

Modelling Idiomatic Expressions in Abstract Meaning Representation

Venera Gareeva¹, Johannes Heinecke²

¹Université de Strasbourg, 67000 Strasbourg, France

²Orange Research, 22300 Lannion, France

venera.gareeva@etu.unistra.fr, johannes.heinecke@orange.com

Abstract

Idiomatic expressions, a subclass of multiword expressions (MWE), pose persistent challenges for semantic parsing, as their meanings often diverge from the compositional semantics of their constituent words and depend strongly on contextual cues. While Abstract Meaning Representation (AMR) parsers aim to capture sentence-level semantics in a structured graph form, existing datasets provide limited coverage of idiomatic language, constraining their ability to model such expressions accurately. To address this gap, we extended a subset of the MAGPIE dataset by constructing a corpus of potentially idiomatic expressions (PIE) annotated with their corresponding AMR graphs. The dataset includes both naturally occurring and synthetically generated sentences, covering idioms in literal and idiomatic contexts. We fine-tune a state-of-the-art AMR parser on this dataset and evaluate its capacity to generate context-sensitive graphs that correctly reflect idiomatic versus literal interpretations. Our results show that standard parsers often capture only literal meanings of such expressions, while fine-tuning on our dataset improves alignment with the intended interpretations.

Keywords: idiomatic expressions, Abstract Meaning Representation, idiom corpus, parsing

1. Introduction

Several formalisms have been proposed to capture semantic structures. Among them, Abstract Meaning Representation (AMR) is a formalism whose aim is to provide an abstract, standardized representation that is independent of syntax and structured as directed graphs (Banarescu et al., 2013). While AMR is designed to represent the meaning of sentences, a challenge arises in how to capture idiomatic expressions, whose interpretations cannot be derived literally, for example, *kick the bucket* (“to die”) or *pull someone’s leg* (“deceive someone jokingly”).

Within the AMR framework, different types of MWEs (combinations of words whose meaning cannot always be directly inferred from the meanings of their individual components (Ramisch, 2023)) are handled through the semantic role structure provided by PropBank (Kingsbury and Palmer, 2002), which supplies verb-specific frames that guide annotation. As a result, constructions such as verb–particle combinations and light verb constructions are generally well represented, since their meanings align with established predicate–argument structures. In contrast, the representation of idiomatic expressions in PropBank, and consequently in AMR, remains limited.

This limitation is particularly critical because understanding idiomatic expressions in context is essential for capturing the intended meaning of a text. Unlike other types of MWEs (like colloca-

tions or light verb constructions), idioms cannot be resolved solely through lexical or syntactic cues; their interpretation depends on recognizing pragmatic and contextual information (Beck and Weber, 2020). Human processing of idioms relies heavily on interpreting the surrounding context, which determines whether an figurative or literal reading is activated. For example, “*in hot water*” can denote a physical situation (e.g., cooking vegetables) or a figurative state of trouble, depending on context.

In this paper, we investigate how idiomatic expressions are represented in AMR corpora and whether text-to-AMR parsers distinguish figurative from literal readings. We focus on verbal idioms and introduce a dedicated English corpus to fine-tune state-of-the-art AMR parsers for evaluating context-sensitive interpretations. While adaptations of PropBank exist for other languages, such as Universal Propositions (Akbik et al., 2015)¹, their limited coverage (e.g., 533 verbs for French), together with AMR’s reliance on PropBank frameworks, motivates English as a natural starting point before multilingual extension.

1.1. Abstract Meaning Representation

In the AMR formalism, sentences are represented as graphs whose nodes correspond to concepts and edges to semantic relations. AMR graphs can be encoded in several formats, including PENMAN notation (Kasper, 1989), a textual representation

¹<https://github.com/UniversalPropositions/UP-1.0>

that uses nested parentheses to hierarchically organize concepts and their semantic roles, making graph structure explicit.

AMR is based largely on semantic frames derived from the PropBank project (Kingsbury and Palmer, 2002; Palmer et al., 2005), that provides a detailed inventory of verb-specific roles and their typical arguments. PropBank thus allows for a clear identification of “who does what to whom” in a sentence.

To train models that predict AMR graphs from text, several datasets are available, the largest being AMR 3.0 (Knight et al., 2020), distributed by the Linguistic Data Consortium (LDC2020T02). The corpus contains nearly 60,000 sentences, divided into 55,635 training, 1,722 validation, and 1,898 test instances.

