The Proper Treatment of Verbal Idioms in German Discourse Representation Structure Parsing

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

Abstract

Existing datasets for semantic parsing lack adequate representations of potentially idiomatic expressions (PIEs), i.e., expressions consisting of two or more lexemes that can occur with either a literal or an idiomatic reading. As a result, we cannot test semantic parsers for their ability to correctly distinguish between the two cases, and to assign appropriate meaning representations. We address this situation by combining two semantically annotated resources to obtain a corpus of German sentences containing literal and idiomatic occurrences of PIEs, paired with meaning representations whose concepts and roles reflect the respective literal or idiomatic meaning. Experiments with a state-of-the-art semantic parser show that given appropriate training data, it can learn to predict the idiomatic meanings and improve performance also for literal readings, even though predicting the correct concepts in context remains challenging. We provide additional insights through evaluation on synthetic data.

1 Introduction

Meaning representations such as Minimal Recursion Semantics (Copestake et al., 2005), Abstract Meaning Representations (Banarescu et al., 2013) or Discourse Representation Structures (Kamp and Reyle, 1993) form a link between natural language and the realm of symbolic computation, including ontologies and logical reasoning. They have uses in tasks such as information extraction, dialogue systems, and computer-assisted study of natural language semantics (Sadeddine et al., 2024). Meaning representations have traditionally been constructed from text using rule-based precision grammars or combinations of statistical syntactic parsers and rule-based interpretation systems (Copestake and Flickinger, 2000; Curran et al., 2007). More recently, larger quantities of annotated sentence-meaning pairs have made it possible to perform accurate data-driven text-to-meaning parsing (semantic parsing) and meaning-to-text generation (e.g., Flanigan et al., 2014; van Noord et al., 2020; Wang et al., 2023).

Decomposable verbal idiom: *Don't **spill** the **beans**!*

	x e
¬	**spill.v.05**(e) Agent(e, hearer)
	Theme(e, x) **secret.n.01**(x)

Non-decomposable verbal idiom:
*Are you **pulling** my **leg**?*

e
pull_the_leg_of(e) Agent(e, hearer)
Theme(e, speaker)

Literal occurrence of a verbal potentially idiomatic expression: *They like **playing games** on the PlayStation 2.*

x e f y z
person.n.01(x) like.v.02(e) Experiencer(e, x)
Stimulus(e, f) **play.v.01**(f) Agent(f, x)
Theme(f, y) Instrument(f, z) **game.n.01**(y)
entity.n.01(z) Name(z, "PlayStation 2")

Figure 1: Discourse representation structures for three sentences, containing different occurrences of potentially idiomatic expressions (PIEs). The bolded words are the components of the PIEs, the bolded concepts express their meanings in the respective context.

Datasets that have been constructed using computational grammars typically have a more or less strong built-in assumption that each occurrence of a content word is associated with exactly one occurrence of a *concept* (i.e., of a word sense from an ontology such as WordNet; Fellbaum, 1998). Furthermore, one typically assumes that while lex-

Proceedings of the 16th International Conference on Computational Semantics, pages 156–165

emes can be ambiguous, their senses do not depend on co-occurrence with specific other lexemes.

These assumptions break down in the case of *phrasemes* or *multiword expressions* (MWEs), i.e., combinations of two or more words expressing a single sense (e.g., *pull someone's leg*: `kid.v.01`), or being associated with different but specific senses when occurring together (e.g., *spill the beans*: `talk.v.04` and `secret.n.01`). MWEs occur in a variety of forms (Baldwin and Kim, 2010). In this paper, we focus on *verbal* MWEs, i.e., MWEs whose syntactic head is a verb. In particular, we focus on the subtype of *verbal idiom*. Ramisch et al. (2018) define verbal idioms (VIDs) as MWEs with at least two lexicalized components including the head verb and at least one of its dependents, excluding special cases like light verb constructions, verb-particle constructions, inherently adpositional verbs or inherently reflexive verbs. Following Nunberg et al. (1994), we further distinguish two subtypes of verbal idioms: *decomposable VIDs* such as *spill the beans* where the lexicalized components still have individual meanings even though they are specific to the combination, and *nondecomposable VIDs* such as *pull someone's leg* where all lexical components express a single concept together. Note also that even when two or more lexemes can form a VID together, they can still occur in the same syntactic configuration with a literal, non-idiomatic, compositionally derivable meaning. For example, the phrase *playing games* can occur with an idiomatic but also with a literal meaning. We are therefore dealing with *potentially idiomatic expressions* (PIEs; Haagsma, 2020) with both idiomatic and literal occurrences. It should be noted that PIE occurrences need not be contiguous but exhibit syntactic flexibility as in *the beans were spilled* or *the games that we played*.

