To Leave No Stone Unturned: Annotating Verbal Idioms in the Parallel Meaning Bank

Rafael Ehren, Kilian Evang, Laura Kallmeyer
Heinrich Heine University Düsseldorf
Universitätsstr. 1, 40225 Düsseldorf, Germany
{rafael.ehren, kilian.evang, laura.kallmeyer}@hhu.de

Abstract

Idioms present many challenges to semantic annotation in a lexicalized framework, which leads to them being underrepresented or inadequately annotated in sembanks. In this work, we address this problem with respect to verbal idioms in the Parallel Meaning Bank (PMB), specifically in its German part, where only some idiomatic expressions have been annotated correctly. We first select candidate idiomatic expressions, then determine their idiomaticity status and whether they are decomposable or not, and then we annotate their semantics using WordNet senses and VerbNet semantic roles. Overall, inter-annotator agreement is very encouraging. A difficulty, however, is to choose the correct word sense. This is not surprising, given that English synsets are many and there is often no unique mapping from German idioms and words to them. Besides this, there are many subtle differences and interesting challenging cases. We discuss some of them in this paper.

Keywords: verbal idioms, semantic annotation

1. Introduction

Despite being one of the most discussed multiword expression (MWE) types, verbal idioms (VIDs) are surprisingly challenging to define. Actually, it seems to be easier to define them in terms of what they are not, as it is done by the PARSEME annotation guidelines (Ramisch et al., 2020)[1]. According to these guidelines, VIDs consist of a head verb and at least one lexicalized dependent which is neither a reflexive pronoun nor a particle. If the dependent is a verb or a noun, fine-grained tests need to be applied to discriminate the expression from multiverb expressions or light-verb constructions (LVCs). Another defining – and probably the most challenging – characteristic of an idiom is its non-compositionality, i.e. the meanings of its parts do not combine to form the meaning of the whole expression. However, since Nunberg et al. (1994), it is commonly acknowledged that there exists another dimension w.r.t. non-compositionality. We now make the distinction between decomposable and non-decomposable idioms. Both types are non-compositional, but for the former we can establish a mapping from its parts to their respective idiomatic meanings which in turn combine to form the meaning of the whole. Or, if we reverse the direction: We can decompose the idiomatic meaning and map these individual meanings to the components of the expressions.[2] This, however, is not possible for non-decomposable idioms whose meanings do not allow for this kind of distribution over their parts. For illustration, consider the following two classic examples:

(1) After a long interrogation the spy **spilled the beans.**

(2) After a long illness, he finally **kicked the bucket.**

Example (1) shows an instance of the idiom *spill the beans* which means 'to reveal a secret'. We consider this decomposable because the individual meanings can be mapped to the different components of the expression: 'reveal' to *spill* and 'secret' to *beans*. Such a mapping does not exist for 'kick the bucket' in example (2) because the idiomatic meaning 'to die' cannot be decomposed into individual meanings.

Because of this behavior, non-decomposable idioms are more challenging when it comes to semantic annotation (and consequently semantic parsing) than decomposable ones. For the latter, there exists a one-to-one mapping from words to concepts, but not for the former. This might be the reason why they are often ignored during semantic annotation and receive a literal treatment. Consider the following example from the English partition of the Parallel Meaning Bank (PMB):

(3)

[1] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=050_Cross-lingual_tests/030_Verbal_idioms__LB_VID_RB_

[2] Nunberg et al. (1994) spoke of idiomatically combining expressions, which reflects the initial direction of the analysis (starting from its parts), but since then the terminology changed in order to favor the other direction (starting from the whole expression).

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@*LREC-COLING-2024*, pages 115–124. May 25, 2024.

x_1, e_1
pull.v.01(e_1), Agent(e_1, hearer), Theme(e_1, x_1), leg.n.01(x_1), Of(x_1, speaker)

Discourse representation structure (DRS) for English PMB sentence 01/1871 *Are you pulling my leg?* (not gold).

The non-decomposable idiom *pull sb's leg* has the meaning 'to tease sb', but in the DRS above it is treated literally as *leg* is a discourse referent (x_1) which it should not be. Thus, the DRS actually represents a leg pulling event which is not the desired analysis in this case.

The goal of this work is to improve the coverage of VIDs in the PMB, so that ultimately semantic parsers trained on its data can benefit from it. Furthermore, as a byproduct, we created a dataset of potentially idiomatic expressions (PIEs; Haagsma et al., 2020), since we also labeled instances of literal counterparts of VIDs. This will be further elaborated at the end of section 4.