AMR graphs are commonly evaluated using the Smatch metric (Cai and Knight, 2013), which measures overlap between predicted and gold graphs by comparing sets of graph triples. Because node variables are arbitrary, Smatch computes an optimal alignment that maximizes F1, based on precision and recall over matched triples. A more recent evaluation library is Smatch++ which provides a more accurate mapping of graphs (Opitz, 2023) or takes into account semantic similarity (Opitz et al., 2020).

2. Related Work

AMR is a powerful formalism for capturing propositional content, but its handling of figurative and idiomatic language is limited. Mansouri (2025) claims that AMR struggles with idioms and metaphors, often representing them literally or abstracting them away without capturing their figurative meaning. Thus, an idiomatic expression such as “kick the bucket” may be represented as an action involving a bucket rather than the intended metaphorical sense of death.

That said, it would be overly reductive to claim that AMR is completely incapable of handling idioms. The AMR 3.0 corpus (cf. section 1.1 above) includes some annotated idiomatic expressions. While coverage is far from exhaustive, these examples show that the framework can, at least in part, encode certain figurative meanings rather than collapsing them into literal interpretations. While there has been substantial work on multiword expressions (Baldwin and Kim, 2010; Zeng and Bhat, 2021; Ramisch, 2023), relatively little research has explored the use of semantic formalisms for idiomatic expressions. A closely related line of work by Evang et al. (2025) in Discourse Representation Structure (DRS) parsing shows that, with suitable training data, semantic parsers can learn to distinguish between literal and idiomatic read-

ings, although predicting idiomatic meanings in context remains challenging. Otherwise, several non-semantic approaches have been proposed to identify idiomatic expressions (Zeng and Bhat, 2021; He et al., 2024). In chapter 3 we will notably use the work of Haagsma et al. (2020).

3. Multiword expressions

Multiword expressions (MWE) represent a complex linguistic phenomenon and pose challenges for natural language processing. Following the definition of Baldwin and Kim (2010), MWEs are defined as lexical items composed of multiple lexemes that exhibit lexical, syntactic, semantic, pragmatic, and/or statistical idiomaticity. Unlike named entities, which can be identified through formal indicators such as capitalization or trigger words, MWEs lack systematic markers, making their detection difficult (Savary et al., 2019). Their variability further complicates processing: verbal MWE, for instance, may appear in diverse syntactic forms through passivization, modification, or nominalization. However, in the context of AMR parsing, syntactic variability is less central, as AMR abstracts away from surface syntax and word order.

Beyond formal variability, the main difficulties posed by MWEs are semantic ambiguity and non-compositionality: they may be interpreted literally or figuratively depending on context, and their meanings are often not derivable from their parts.

Idiomatic expressions form a subclass of MWEs, sharing these properties. Their interpretation, however, relies heavily on shared cultural knowledge (Chung, 2024), making them opaque to non-native speakers and difficult to translate across languages. In line with (Nunberg et al., 1994), a distinction can be drawn between idiomatically combining expressions, whose meanings are partly distributed across their components, and idiomatic phrases, whose meanings are fixed and not compositionally derived from their parts. In our work, the set of verbal expressions under study includes examples of both types.

The MAGPIE corpus (Haagsma et al., 2020) is the largest annotated resource for potentially idiomatic expressions (PIEs) in English, containing over 56,000 instances across nearly 1,800 expression types. PIEs are defined as fixed or semi-fixed expressions that may be interpreted literally or figuratively depending on context.

For the purposes of this study, a subset of the MAGPIE corpus was manually selected, retaining only frequent idioms with both literal and idiomatic attestations. This resulted in a working corpus of 3,582 sentences covering 41 mostly verbal idioms. This focus reflects their frequency and suitability for studying contextual meaning variation,

and serves as a starting point for future extensions beyond verbal idioms. The selected subset was annotated according to the process described in Section 4.

4. Annotation

As a preliminary step, the AMR 3.0 corpus (Knight et al., 2020) was examined via the metAMoR-phosED tool (Heinecke, 2023) to identify idiomatic expressions and assess how they were represented by the expert annotators. This exploration revealed that idioms were relatively rare and inconsistently treated: some were covered by PropBank frames (Palmer et al., 2005) (e.g., *steer clear* encoded as *steer-clear-02*), others were paraphrased into equivalent concepts (e.g., *goes without saying* encoded as *obvious-01*), while many were either absent or interpreted literally. Based on these observations, a consistent annotation strategy was established for our corpus. Each idiomatic expression identified in a sentence was handled according to the following principles:

1. If a corresponding semantic frame was already defined in PropBank, this frame was used directly to encode the idiomatic expression. This ensured compatibility with AMR standards. For instance, *follow suit* can be represented by *follow-suit-06*, which directly captures its figurative meaning without additional reformulation. A selection of such idiomatic frames for verbal expressions is presented in Table 1.
2. When no such frame is available, the idiomatic expression was paraphrased (when used in its figurative meaning) to capture its implicit meaning, and the AMR annotation was based on this paraphrase (e.g., *tie the knot* as *marry-01* or *bear fruit* as *produce-01*).