In the context of semantic parsing, PIEs present specific challenges: 1) on encountering a PIE, the parser has to decide whether it indeed has the idiomatic meaning in this context, and 2) if so, it must produce the correct meaning representation, meaning one or more concepts that are specific to the idiom, and no additional concepts for additional components of non-decomposable idoms (see Figure 1).

In this paper, we demonstrate that existing semantic parsers for discourse representation structures underperform on sentences containing literal and idiomatic PIEs. We also show a way to remedy this situation. To this end, we combine two semantically annotated resources, the Parallel Meaning Bank (PMB; Abzianidze et al., 2017, 2020) and the dataset of Ehren et al. (2024), to obtain a corpus of German sentences containing literal and idiomatic occurrences of PIEs, annotated with meaning representations that reflect the correct meaning in context (Section 2). We then show that enriching the training data of a DRS parser with such data improves its performance on sentences containing idiomatic occurrences of PIEs, and in some cases its performance overall. Nevertheless, it remains challenging for the parser to reliably distinguish between literal and idiomatic uses, and also to choose the correct concepts for idioms (Section 3). We provide further insights with an evaluation on synthetic data (Section 4). We conclude in Section 5.

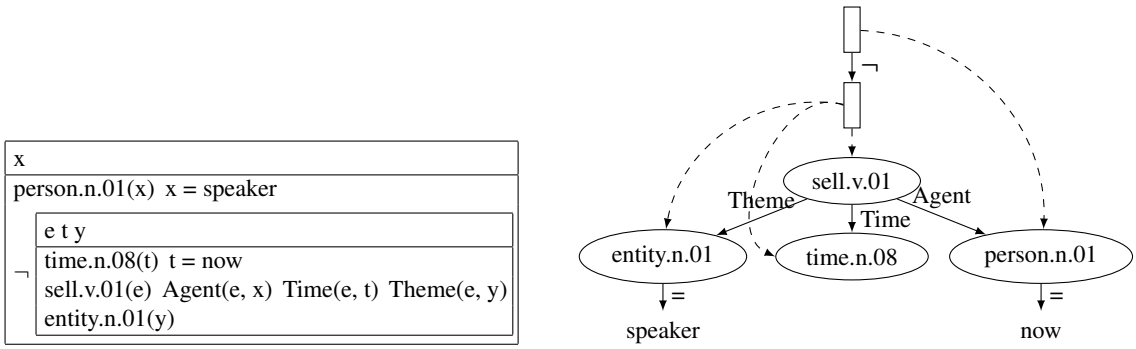
Besides these experimental designs and findings, our contributions include several reusable datasets which will be released upon publication, including an adjudicated version of Ehren et al.'s semantically annotated idiom dataset, an accordingly reannotated version of sentences containing PIEs in the Parallel Meaning Bank, and a synthetic dataset containing the annotated idioms isolated in canonical form, annotated with meaning representations.