The structure of the paper is as follows: First, we will discuss related work and the PMB. Then, we will detail the extraction of candidate sentences and the annotation process. Finally, we will present the results and discuss especially challenging cases before we draw our conclusions.

2. Related Work

Arguably the most well-known MWE corpora are the four editions (1.0–1.3) of the PARSEME corpus (Savary et al., 2015); (Ramisch et al., 2018, 2020; Savary et al., 2023). What sets them apart from other corpora is their scope and homogeneity: The PARSEME corpora consist of a large number of datasets from different languages that were all annotated for verbal MWEs according to the same annotation guidelines. PARSEME corpora are not sense annotated, but these guidelines are highly relevant to us, too, as we used their definitions of the different verbal MWE types to decide which candidate expressions to annotate.

A corpus that contains semantic annotation of MWEs is the STREUSLE corpus (Schneider and Smith, 2015). It is a 55,000 words English web corpus consisting of reviews which were annotated for MWEs, but without restrictions to specific kinds of syntactic constructions. Furthermore, it distinguishes between *strong* and *weak* expressions, the former being opaque idioms (*shoot the breeze*) while the latter are more transparent collocations (*traffic light*). On top of that, they added a level of supersenses which are the top-level hypernyms in the WordNet taxonomy. There is no explicit mention of decomposable and non-decomposable idioms, but the aforementioned *strong* expressions receive a supersense as a unit while *weak* ones do not. So it is probable that non-decomposable expressions received the appropriate treatment w.r.t. to supersense tagging. However, since there were no guidelines to differentiate decomposable and non-decomposable idioms, it is not unlikely that some of the former were annotated as strong and thus erroneously received a holistic treatment.

Sembanks (corpora with deep meaning representations) treat idioms in different ways. Abstract Meaning Representations (AMR; Banarescu et al., 2013) and Uniform Meaning Representations (UMR; van Gysel et al., 2021) are not lexically anchored, so usually introduce a single concept node for an idiom consisting of several words (Bonn et al., 2023). On the other hand, sembanks with lexical anchoring need explicit mechanisms for dealing with cases where the word-concept mapping is not one-to-one, such as idioms. For HPSG, such mechanisms have been proposed, e.g. by Richter and Sailer (2014), but not, to our knowledge, applied in sembanks such as LinGO Redwoods (Oepen et al., 2002).

3. The PMB

The Parallel Meaning Bank (PMB; Abzianidze et al., 2017, 2020) is a partially parallel corpus of text in English, German, Italian, and Dutch, with semantic annotations. These include WordNet senses (Fellbaum, 1998) and VerbNet semantic roles (Kipper Schuler, 2005), among others. All semantic annotation layers are integrated into a meaning representation language based on Discourse Representation Theory (Kamp and Reyle, 1993) which places more emphasis than other frameworks such as AMR on precisely representing the scope of quantifiers as well as modal and logical operators. The semantic representations in this formalism are called Discourse Representation Structures (DRS).

The PMB is built using a dynamic annotation methodology (Oepen et al., 2002) based on a strongly lexicalized theory of the syntax-semantics interface. Statistical models produce an initial syntactic analysis of each sentence using Combinatory Categorial Grammar (CCG; Steedman, 2001) as well as an assignment of semantic tags, roles, senses, etc. to tokens. These annotation layers are corrected by human annotators by adding constraints called *bits of wisdom*. Bits of wisdom are stored in a database so they can be automatically reapplied to the output of the new versions of the statistical models in the future. The result is then fed into a rule-based component named Boxer which assigns a partial meaning representation (λ-DRS) to each token and then computes a DRS for the entire sentence. Automatically pre-

annotated documents are said to have 'bronze' status, documents with at least one bit of wisdom are 'silver', and documents marked as completely corrected by a human are 'gold'.