Subgraphs corresponding to idiomatic expressions in their figurative sense were systematically replaced using a rule-based script. Each idiom was associated with transformation rules that specify the patterns to delete and the semantic structures to insert (e.g., *tie the knot* → *marry-01*).

To illustrate, Fig. 1 and Fig. 2 show the expression *tie the knot* in its two different readings. Thus, in the figurative example, the subgraph corresponding to the idiom is replaced by the single semantic unit *marry-01*, whereas in the literal usage, the compositional structure is preserved, with *tie-01* taking *knot* as its (:ARG1) argument.

From the MAGPIE corpus, we initially gathered 3,352 sentences for the 41 selected expressions. During data preparation, literal usages of PIEs were underrepresented in the original MAGPIE

PropBank	Arguments
follow-suit-06	ARG0: imitator ARG1: thing imitated ARG2: action of following suit
change-hands-06	ARG1: thing changing hands ARG2: giver ARG3: getter
steer-clear-02	ARG0: avoider ARG1: avoided

Table 1: Semantic frames for selected idiomatic expressions in PropBank

```
(t / together
 :domain (a / and
 :op1 (p / person
 :name (n / name
 :op1 "Fiona"))
 :op2 (p2 / person
 :name (n2 / name
 :op1 "Paul")))
 :duration (t2 / temporal-quantity
 :quant 6
 :unit (y / year))
 :time (b / before
 :op1 (d / decide-01
 :ARG0 a
 :ARG1 (m / marry-01
 :ARG0 a))))
```

Figure 1: AMR graph for the sentence: *Fiona & Paul had been together for six years before deciding to tie the knot.* (the expression *tie the knot* is used in its figurative meaning)

```
(s / show-01
 :ARG0 (p / person
 :ARG0-of (s2 / sail-01))
 :ARG1 (t / thing
 :manner-of (t2 / tie-01
 :ARG0 p
 :ARG1 (k / knot)
 :ARG1-of (c / correct-02)))
 :ARG2 (w / we))
```

Figure 2: AMR graph for the sentence *The sailor showed us how to tie the knot correctly* (the expression *tie the knot* is used in its literal meaning)

corpus, so the dataset was augmented with 250 GPT-4.1-nano-generated sentences emphasizing literal meanings using a one-shot prompting strategy. The final dataset contains 3,582 sentences (2,886 train, 360 dev, 336 test). We do not report inter-annotation agreement as the corpus was annotated by a single annotator.

5. Experiments

To evaluate whether text-to-AMR parsers are capable of distinguishing between the literal and figurative interpretations of expressions, pretrained *seq2seq* models were fine-tuned following the annotation process. By exposing the model to idiom-annotated data, we aim to assess its capacity to become idiom-aware, that is, to capture both figurative and literal meanings of PIE expressions.

However, *seq2seq* models cannot directly handle graph structures, which makes data preprocessing essential. Thus, data preprocessing was necessary to adapt AMR graphs to *seq2seq* models, which require linear sequences as input and output. To this end, AMR graphs originally represented in PENMAN were serialized into a simpler format while preserving semantic structure.

For the purpose of fine-tuning, we used the Flan-T5 model (Chung et al., 2022), a variant of T5 (Text-to-Text Transfer Transformer) pretrained under the instruction tuning paradigm, which has demonstrated strong performance across a wide range of NLP tasks and shown to achieve SoTA results in AMR parsing (Lee et al., 2023). We chose the Flan-T5 model since it proved to be the most reliable model of the T5 family (T5, Flan-T5 and the multilingual MT5). All these models come in different sizes (small, base, large, xl, xxl). Even though Flan-T5 xl and xxl gave slightly better results, we kept Flan-T5 base (250M parameters and a vector length of 768) since it needed less computational power (in terms of memory of the GPU device and compute time).

We also tried to use more recent LLMs like Qwen 2.5 (0.5B, 1.5B, 3B and 7B parameters) or Gemma 2 (1B and 2B), but again these models trained on the AMR 3.0 dataset performed slightly less well than Flan-T5 base on AMR 3.0.