2 Data

2.1 The Parallel Meaning Bank

The Parallel Meaning Bank (PMB; Abzianidze et al., 2017, 2020) is a partially parallel corpus of English, German, Italian, and Dutch texts, annotated with discourse representation structures (DRS) following Discourse Representation Theory (Kamp and Reyle, 1993), including word senses, semantic roles, discourse connectives, scope, coreference, etc. The annotations were created by an NLP pipeline and hand-corrected by human annotators. Completely checked documents have the status "gold", partially checked ones, "silver", and unchecked ones, "bronze". Even silver and bronze documents have been shown to be useful for training data-driven DRS parsers (van Noord et al., 2018).

In the PMB, a *document* consists of one or more sentences, paired with one DRS. Traditionally, DRSs are drawn as boxes as shown in Figure 2a. The top part of a box contains the *discourse referents*, which represent events, things, and other entities. The bottom part contains *conditions*, including a *concept condition* for each discourse ref-



(a) (b)

```
person.n.01 EQU speaker             % I am not
NEGATION <1
time.n.08 EQU now
sell.v.01 Agent -2 Time -1 Theme +1 % selling
entity.n.01                         % anything
```

(c)

Figure 2: DRS for the sentence "I am not selling anything" in (a) box notation, (b) graph notation, and (c) sequence notation. The sequence notation additionally shows concept-token alignment information. Adapted from Wang et al. (2023).

erent, saying what type of entity it is, *relation conditions* encoding semantic roles and other relations between entities, *equality conditions* linking referents to discourse constants such as `speaker` or `now`, and *complex conditions* consisting of a logical connective such as negation and one or two embedded boxes. DRSs can also be represented as *discourse representation graphs* (DRGs) as shown in Figure 2b. Here, boxes and discourse referents are represented as nodes. Referent nodes are labeled with their concepts. Box nodes have outgoing edges to the introduced referents and complex conditions, the latter labeled with discourse connectives. Relation conditions are encoded as edges between the referent nodes. Constants are encoded as nodes with incoming edges labeled =. Finally, this graph structure can be linearized as a sequence of tokens (Bos, 2023) as shown in Figure 2c. Here, nodes are encoded by their label, and edges are encoded following their source node by their label followed by a *pointer* indicating the relative position of the sink node. Concept nodes are aligned to the natural-language tokens that evoke them, also shown in Figure 2c.

2.2 German Verbal PIE Data

Ehren et al. (2024) argue that idioms are underrepresented in the gold part of the PMB, and they released a dataset of 6 187 sentences from the PMB's German part that contain verbal potentially idiomatic expressions (PIEs), annotated for whether the instance is idiomatic, and if so, for its sense, assigned roles, and, in the case of decomposable idioms, internal senses and roles. The following are examples of the annotations in this dataset. (1) is a PIE annotated as literal, (2) is a decomposable idiom annotated with senses and internal and external roles, and (3) is a non-decomposable idiom annotated with a sense and external roles.

(1) Tom lag bewusstlos auf dem Operationstisch
Tom lay unconscious on the operating table
PIE: auf dem Tisch liegen ("to be available; shown, offered")
Reading: literal

(2) Ich_[Theme] sitze_[be.v.01] im_[] selben_[]
I sit in the same
Boot_[Attribute]_[situation.n.02] wie du
boat as you
PIE: im selben Boot sitzen ("to be in the same boat")
Reading: idiomatic

(3) Tom_[Theme] kämpft, um über_[] die_[]
Tom fights, to over the
Runden_[] zu kommen_[survive.v.03]
rounds to come
PIE: über die Runden kommen ("to support oneself")
Reading: idiomatic

2.3 Adjudication

We extracted from Ehren et al.'s dataset the 2 204 sentences with a PIE annotated by at least one annotator as idiomatic, and thus with a semantic an-

Table 1: Four types of adjudication decisions, with examples. Struck out lines represent annotations that were discarded in favor of the other annotation. Struck out (underlined) spans represent parts of selected annotations that were removed (added) by the adjudicator. Annotations without struck out or underlined spans represent annotations that were approved unchanged in adjudication, shown here for comparison.

