AMR Parsing beyond English: an Experiment on Bulgarian, French, Hungarian and Ukrainian

Ivaylo Mitov[1], **Tadzhat Marharian**[1], **Zsofia Hauk**[1], **Samba Fall**[1], **Maxime Amblard**[2], **Bruno Guillaume**[2]
[1]IDMC, Université de Lorraine, France;
[2]Université de Lorraine, CNRS, Inria, LORIA, France;

Abstract

Under the assumption that the meaning of a sentence should be unchanged when it is translated into another language, recent work has developed on cross-lingual semantic parsing in an effort to extend the access to semantic resources beyond English. In this paper, we develop the automatic production of Abstract Meaning Representations (AMR), a graph-based semantic formalism, for four languages – Bulgarian, French, Hungarian and Ukrainian. We achieve high-performance on French and Hungarian, and execute, to our knowledge, the first semantic parsing of Bulgarian and Ukrainian on translations of the AMR3.0 corpus (Knight et al., 2020). Furthermore, we perform a complementary experiment on a novel parallel corpus of gold AMR annotations of the first chapter of *"The Adventures of Pinocchio"* in Bulgarian and Ukrainian. The experiment reveals that, despite their above-average performance, the models' performance decreases when probed on texts outside of the domain of the training data.

Keywords: cross-lingual, parsing, AMR, semantics, corpus

1. Introduction

Semantic representation is a fundamental task in Natural Language Processing, that aims to encode linguistic meaning in a manner that is interpretable by machines. Recent advances have enabled self-supervised models to project language into high-dimensional vector spaces, under the assumption that the meaning of linguistic content can be defined by its surrounding context (Devlin et al., 2019). Although fruitful, this approach raises concerns regarding its environmental and financial costs, as well as the faithfulness and interpretability of the vector embeddings, due to their data-driven and implicit nature. In order to address these issues, there is a call for the integration of symbolic semantic representations as a way to increase the transparency of modern AI systems. This approach typically falls within the broader category of *neuro-symbolic AI* (Conia et al., 2024).

The formal representation of meaning has been a long-standing area of research in language modelling and can be largely divided into logic-based and graph-based approaches. While the former makes use of tools such as higher-order logic and lambda calculus, its interpretation is challenging for non-experts. On the other hand, the latter offers a more readable representation, which ties into the need for models to be easier to interpret. A popular graph-based formalism is Abstract Meaning Representation (AMR) (Banarescu et al., 2013), which is going to be the focus of this paper.

Unlike the aforementioned vector embedding techniques, symbolic semantic representations make use of lexical resources that list predicate senses, used as nodes in the constructed graphs. While such resources exist for English, this is not the case for the majority of the world's languages, due to the extensive costs and task complexity linked to their creation. These factors are especially prevalent for less-resourced languages, where technological developments in the field remain out of reach for such language communities, thus creating or exacerbating pre-existing inequalities (Vanroy and Van de Cruys, 2024). Today, research conducted in multilingual NLP aims to reduce these gaps by leveraging transfer learning techniques. Given the advent of Neural Machine Translation (NMT) and cross-lingual AMR parsing (Blloshmi et al., 2020; Uhrig et al., 2021), semantic parsing has become more feasible for less-resourced languages.

In this paper we propose to explore cross-lingual AMR parsing for Bulgarian, French, Hungarian, and Ukrainian by leveraging state-of-the-art NMT and semantic parsing models on gold AMR annotations from the AMR3.0 corpus. To our knowledge, no work has yet explored AMR parsing for Bulgarian and Ukrainian inputs, and there has been no work on Hungarian that integrated gold AMR annotations for training. Furthermore, we evaluate the Bulgarian and Ukrainian models on our novel corpus of sentences annotated in AMR from the first chapter of Carlo Collodi's "The Adventures of Pinocchio" in order to test how they perform on real sentences from a domain that is underrepresented in their training data. The Pinocchio annotations are available at `https://github.com/YARN-World/YARN_Pinocchio`.

The remainder of this paper is structured as follows: In Section 2, we introduce the Abstract Meaning Representation (AMR) formalism and previous

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9733–9744
11–16 May 2026.

work done on the topic of cross-lingual AMR parsing. Section 3 lays out the evaluation metrics used and Section 4 describes the AMR3.0 corpus along with its translation into our four target languages, and our novel Pinocchio AMR corpus. Following, Section 5 describes the fine-tuning process for the parser models, while Section 6 presents the results for their performance on both the AMR3.0 test set and the Pinocchio AMR corpus. In Section 7 we discuss our findings, coupled with a manual inspection of the models' outputs. Finally, Section 8 summarizes our contributions and outlines avenues for future work.