The first experimental step was to train the Flan-T5 model on AMR 3.0, producing a baseline system capable of generating AMR structures from natural language. This baseline was then further fine-tuned on our idiom-specific corpus. The fine-tuning was implemented using the Huggingface’s Seq2Seq Trainer framework. Two experimental setups were then conducted. In the first, the FlanT5 model was fine-tuned sequentially—first on the AMR 3.0 corpus and then on the idiom-specific AMR-annotated corpus. In the second, both corpora were concatenated into a single dataset to enable joint fine-tuning. Overall, no major differences in performance were observed between the two approaches.

5.1. Results

To assess the performance of our fine-tuned models, we first evaluate the models trained on the

AMR 3.0 training set and tested on the AMR 3.0 test set. Table 2 summarizes the results of AMR parsing experiments conducted with different learning rates and beam search configurations across various sizes of the Flan-T5 model (base and large). We choose the base model for future experiments since it performs nearly as well as the large model while remaining more computationally efficient and cost-effective.

model	init. LR	epochs	beam search	Smatch score
base	0.00005	15	No	75.12
base	0.00005	15	Yes	75.15
base	0.0001	10	No	81.72
base	0.0001	10	Yes	82.12
large	0.0001	5	No	82.29

Table 2: AMR parsing experiments with Flan-T5 models using different learning rates and beam search configurations

The next step was to fine-tune the models on our idiom-annotated dataset. Test showed a Smatch score of 84.15%.

Subsequently, we compared Smatch scores for each expression against its distribution of literal and figurative occurrences, revealing a clear trend: expressions with balanced examples in both senses generally achieve higher scores, indicating that the model benefits from diverse, evenly distributed training data.

To illustrate this observation, Table 3 provides an overview of selected expressions, showing both the distribution of examples by sense (literal/figurative) and the resulting Smatch scores.

idiom. expr.	total	fig. lit.		Smatch	
		usage	usage	fig.	lit.
go without saying	99	95	4	85	60
bear fruit	103	99	4	90	50
behind bars	98	27	71	82	91
tie the knot	42	18	24	91	93
spill the beans	31	30	1	87	0
ring a bell	122	16	106	100	90
cold feet	14	6	8	40	75
hald water	50	19	31	69	87

Table 3: Individual Smatch Scores by literal and figurative sense for a selection of idiomatic expressions

The corpus was then adjusted, as described in the annotation section (mostly to address cases where it lacked examples of expressions in their literal usage). A new fine-tuning was subsequently carried out on this rebalanced corpus. The performances were re-evaluated with Smatch score of

85.73%, which is an increase of nearly 1.6 points from 84.15% before the adjustments.

Comparing these individual scores to those obtained before the dataset adjustment (cf. Table 3) reveals a clear pattern: the parsing of literal usages improved substantially in most cases, particularly for expressions that initially had very few literal examples (e.g., the expressions *go without saying* and *bear fruit*). For instance, the Smatch score for literal usages of *go without saying* increased from 60% to 98%, while the score for the figurative sense remained nearly stable. Thus, bold figures in the table highlight significant improvements in literal usage parsing following dataset rebalancing, showing that the model learned to better capture literal meanings without significantly affecting performance on figurative usages. Therefore, the injection of idiomatic data into the training pipeline already yielded some improvements across all cases, and even where gains are more modest, a positive tendency is observed, suggesting that enriching the parser with dedicated idiomatic expressions is a promising direction for future development.

idiom. expr.	total	fig. usage	lit.	Smatch	
				fig.	lit.
go without saying	159	95	64	82	98
bear fruit	153	99	54	92	91
tie the knot	65	18	47	92	93
spill the beans	52	30	22	67	100
ring a bell	122	16	106	100	90
miss the boat	36	14	22	91	89
off the hook	42	31	11	86	80
green light	109	63	46	84	85
cold feet	38	15	23	100	70

Table 4: Individual Smatch score for some idiomatic expressions after adjusting the distribution in the dataset. Bold figures indicate improved parsing of literal usages, while the performance on figurative usages remains nearly unaffected.

6. Conclusion

In this paper, we analyzed how idiomatic expressions are represented in AMR and evaluated whether text-to-AMR parsers can distinguish literal from figurative readings. An analysis of the AMR 3.0 corpus showed that idioms are under-represented and inconsistently annotated.