Consistent concepts and roles per sentence and per PIE type:
Das abgestürzte Flugzeug_[Patient] ging_[go_up.v.06] in_[] Flammen_[] auf_[]
~~Das abgestürzte Flugzeug_[Patient] ging_[go_up.v.01] in_[] Flammen_[] auf_[]~~
"The crashed plane went up in flames"
Die ,, Hindenburg_[Patient] " ging_[go_up.v.~~01~~06] plötzlich in_[] Flammen_[] auf_[]
"The 'Hindenburg' suddenly went up in flames"
Marking negation as part vs. not part of the idiom:
Käse_[Theme] und andere Milchprodukte_[Theme] bekommen_[agree.v.06] mir_[Attribute] nicht~~_[]~~
"Cheese and other dairy products do not agree with me"
Ich_[Experiencer] kann_[] sie_[Stimulus] nicht_[] ausstehen_[loathe.v.01]
"I cannot stand her/them"
Treatment of auxiliary (including modal) verbs as not head of clause:
Ich_[Experiencer] kann_[~~interest.v.01~~] mit_[] diesem Test_[Stimulus] nichts anfangen_[interest.v.01]
"I am not interested in this text" (lit. "I cannot begin anything with this text")
Treatment of adjective copula and auxiliary sein *as not part of idiom:*
Du_[Agent] musst auf alles_[Beneficiary] gefasst_[prepare_for.v.01] sein~~_[prepare_for.v.01]~~
"You have to be prepared for anything"

notation. This results in a total of 4 600 annotations (the number of annotations per sentence is slightly above 2), across 957 different PIEs. The first author went through the dataset manually and resolved divergent annotations according to Ehren et al.'s annotation manual. We also made sure that the same PIE was annotated consistently across occurrences, using the same WordNet sense and the same VerbNet roles for corresponding arguments. Because it was a frequent source of disagreement and affects automatic combination with the PMB data through word-concept alignment information (see next section), we made special adjudication passes to ensure conformance with the annotation guidelines wrt. the treatment of copulas, auxiliary verbs, and negation words like *nicht* "not". Examples are shown in Table 1.

2.4 Combining the Annotations with the PMB Data

The resulting unique annotations were automatically combined with the PMB 5.1.0, matching the annotations by sentence, using the alignment between tokens and concepts provided with the PMB. Examples are shown in Figure 3. For tokens annotated with a sense, the corresponding node was relabeled with that sense. For tokens annotated with a role, the incoming edge was relabeled with that role. For tokens annotated with an empty pair of brackets, the corresponding node and its incoming and outgoing edges were removed. As a result, we obtained 2 186 reannotated sentence-DRS pairs with idiomatic readings of PIEs.

2.5 Datasets

We prepared the following datasets for training and evaluating DRS parsers:

$\mathcal{I}$: the 2 186 automatically reannotated sentences containing idiomatic PIE instances, as described in Section 2.4.

$\mathcal{L}$: the 455 sentences marked by at least one annotator in Ehren et al.'s dataset as containing a literal reading of a PIE.

Note that both datasets may contain errors, as most of them are "bronze" or "silver", and reannotation only fixes the annotation of the PIE instance.

3 Targeted Training on PIE Instances

We assess the performance of a seq2seq parser on PIEs, comparing four different training conditions: training on the unmodified PMB data (baseline), adding available PIE instances into the training data (enhanced), adding a balanced mix of literal and idiomatic PIE instances into the training data (bal-

Er_[Experiencer] **schwimmt**_[buck.v.02] **gegen**_[] **den**_[] **Strom**_[Stimulus]_[trend.n.01]
"He bucks the trend" (lit. "He swims against the tide")

```
male.n.02                                                          % Er
~~schwimmt.v.01~~buck.v.02 ~~Agent~~Experiencer -1 Time +1 ~~Location~~Stimulus +2 % schwimmt gegen den
time.n.08              EQU now                                     %
~~tide.n.01~~                                                      % Strom.
```