While the syntax-based annotation methodology of the PMB helps ensure consistency, it is challenged by multiword expressions where the mapping between lexical meanings and tokens is not one-to-one. Some types of verbal multiword expressions are already handled adequately. For example, in the verb-particle construction (4) and in inherently reflexive verbs (5), the meaning is assigned to the head, and the other element is treated as semantically empty. Decomposable verbal idioms as in (6) are treated by assigning each component a suitable non-literal meaning. Of course, this is only true for documents that have already been annotated by humans; the automatic pre-annotation usually fails to pick correct non-literal senses, as shown for a German idiom in (7). Furthermore, not much attention has so far been given to light verb constructions and non-decomposable idioms. As a result, most sentences containing such constructions do not have a gold annotation in the PMB yet, but only an automatically generated (i.e., bronze status) and semantically inadequate annotation using a literal sense of each word. Examples of this are shown in (8) and (3).

(4)

x_1, e_1, t_1
wedding.n.01(x_1), take_place.v.01(e_1), Theme(e_1, x_1), Time(e_1, t_1), DayOfWeek(t_1, saturday)

DRS for English PMB sentence 01/2506 *The wedding will take place on Saturday* (gold).

(5)

¬	s_1 ashamed.a.01(s_1), Experiencer(s_1, speaker)

DRS for German PMB sentence 03/2800 *Ich schäme mich nicht* "I'm not ashamed" (gold).

(6)

¬	x_1, e_1 spill.v.05(e_1), Agent(e_1, hearer), Theme(e_1, x_1), secret.n.01(x_1)

DRS for English PMB sentence 11/0958 *Don't spill the beans* (gold).

(7)

x_3, x_4, s_1
Order(x_3, "inneren"), Role(x_3, x_4), person.n.01(x_3), schweinehund.n.01(x_4), Patient(s_1, x_3), besiegen.a.01(s_1)

Partial DRS for German PMB sentence 17/1163 *den inneren Schweinehund zu besiegen* "to overcome one's weaker self" (not gold).

(8)

x_1, e_1
take.v.01(e_1), Agent(e_1, speaker), Theme(e_1, x_1), bath.n.02(x_1)

DRS for English PMB sentence 58/2404 *I'm taking a bath* (not gold).

In this work, we aim to improve the coverage of idioms in the PMB. This requires creating annotation guidelines that capture the semantics of such cases adequately while still fitting in with the lexicalized annotation framework of the PMB. It furthermore requires looking for idiom instances in the PMB and targeting them for annotation.

4. Extraction

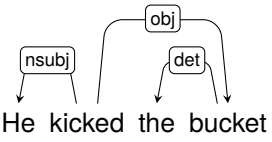
The first step was to find potential candidates for the annotation, i.e. sentences that contained German VID instances. To this end, we collected VID types from the *Redensarten-Index*[3] (transl. *Proverb-Index*), an electronic, privately maintained dictionary, which, contrary to the name, not only contains German proverbs but also an even larger number of idioms. At the time of this writing, the database comprises 15,661 entries. Since a lot of entries consist of several variants of the same expression, this number rises to 54,936 when counting every variant as a different type. After filtering out all the non-verbal expressions using parsing, 39,521 verbal ones remained.

After compiling a list of VID types, the next step was to find sentences in the PMB that contained instances of those VID types. We employed the parsing-based extraction method described in Haagsma (2020). This method only extracts sentences that contain the lemmata in the same dependency relations as the VID type, thus the focus of this approach is to increase precision by not extracting sentences that coincidentally comprise the same lemmata. Figure 1 shows two sentences that contain the tokens *kicked*, *the* and *bucket*, but only in (a) they have the desired dependency relations: NSUBJ between *bucket* and *kick* and OBJ between *kick* and *bucket*. In (b), the relation that holds between *kick* and *bucket* is OBL (for *oblique*) and accordingly the sentence would not be extracted, since it does not contain an instance of *kick the bucket* but only an accidental co-occurrence.

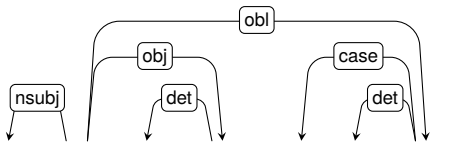
We employed UDPipe 2.12[4] (Straka, 2018) to

[3] https://www.redensarten-index.de/suche.php

[4] More specifically, the German model *german-gsd-ud-2.12-230717*



(a) PIE instance



(b) Not a PIE instance

Figure 1: Parsing-based extraction.

parse the gold, silver and bronze sentences of the German part of the PMB and subsequently used the method described above to extract sentences with VID candidates. This resulted in 6,187 sentences being extracted which were then prepared for annotation.