To address this gap, we constructed a dedicated corpus of potentially idiomatic expressions from MAGPIE, covering over 3,500 instances of 41 frequent idioms with both literal and figurative uses. Each expression was semantically normalized using a rule-based procedure, either by mapping it to an existing PropBank frame or by replacing it with a

figurative paraphrase. This corpus was then used to fine-tune Flan-T5 models for idiom-aware AMR parsing.

Experimental results show that balanced distributions of literal and figurative examples improve model performance, suggesting that generalization across idiomatic senses depends on exposure to both readings.

6.1. Limitations

However, several limitations remain. First, only a restricted set of idiomatic expressions was covered, which limits the generalization of the trained models. A larger and more diverse corpus would likely improve the robustness of idiom-aware parsing. Second, the annotation process lacked multiple independent annotators, so inter-annotator reliability was not assessed. Third, some idioms that are difficult to paraphrase still require the definition of additional PropBank frames to produce accurate AMR graphs (e.g., *break the ice*). Finally, the corpus is limited to English, restricting linguistic scope. Future work is needed to expand idiom coverage and other languages. Nonetheless, simply translating the existing corpus is unlikely to preserve idiomatic equivalence, since each language possesses its own distinct system of idioms.

7. Bibliographical References

- Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. [Generating high quality proposition Banks for multilingual semantic role labeling](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China. Association for Computational Linguistics.
- Timothy Baldwin and Su Nam Kim. 2010. [Multi-word expressions](#). In *Handbook of Natural Language Processing*.
- 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.

- Sara Beck and Andrea Weber. 2020. [Context and literality in idiom processing: Evidence from self-paced reading](#). *Journal of Psycholinguistic Research*, 49.
- 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.
- Dang Chung. 2024. [Challenges of translating idiomatic expressions: A cross-linguistic analysis at a university in hanoi, vietnam](#). *International Journal of Social Science and Human Research*, 07.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#).
- Kilian Evang, Rafael Ehren, and Laura Kallmeyer. 2025. [The proper treatment of verbal idioms in German discourse representation structure parsing](#). In *Proceedings of the 16th International Conference on Computational Semantics*, pages 156–165, Düsseldorf, Germany. Association for Computational Linguistics.
- 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.
- Wei He, Marco Idiart, Carolina Scarton, and Aline Villavicencio. 2024. [Enhancing idiomatic representation in multiple languages via an adaptive contrastive triplet loss](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12473–12485, Bangkok, Thailand. Association for Computational Linguistics.
- Johannes Heinecke. 2023. [metAMoRphosED, a graphical editor for Abstract Meaning Representation](#). In *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantics (ISA-19)*, pages 27–32, Nancy, France. Association for Computational Linguistics.
- Robert T. Kasper. 1989. [A flexible interface for linking applications to Penman’s sentence generator](#). In *Speech and Natural Language: Proceedings of a Workshop Held at Philadelphia, Pennsylvania, February 21-23, 1989*.
- Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1989–1993, Las Palmas, Canary Islands, Spain. European Language Resources Association.
- Young-Suk Lee, Ramón Fernández Astudillo, Radu Florian, Tahira Naseem, and Salim Roukos. 2023. [Amr parsing with instruction finetuned pre-trained language models](#).
- Behrooz Mansouri. 2025. [Survey of abstract meaning representation: Then, now, future](#).
- Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. [Idioms](#). *Language*, 70(3):491–538.
- Juri Opitz. 2023. [SMATCH++: Standardized and extended evaluation of semantic graphs](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1595–1607, Dubrovnik, Croatia. Association for Computational Linguistics.
- Juri Opitz, Anette Frank, and Letitia Parcalabescu. 2020. [Amr similarity metrics from principles](#). *Transactions of the Association for Computational Linguistics*, 8(0):522–538.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.
- Carlos Ramisch. 2023. [Multiword expressions in computational linguistics](#). Habilitation à diriger des recherches, Aix Marseille Université (AMU).
- Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019. [Without lexicons, multiword expression identification will never fly: A position statement](#). In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.
- Ziheng Zeng and Suma Bhat. 2021. [Idiomatic expression identification using semantic compatibility](#). *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

8. Language Resource References

Kevin Knight and Bianca Badarau and Laura Baranescu and Claire Bonial and Madalina Bar-docz and Kira Griffitt and Ulf Hermjakob and Daniel Marcu and Martha Palmer and Tim O’Gorman and Nathan Schneider. 2020. *Abstract Meaning Representation (AMR) Annotation Release 3.0*. Linguistic Data Consortium. distributed via LDC: LDC2020T02, 3.0, ISLRN [676-697-177-821-8](#).