Sie **steckt**_[despair.v.01]_[Experiencer] **den**_[] **Kopf**_[] **in**_[] **den**_[] **Sand**_[]
"She despairs" (lit. "She puts the head into the sand")

```
female.n.02                                                          % Sie
~~steckt.v.01~~despair.v.01 ~~Agent~~Experiencer -1 Time +1 ~~Theme +3 Location +4~~ % steckt
time.n.08                EQU now                                     %
~~female.n.01~~                                                     % den
~~head.n.01~~                ~~Participant -1~~                     % Kopf in den
~~sand.n.01~~                                                       % Sand.
```

Figure 3: Automatic combination of semantic idiom annotations with the PMB data via concept-token alignment. The meaning representations are discourse representation structures in sequence notation (cf. Section 2.1). Struck out (underlined) spans represent parts of the meaning representation that were removed (added) compared to the original PMB data. Note that some of the replaced senses, such as `steckt.v.01` or `schwimmt.v.01`, would be incorrect even in a literal reading, since they are not WordNet senses but artifacts of the bootstrapping process for the German DRS data.

anced), and weighing PIE instances more strongly than other training instances (balanced×4).

3.1 Model and Evaluation Metric

We use the seq2seq parser of Wang et al. (2023) as implemented by Zhang et al. (2024), with the pre-trained ByT5 language model (Xue et al., 2022). We further pre-train on PMB gold, silver, and bronze data for 3 epochs, then fine-tune on gold data (plus PIE data) for 10 epochs. We also follow these papers by using Smatch (Cai and Knight, 2013), adapted to DRS, as the evaluation metric.

3.2 Data Splits

We split $\mathcal{I}$ and $\mathcal{L}$ randomly into five equal parts and use a different part in each run for testing, reporting results as the median of five runs. We call this part $\mathcal{I}_{test}$ ($\mathcal{L}_{test}$) and the remainder $\mathcal{I}_{train}$ ($\mathcal{L}_{train}$).

For pre-training the **baseline** model, we use the PMB 5.1.0 German bronze, silver, and gold training portions, but with sentences in $\mathcal{I}_{test}$ and $\mathcal{L}_{test}$ removed. For fine-tuning the baseline model, we use the PMB 5.1.0 German gold training portion, which does not overlap with $\mathcal{I}_{test}$ or $\mathcal{L}_{test}$.

For pre-training the **enhanced** model, we use the same pre-training data as above except that sentences in $\mathcal{I}_{train}$ have their annotations replaced by the modified ones. For fine-tuning, we additionally add $\mathcal{I}_{train}$ and $\mathcal{L}_{train}$ to the fine-tuning data.

For fine-tuning the **balanced** model, we do the same but add $\mathcal{L}_{train}$ five times so that the count of idiomatic and literal training instances is approximately equal.

For fine-tuning the **balanced×4** model, we again multiply all the idiomatic and literal training instances by 4, thus weighting idiomatic PIE instances 4 times and literal ones 20 times as heavily as the standard training data. The value 4 was found in preliminary experimentation to improve accuracy for parsing idioms compared to 2 and to be on par with 8.

We then evaluate 1) on the PMB 5.1.0 standard gold test and dev sets; 2) on $\mathcal{I}_{test}$ and various subsets, *viz.* sentences with idioms seen in training, sentences with idioms not seen in training, and sentences sampled from shortest to longest to have the same mean length (in characters) as the standard test set; 3) on $\mathcal{L}_{test}$. The output DRSs are evaluated against the corresponding DRSs in the test sets using the Smatch metric.