During this process not only instances of VID types were extracted, but also instances of their literal counterparts:

(9) Beth wurde von ihrem faulen Freund gefragt,
Beth was by her lazy friend asked,
ob sie **seine Hausaufgaben** für Geschichte
if she his homework for history
machen würde.
do would.
'Beth was asked by her lazy friend if she would do his homework for history.

In (9) we have an instance of *seine Hausaufgaben machen* (*to do one's homework*), but since it is the literal reading of this expression, we do not have an instance of the VID type (which means 'to prepare oneself'). These kind of literal instances are not relevant to the annotation of the PMB[5], but we decided to label them anyway in order to create a dataset of potentially idiomatic expressions (PIEs) as a byproduct. The term PIE encompasses both the literal and idiomatic meaning of an expression, thus we will use it from here on out when we talk about both at the same time.

5. Annotation

The annotation was conducted by three linguistically trained native speakers, with every sentence being annotated twice. Annotators were given text files where each instance to annotate came with a "form" with several questions they had to work through step by step (cf. Fig. 2).

In a first step, the guidelines were written and subsequently revised after a trial annotation of 50 sentences. However, due to the complex nature of the task, the guidelines kept on being revised multiple times throughout the whole process. To ensure consistency there was a subsequent correction step where every annotator revised their work once again. Weekly meetings with annotators were conducted throughout to discuss difficult cases and clarify the annotation guidelines.

The annotation consisted of several objectives:

1. Filter out false positives
2. Annotate the degree of idiomaticity
3. Judging the (non-)decomposability
4. Sense and role annotation

We will discuss these steps in more detail in the following.

Firstly, due to errors during the extraction and the fact that we did not filter the list of idiomatic expressions other than for verbal types[6], there was a large number of false positives, i.e. types of expressions not of interest to us. Our focus was exclusively on what can be considered verbal idioms (VIDs) or, in rarer instances, light-verb constructions according to the PARSEME annotation guidelines 1.2, so verb senses that are only considered "multiword" because they obligatorily occur with a certain function word were to be ignored. These include verb-particle constructions (VPCs, e.g. *jmdm. etwas **an**tun* 'do something to somebody'), and inherently adpositional verbs (IAVs, e.g. ***zu** jmdm. halten* 'stand by sb.'). As we have seen in Section 3, VPCs are already handled satisfactorily in the PMB, and likewise IAVs, where the adposition is treated as part of the argument and does not contribute a sense on its own. Furthermore, proverbs were also not considered as these do not have free argument slots, contrary to idioms (e.g. *A watched pot never boils.*).

In the next step, the annotators had to decide whether the PIE instance fell into one of the following categories: IDIOMATIC, PROBABLY IDIOMATIC, PROBABLY LITERAL, LITERAL or BOTH. We gave the annotators the possibility to express uncertainty with the qualifier *probably* in order to account for the fact that some sentences did not have enough context to allow for maximum certainty regarding the reading - even if the annotator happened to be rather sure[7]. The label BOTH was intended for

[5]Because they usually can be treated compositionally.

[6]Manually filtering a list of 39,521 expressions would have been too time consuming.

[7]For example, because a certain PIE type was known to have one predominant reading.

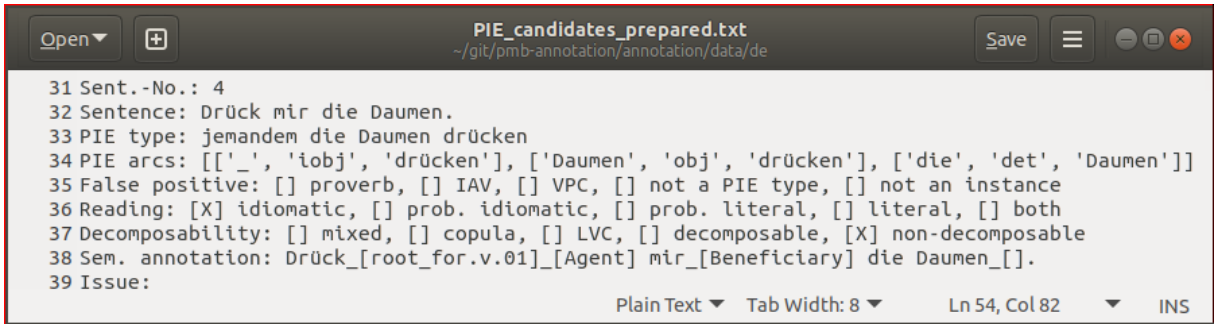


Figure 2: Text-based annotation interface, showing the sentence *Drück mir die Daumen!*, lit.: 'Squeeze your thumbs for me!', fig.: 'Wish me luck!'

cases in which both readings (IDIOMATIC and LITERAL) are active at the same time.

