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proc.of LREC*, pages 279–287.

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.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Franziska Pannach and Tillmann Dönicke. 2021. Cracking a Walnut with a Sledgehammer: XLM-RoBERTa for German Verbal Idiom Disambiguation Tasks. In *Proc. of the Shared Task on the Disambiguation of German Verbal Idioms at KONVENS 2021*.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proc.of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions. In *Proc.of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118.

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.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME Shared Task on Automatic Identification of Verbal Multi-word Expressions. In *Proc. of the 13th Workshop on Multiword Expressions (MWE 2017)*.