3.3 Results

Results are shown in Table 2. We see that, compared to the baseline model, the enhanced model improves scores significantly even on the standard test and dev sets. This could be due to additional data helping even when it is not gold and does not directly address phenomena found in the test set. We see that compared to the standard dev and test sets, both models perform much worse on sen-

	baseline	enhanced	balanced	balanced×4
standard test	.815	.828*	.824*	.810
standard dev	.827	.835*	.832*	.819
idiomatic	.520	.572*	.567*	.606*
idiomatic seen	.530	.580*	.568*	.604*
idiomatic unseen	.518	.550*	.522*	.555*
idiomatic short	.614	.679*	.670*	.739*
literal	.650	.642	.658	.613

Table 2: Performance comparison of different models on the PMB 5.1.0 official test/dev data, on sentences with idioms, on sentences with seen and unseen idioms, on short sentences with idioms, and on sentences with literal PIE occurrences. Scores are macro-averages (mean) over the Smatch scores of the sentences; averaged (median) over five runs, each with a different test fold of idiom and literal sentences. * indicates statistically significant improvement over the baseline ($p \leq 0.05$) according to a permutation test (Dror et al., 2018).

tences containing PIEs and especially idiomatic occurrences, even when we downsample the latter to only contain idioms seen in training, or to contain only short sentences. This is partly due to idioms being challenging to handle, but also points to the test sentences' bronze/silver status, which we address in Section 4. But we also see that the enhanced model, additionally trained on reannotated idiomatic instances as well as literal PIE instances, performs several percentage points better on the idiomatic instances. As may be expected, the improvement on idioms that have not been seen in training is comparatively small. Performance on literal PIE instances is worse than the baseline model, suggesting that the enhanced model is biased towards idiomatic readings. This is not very surprising given the much larger size of $\mathcal{I}$ compared to $\mathcal{L}$. By contrast, the balanced model avoids degradation on literal instances and in fact improves accuracy (though not significantly) while also still improving over the baseline significantly on all other test sets, albeit slightly less than the enhanced model. It is worth noting that targeted training on PIEs can thus maintain performance on literal instances although literal interpretations are the default case in the standard training data. Finally, the balanced×4 model achieves the best accuracy on idiomatic readings, but performs worse than the baseline on literal readings, and also degrades on the standard test sets.

4 Evaluation on Synthetic Data

The data in $\mathcal{I}$ and $\mathcal{L}$ is based on bronze and silver documents in the PMB and thus contains errors, even though concepts and roles representing the meanings of idioms have been automatically fixed in $\mathcal{I}$. The above experiments thus give a somewhat misleading view of the parsers' performance. To better understand the performance on idioms without any unrelated error sources, we perform an evaluation on a synthetic 'test set' of minimal sentences containing idioms, paired with DRSs that are correct by construction, assuming the idiomatic reading.

4.1 Construction of the Synthetic Dataset

We went through the adjudicated idiom sentences and reduced each idiom to a natural-language canonical form, similar to Odijk and Kroon (2024). In our case, canonical forms are main clauses, manually chunked, and decorated with senses and roles. DRSs can be automatically generated from the canonical forms by mapping placeholder words such as *etwas* "something", *irgendwie* "somehow", *irgendwohin* "somewhere", or *jemand* "somebody" to arbitrary fillers; we simply use general concepts representing the meaning of the placeholders, *viz.* `entity.n.01`, `manner.n.01`, `location.n.01`, and `person.n.01`, respectively. For example:

(4) [Jemand]$_{\text{Patient}}$ kommt$_{\text{die.v.01}}$ [ums Leben]
somebody comes around the life
"Somebody dies"
`person.n.01 die.v.01 Patient -1 Time +1 time.n.08 EQU now`

(5) [Etwas]$_{\text{Stimulus}}$ geht$_{\text{annoy.v.01}}$
something goes
[jemandem]$_{\text{Experiencer}}$ [auf die Nerven]
somebody on the nerves
"Something annoys somebody"
`entity.n.01 annoy.v.01 Stimulus -1 Time +1 Experiencer +2 time.n.08 EQU now person.n.01`

	baseline	enhanced
dev	.695	.767*
test	.689	.765*
dev decomposable	.706	.722*
dev non-decomposable	.678	.752*
test decomposable	.720	.736*
test non-decomposable	.692	.757*

Table 3: Results on synthetic test data. * indicates statistically significant improvement over the baseline ($p \leq 0.05$) according to a permutation test (Dror et al., 2018).