After that, the goal was to judge the level of decomposability of the expression. Besides the obvious labels, DECOMPOSABLE and NON-DECOMPOSABLE, the annotators could also choose the labels LVC, COPULA and MIXED. The latter three categories will be discussed in the next section in greater detail.

Strictly speaking, the previous step was not really necessary, but served as a kind of priming for the last step: the semantic annotation of the idiom and its arguments. During this step, the annotators were supposed to choose the WordNet sense (Fellbaum, 1998) that most closely corresponded to the meaning of the idiom and add it to the sentence. In order to do this, the annotators had to decide on the level of decomposability anyway because the number of senses added to the VID depended on this. Consider the next two examples for illustration:

(10) Er_[Experiencer] **schwimmt**_[buck.v.02]
He swims
gegen_[] **den**_[]
against the
Strom_[Stimulus]_[trend.n.01].
tide.
'He bucks the trend.'

(11) **Stecke**_[despair.v.01]_[Experiencer] nicht
Bury not
den Kopf_[] **in den Sand**_[]!
the head in the sand!
'Don't despair!'

Example (10) shows an instance of the VID *gegen den Strom schwimmen* (*swim against the tide* ⇒ 'buck the trend'), which is decomposable as we can map the individual idiomatic meanings to the components: 'buck' → *swim* and 'trend' → *tide*. Consequently, the two WordNet senses *buck.v.02* and *trend.n.01* were added. The example furthermore shows that in addition to the senses we also added the semantic roles of the predicate's arguments, in this case *Experiencer* and *Stimulus*. Annotators were instructed to use WordNet Search 3.1[8] for finding senses, and VerbAtlas (Di Fabio et al., 2019) for mapping them to VerbNet-style rolesets, but to prefer PMB-specific conventions when in doubt. As can be seen, the senses were added by suffixing an underscore followed by brackets to a component. If a component was annotated with a sense and a semantic role, the latter always preceded the former (first *Stimulus* then *trend.n.01* in this case).

In example (11), on the other hand, we have an instance of the non-decomposable VID *den Kopf in den Sand stecken* (*to put the head in the sand* ⇒ 'to despair'). It is non-decomposable as it is not possible to decompose the overall idiomatic meaning into individual meanings. For non-decomposable VIDs the WordNet sense (*despair.v.01* in this case) was added to the verbal head of the expression, while the other brackets were left empty.

Apart from VIDs we also annotated for LVCs as they are also not handled in the desired manner in the PMB:

(12) Die Generation_[Theme] der
The generation of
Zeitzeugen **geht**_[end.v.01]
contemporary witnesses goes
zu_[] **Ende**_[] [...]
to end [...]
'The Generation of contemporary witnesses is ending.'

Example (12) contains an instance of the LVC *zu Ende gehen* (*to go to end* ⇒ 'to end'). We consider this a special case of non-decomposability since no part of the meaning could ever be mapped to the semantically bleached verbal part. To ensure consistency we nevertheless add the sense

[8] http://wordnetweb.princeton.edu/perl/webwn

(*end.v.01*) to the verbal part of the expression. Please note that we did not annotate for expressions that according to the PARSEME annotation guidelines would be considered LVC.cause, i.e. the verb indicates the cause of the event (e.g. *to grant rights* or *to provoke a reaction*).

6. Annotation Results and Discussion

6.1. Inter-annotator agreement

For computing agreement, we excluded 341 sentences that had been discussed in annotation meetings, thus had not been annotated by two annotators independently. For simplicity, we also excluded 18 sentences that for various reasons did not have exactly 2 annotations and 7 sentences where one or both annotators detected more than one instance of the same idiom.

On the remaining 5,821 sentences, we classified annotators' decisions both broadly into "idiom" or "not an idiom", and more finely by, e.g. decomposability class or false positive class. On the coarse-grained comparison, annotators agreed in 3,448 cases that something is not an idiom and should thus not receive a detailed semantic annotation. In 1,945 cases they agreed it is an idiom. And in 428 cases they disagreed on this. Coarse-grained agreement is strong (Cohen's $\kappa = .8433$).