(6) [Etwas]$_{\text{Patient}}$ geht$_{\text{come.v.04}}$ [in Erfüllung]$_{\text{true.a.01}}$
something goes into fulfillment
"Something comes true"
`entity.n.01 come.v.04 Patient -1 Time +1 Result +2 time.n.08 EQU now true.a.01`

We obtained 890 sentence-DRS pairs in this way and split them randomly into an even-sized development set and test set.

4.2 Experiments

We use the baseline model and the enhanced model from Section 3.2. Now, instead of 5-fold cross validation, we add all of $\mathcal{I}$ and $\mathcal{L}$ to the fine-tuning data and evaluate on the synthetic data.

4.3 Results

Results are shown in Table 3. With the syntactic structure and the interpretation of arguments trivial in the synthetic data, scores now almost exclusively reflect the model's ability to map the idiom to the correct sense(s) and roles. Again, the enhanced model does significantly better than the baseline model.

We also see that the baseline model does better on decomposable than on non-decomposable idioms. This makes sense, as the correct interpretations of decomposable idioms are structurally closer to literal readings, with two senses rather than one. In the enhanced model, this is reversed: it does better on non-decomposable idioms. This shows that the model has learned to predict the non-canonical structure of non-decomposable idioms. The better scores are probably also due to one sense being statistically easier to predict correctly than two, and to the stronger representation of non-decomposable idioms in our training data.

We show here some examples that the baseline model parses wrongly and the enhanced model parses correctly:

(7) Jemand macht sich über jemanden
Somebody makes themself about somebody
lustig
funny
"Somebody mocks somebody"
Baseline: `person.n.01 make.v.01 Agent -1 Time +1 Product +2 Theme +3 time.n.08 EQU now male.n.02 person.n.01 funny.a.01 AttributeOf -1`
Enhanced: `person.n.01 mock.v.01 Agent -1 Time +1 Theme +2 time.n.08 EQU now person.n.01`

(8) Jemand setzt jemanden über etwas in
Somebody sets somebody about something in
Kenntnis
knowledge
"Somebody informs somebody"
Baseline: `person.n.01 put.v.01 Agent -1 Time +1 Theme +2 Theme +3 time.n.08 EQU now person.n.01 entity.n.01`
Enhanced: `person.n.01 inform.v.01 Agent -1 Time +1 Recipient +2 Topic +3 time.n.08 EQU now person.n.01 entity.n.01`

(9) Etwas geht vor sich
Something goes before itself
"Something happens"
Baseline: `entity.n.01 go.v.01 Theme -1 Time +1 time.n.08 EQU now`
Enhanced: `entity.n.01 happen.v.01 Theme -1 Time +1 time.n.08 EQU now`

(10) Jemand weiß etwas zu schätzen
Somebody knows something to value
"Somebody appreciates something"
Baseline: `person.n.01 know.v.01 Experiencer -1 Time +1 Stimulus +2 time.n.08 EQU now entity.n.01 appreciate.v.01 Agent -4 Theme -1`
Enhanced: `person.n.01 appreciate.v.01 Experiencer -1 Time +1 Stimulus +2 time.n.08 EQU now entity.n.01`

As for the effect of the frequency of an idiom in the training data, a scatterplot (Figure 4) shows that although high scores on the synthetic data are already achieved with as little as one training example, reliably decent scores are only seen around 10 or more training examples.

5 Conclusions, Limitations, and Future Work

Potentially idiomatic expressions (PIEs) present a special challenge in semantic parsing due to their

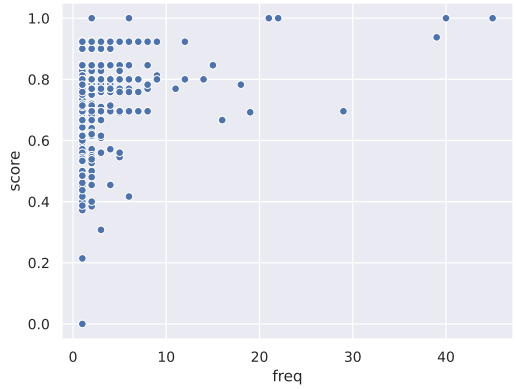


Figure 4: Scatterplot of idiom frequency in the training data against smatch score in the synthetic development data.

idiomatic meanings in some contexts, often with a single concept expressed by two or more content words. For an existing state-of-the-art system for parsing German text to discourse representation structures, we have shown that it struggles with sentences containing verbal PIEs more than with the average sentence. We have also shown that the gap can be partially closed without changing the parser's sequence-to-sequence architecture, simply by injecting sentences with PIEs into the training data, where sentences with idiomatic readings have been reannotated to reflect these.

Our contributions also include an adjudicated version of Ehren et al.'s German semantically annotated verbal PIE dataset, a correspondingly reannotated version of 2 186 German sentences in the Parallel Meaning Bank which we intend to submit for inclusion into the next release of the PMB, and a synthetic dataset of 890 idioms in isolated canonical form, with corresponding meaning representations.

There are two main limitations: the first concerns evaluation. Because we had only partially corrected test data at our disposal, we still only have an approximate picture of how accurately models handle PIEs. We partially addressed this by evaluating on synthetic data, but future work should aim to get an accurate picture on idiom semantic parsing accuracy on real data.

The second limitation applies to semantic parsing in general: meaning representations are expensive to annotate, thus the training data is limited in quantity and quality, with model training having to rely on partially corrected data. Although we achieved improved scores on sentences containing idioms, in many cases the models still struggle to pick the correct sense. As performance grows on idiomatic instances, it goes down on literal ones, suggesting that models seem to prefer one or the other and struggle with distinguishing between literal and idiomatic occurrences in context.

Future work should build on our synthetic dataset by using it not just for testing but also for training, automatically generating from the canonical forms sentences more varied in clause type, embedding complexity, fillers for placeholders, negation, modality, tense, etc. In addition, it may be worth making the decision between idiomatic and literal readings explicit and delegating it to a specialized model.

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. Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf.

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.

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

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.

Johan Bos. 2023. The sequence notation: Catching complex meanings in simple graphs. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 195–208, Nancy, France. Association for Computational Linguistics.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Ann Copestake and Dan Flickinger. 2000. An open source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Ann A. Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3:281–332.

James Curran, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale NLP with C&C and boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, Prague, Czech Republic. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Rafael Ehren, Kilian Evang, and Laura Kallmeyer. 2024. To leave no stone unturned: Annotating verbal idioms in the Parallel Meaning Bank. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 115–124, Torino, Italia. ELRA and ICCL.

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

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. 2014. A discriminative graph-based parser for the Abstract Meaning Representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Baltimore, Maryland. Association for Computational Linguistics.

Hessel Haagsma. 2020. *A Bigger Frish 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. In *Studies in Linguistics and Philosophy*.

Rik van Noord, Lasha Abzianidze, Antonio Toral, and Johan Bos. 2018. Exploring neural methods for parsing discourse representation structures. *Transactions of the Association for Computational Linguistics*, 6:619–633.

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.

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

Jan Odijk and Martin Kroon. 2024. A canonical form for flexible multiword expressions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 91–101, Torino, Italia. ELRA and ICCL.

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 *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zacchary Sadeddine, Juri Opitz, and Fabian Suchanek. 2024. A survey of meaning representations – from theory to practical utility. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2877–2892, Mexico City, Mexico. Association for Computational Linguistics.

Chunliu Wang, Huiyuan Lai, Malvina Nissim, and Johan Bos. 2023. Pre-trained language-meaning models for multilingual parsing and generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5586–5600, Toronto, Canada. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Xiao Zhang, Chunliu Wang, Rik van Noord, and Johan Bos. 2024. Gaining more insight into neural semantic parsing with challenging benchmarks. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024*, pages 162–175, Torino, Italia. ELRA and ICCL.