On the fine-grained comparison, annotators agreed in 4,230 cases and disagreed in 1,591 cases, yielding a moderate $\kappa = .6311$. Table 1 shows how frequent each class is, looking only at instances where annotators agree. We can see that most instances extracted are false positives, in particular cases where the extracted structure is not an instance of the idiom type, as in Figure 1b. Among the instances unanimously classified as idioms, a large majority is annotated as non-decomposable.

Table 2 shows the ten most frequently disagreed upon classes. In many cases, annotators agree that the items are not relevant to our annotation goal, they just disagree on why (e.g., IAV vs. not an instance). In other cases, annotators came to different conclusions regarding decomposability. Finally, there are cases where one annotator annotated the item as a non-decomposable idiom whereas the other deemed it not an instance, an IAV, not a verbal PIE type, or literal.

For the sense and role annotation of items that both annotators classified as an idiom, we look at whether both annotators selected the same word as the syntactic head of the idiom (head selection), whether they assigned the selected head the same sense (head sense classification), and for each word in the sentence whether they marked it as the head of an argument that is part of the (decomposable) idiom (internal argument identification), or as an argument that is not part of the idiom (external argument identification). For unanimously identified internal arguments, we also look at role and sense classification, and for unanimously identified external arguments, at role classification. Table 3 shows the results, with strong agreement for head selection and argument identification, weak to moderate agreement for head sense classification, and moderate to strong agreement for argument role and sense classification scores.

not an idiom	**3,448**
not an instance	1,968
IAV	194
VPC	149
proverb	142
literal	121
not a verbal PIE type	90
idiom	**1,945**
non-decomposable	1,335
decomposable	186
LVC	24
copula	19
mixed	2

Table 1: Unanimously classified PIEs by frequency. Numbers in bold represent coarse-grained agreement.

IAV, not an instance	349
literal, not an instance	195
decomposable, non-decomposable	181
non-decomposable, not an instance	**136**
LVC, non-decomposable	108
not a verbal PIE type, not an instance	91
IAV, non-decomposable	**73**
non-decomposable, not a verbal PIE type	**43**
IAV, literal	41
literal, non-decomposable	**39**

Table 2: Most frequent disagreements in PIE classification. Entries in bold are not only fine-grained but also coarse-grained disagreement.

6.2. Challenges to the annotation

In the following we will discuss some of the reasons that made the task quite challenging. As mentioned above, the guidelines were revised multiple times during the annotation process.

Decomposability One of these revisions consisted of adding another category w.r.t. decomposability. During the annotation it became clear that some expressions do not fit the binary distinction of decomposability presented above:

Head selection	.9769
Head sense classification	.5862
Internal argument identification	.9914
Internal argument role classification	.7296
Internal argument sense classification	.6824
External argument identification	.9845
External argument role classification	.8352

Table 3: Agreement scores for semantic annotation of idioms. Head selection is given in terms of raw agreement; the other scores are Cohen's κ scores.

(13) Tom_[Agent] **legte**_[reveal.v.02] **die**_[]
Tom laid the
Karten_[Topic]_[intention.n.01] **auf**_[]
cards on
den_[] **Tisch**_[].
the table.
'Tom revealed his intentions'.

Example (13) shows an instance of the VID *die Karten auf den Tisch legen* (*to lay the cards on the table* ⇒ 'to reveal one's intentions'). It is decomposable in the sense that we can map 'reveal' to *auf den Tisch legen* and 'intentions' to *Karten*, but there is no part of the meaning we can map to *Tisch* individually, i.e *auf den Tisch legen* itself is non-decomposable. To accomodate for these kind of instances, we added the category MIXED to the possible choices for decomposability.

Another frequently discussed question was whether to prioritize decomposition even when a non-decomposable analysis would have been more convenient because a very suitable sense was available:

(14) Der Gouverneur_[Agent] **setzte**_[set.v.05]
The governor set
die Häftlinge_[Patient] **auf freien**
the prisoners on free
Fuß_[Result]_[free.a.01].
foot.
'The governor set the prisoners free'.

Example (14) contains an instance of the VID *jmdn. auf freien Fuß setzen* (*to set sb. on free foot* ⇒ 'to set sb. free'), so the WordNet sense *set_free.v.01* would have been very fitting, but since we decided to prioritize the decomposition of the expression in such cases we opted for a decomposable analysis which seems less elegant.

Missing senses As one can imagine, it is not always straightforward to map a German idiom to an English WordNet sense. Sometimes there are two or more equally plausible possibilities, leading to spurious disagreement, e.g. *dazzle.v.02* or *stagger.v.04* for *jmdm. den Atem rauben* 'to take sb.'s breath away'. In case of missing verbal synsets, we were often able to use a nominal, adjectival, or adverbial one instead, as in (15).

(15) Dichter_[AttributeOf] wie Milton
Poets like Milton
sind_[rare.a.03] **dünn**_[] **gesät**_[].
are thinly sowed.
'Poets like Milton are few and far between.'

But sometimes we were hardly able to find any fitting sense at all.

(16) Tom **hat nichts zu verlieren**.
Tom has nothing to lose.
'Tom has nothing to lose.'

For example, the expression *nichts zu verlieren haben* 'to have nothing to lose' means something along the lines of being desperate and prone to dangerous behavior, but we were not able to find a synset capturing this, as, e.g. *desperate.a.03* seemed both too general and too specific, so we did not annotate (16), although in cases were we found a synset that was a bit too general but not too specific we usually accepted it, as in (17).

(17) er_[Agent] **gab**_[give.v.20] **ihm**_[Patient]
he gave him
einen tüchtigen Fußtritt_[Theme] **mit**_[]
a hearty kick with
auf_[] **den**_[] **Weg**_[]
on the way
'he gave him a good kick (as he was leaving)'

Some idioms have an emphatic meaning component not captured by the synset we assigned it, as in (18).

(18) Tom_[AttributeOf] **schwimmt**_[rich.a.01]
Tom swims
im Geld_[].
in the money.
'Tom is rolling in money.'

As a last resort when unable to find a roughly fitting synset, we would create a new one:

(19) Mir_[Experiencer] **fällt**_[cabin_fever.n.00]
Me falls
die_[] **Decke**_[] **auf**_[] **den**_[] **Kopf**_[].
the ceiling on the head.
'I'm starting to get cabin fever'.

The expression *jmdm. fällt die Decke auf den Kopf* (*the ceiling falls on sb's head*) alludes to the negative psychological effects someone can experience when confined to a small space for a long period of time. In English, the term *cabin fever* ex-

ists to describe this state, but it is not available in WordNet. And neither is any equivalent sense, so in such cases, we made a sense up which we suffixed with 00 (*cabin_fever.n.00* in (19)).

Collocations Lastly, the status of collocations was discussed frequently. Although we were not aware of it during annotation, we find the distinction between *idioms of encoding* and *idioms of decoding* (Fillmore et al., 1988; Richter and Sailer, 2014) helpful. Idioms of decoding are idioms proper: a listener has to know the expression to understand it, e.g. *ins Gras beißen*, lit. 'bite into the grass', 'kick the bucket'. Idioms of encoding require the speaker to know an expression to encode the meaning idiomatically, e.g. to know to say *Zähne putzen*, lit. 'clean teeth', 'brush teeth', and not *Zähne sauber machen*, lit. 'make teeth clean', although both encode the meaning compositionally and are understandable without having the expression in the mental lexicon. Mere idioms of encoding are sometimes called collocations, and were out of scope for this annotation project. But sometimes the difference is hard to tell.

(20) Endlich **zeigte** er **sein wahres Gesicht**.
Finally shows he his true face.
'Finally he reveals his real personality.'

(21) Wir sollten das wohl **unter vier**
We should that probably among four
Augen besprechen.
eyes talk about.
'We should probably discuss this in private.'

For example, in (20), one can argue that *sein wahres Gesicht zeigen* is an idiom of decoding because *Gesicht* with the sense *personality* is not often, perhaps never found outside of this expression, whereas *zeigen* with the sense *reveal* is quite common. Another example is shown in (21), where one can likewise argue that the adverbial phrase *unter vier Augen* in the sense *in private* usually only occurs with the verb *besprechen* or a small set of near-synonyms like *bereden*, *diskutieren*. We did not annotate these examples in the end and leave defining a sharper criterion for distinguishing idioms from collocations for future work.

7. Conclusions and Future Work

Idioms present many challenges to semantic annotation in a lexicalized framework, which leads to them being underrepresented or inadequately annotated in sembanks. In this work, we have carried out a targeted annotation of German idioms in the Parallel Meaning Bank by automatically detecting instances of potentially idiomatic expressions (PIEs) and annotating them for their idiomatic status, as well as their semantics, including WordNet senses and VerbNet semantic roles. Many automatically detected PIEs were false positives; of the rest, most received non-decomposable analyses, some decomposable ones, and some received special labels like MIXED, COPULA, or LVC. Inter-annotator agreement across the subtasks is very encouraging considering the complexity of the task, with the lowest score achieved for word sense disambiguation, unsurprising given that English synsets are many and there is often no unique mapping from German idioms and words to them. As our qualitative analysis of the results shows, there are also many subtle difficulties in classifying PIEs.

The next challenge will be to actually integrate the produced annotations into the PMB so as to get closer to a gold standard semantic annotation for sentences containing idioms. We are preparing a translation of the annotations into *bits of wisdom*, the format in which human annotator decisions are stored in the PMB and then inserted into the PMB's dynamic annotation workflow. Assigning senses and roles is relatively straightforward; however, for non-decomposable idioms, we also have to make sure that the arguments get assigned λ-DRSs that do not contribute concepts, which will require adding some new rules to Boxer, the rule-based component computing meaning representations based on syntax and token-level annotations. The documents receiving the annotations will automatically receive silver status and have to be checked manually again to receive gold status. This will make the PMB a more comprehensive and challenging testbed for data-driven DRS parsers such as van Noord et al. (2020) or Shen and Evang (2022), whose ability to handle idioms future work will also address. Furthermore, an analogous annotation project is currently underway for English idoms in the PMB.

Acknowledgments

We would like to thank the anonymous reviewers for their feedback. We would also like to thank our annotators for their work. This work was carried out in the MWE-SemPrE project funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 467699802.

8. Bibliographical References

Julia Bonn, Andrew Cowell, Jan Hajič, Alexis Palmer, Martha Palmer, James Pustejovsky,

Haibo Sun, Zdenka Uresova, Shira Wein, Nianwen Xue, and Jin Zhao. 2023. UMR annotation of multiword expressions. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 99–109, Nancy, France. Association for Computational Linguistics.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of *let alone*. *Language*, 64:501–538.

Hessel Haagsma. 2020. *A Bigger Fish to Fry: Scaling up the Automatic Understanding of Idiomatic Expressions*. Ph.D. thesis, Rijksuniversiteit Groningen.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Springer, Dordrecht.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538. Publisher: Linguistic Society of America.

Frank Richter and Manfred Sailer. 2014. Idiome mit phraseologisierten Teilsätzen: eine Fallstudie zur Formalisierung von Konstruktionen im Rahmen der HPSG. In Alexander Lasch and Alexander Ziem, editors, *Grammatik als Netzwerk von Konstruktionen: Sprachwissen im Fokus der Konstruktionsgrammatik*, pages 291–312. De Gruyter, Berlin, Boston.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, and Gyri Smørdal Losnegaard. 2015. PARSEME–PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*.

Minxing Shen and Kilian Evang. 2022. DRS parsing as sequence labeling. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 213–225, Seattle, Washington. Association for Computational Linguistics.

Mark Steedman. 2001. *The Syntactic Process*. The MIT Press.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207.

Jens E. L. van Gysel, Jayeol Chun Meagan Vigus, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmerand James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a uniform meaning representation for natural language processing. *Künstliche Intelligenz*, 35:343–360.

Rik van Noord, Antonio Toral, and Johan Bos. 2020. Character-level representations improve DRS-based semantic parsing even in the age of BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4587–4603, Online. Association for Computational Linguistics.

9. Language Resource References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Lasha Abzianidze, Rik van Noord, Chunliu Wang, and Johan Bos. 2020. The parallel meaning bank: A framework for semantically annotating multiple languages. *Applied mathematics and informatics*, 25(2):45–60.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637, Hong Kong, China. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Karin Kipper Schuler. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Stephan Oepen, Kristina Toutanova, Stuart Shieber, Christopher Manning, Dan Flickinger, and Thorsten Brants. 2002. The LinGO redwoods treebank: Motivation and preliminary applications. In *COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes*.

Carlos Ramisch, Silvio Cordeiro, Agata Savary, Veronika Vincze, Verginica Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, and Voula Giouli. 2018. Edition 1.1 of the parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, and Voula Giouli. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118.

Agata Savary, Chérifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, and Sara Stymne. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35.

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547.