2. Related Work

Abstract Meaning Representation (AMR) is a popular formalism that represents sentence-level semantics as rooted, directed, and acyclic graphs. The nodes are concepts or predicates (marked with a specific sense ID), while the edges depict semantic relations that exist between these concepts. AMR captures "who does what to whom" primarily through core semantic roles like :ARG0 (agent), :ARG1 (patient or theme), and :ARG2 (recipient or benefactive), with additional roles like :ARG3 and :ARG4 as needed. It also encodes adjunct information using relations such as :time, :location, :manner, :cause, :purpose, :mode and more. Notably, AMR abstracts away from the syntactic structure of the sentence, thus allowing the same representation to be valid for multiple paraphrases.

The framework is intrinsically tied to PropBank framesets (Kingsbury and Palmer, 2002; Palmer et al., 2005; Gildea and Palmer, 2002), an English lexical resource of predicate senses and their corresponding actants. As a result, it is biased towards specificities of the English language. Nevertheless, its abstraction from syntactic structure marks it as a useful tool for multilingual applications (Uhrig et al., 2021).

Cross-lingual AMR Parsing Damonte and Cohen (2018) first introduced the task by proposing two main methods: `annotation projection` and `translate+parse`. The `annotation projection` method aligns AMR nodes to target-language words based on word alignments between a source English sentence and its target language equivalent. On the other hand, the `translate+parse` approach leverages machine translation to translate target language sentences into English and applying existing AMR parsers on them. As such, so long as the quality of the translation is high, the method avoids the need for alignment.

Recent work has explored the `translate+parse` approach to cross-lingual AMR parsing for various languages. For example, Mitreska et al. (2022) developed a pipeline for German, Spanish, Italian, Bulgarian, and Macedonian, exploiting parallel Europarl data (Koehn, 2005). Their pipeline entails: (1) translating these sentences into English (here, via Txtai4[1] and Google translate using the DeepTranslator library[2]), (2) parsing the linguistic content of these English translations via an English AMR parser, (3) converting to English text the obtained semantic graphs via AMR-to-text models, and (4) back-translating these texts to their respective original languages. Performance is evaluated using cosine similarity between the embeddings of the original and final texts, obtained via several multilingual sentence encoders. Similarly, Uhrig et al. (2021) applied an English AMR parser to German, Spanish, Italian, and Mandarin Chinese texts that were translated with Opus-MT (Tiedemann and Thottingal, 2020; Tiedemann, 2020), and then evaluated the parsed graphs directly against gold annotations and sentences in those languages from the LDC2020T07 benchmark (AMR2.0 - Four Translations[3]), bypassing AMR-to-text and back-translation steps.

While these works leverage existing English AMR parsers on translations, advances have been made in developing systems that output English AMR graphs for non-English inputs. This approach was proposed by Blloshmi et al. (2020): they first translated English sentences of AMR2.0 into German, Italian, Mandarin Chinese, and Spanish, then used the gold AMR2.0 annotations to train language-specific, bilingual and multilingual parsers, using their XL-AMR sequence-to-sequence model. They also integrated a pre-processing step to filter out bad translations; they measured the cosine between the original English sentences and the back-translations (to English) of the German, Italian, Mandarin Chinese, and Spanish translations.

Barta et al. (2025) apply a similar approach to Hungarian. They used the Europarl parallel corpus (Koehn, 2005) to generate silver training data (Hungarian translations) and an English AMR parser to produce the associated, silver AMR graphs for this data, also exploring how distinct model architectures (`mT5-large` and `LLaMA 3 2.1B`) affect AMR graph generation performance. They achieve a SMATCH F1 score, see section 3.2, of 72.90 on the (translated) Hungarian AMR 3.0 test set, demonstrating that these approaches can be fruitful. Further, Boritchev and Heinecke (2023) and Kang et al. (2024) conducted work on AMR parsing for French, with the former achieving a SMATCH F1 score of 74.00 on the AMR2.0 test set, which is, to

[1] https://neuml.github.io/txtai/pipeline/text/translation/
[2] https://pypi.org/project/deep-translator/#id1
[3] https://catalog.ldc.upenn.edu/LDC2020T07

our knowledge, the best reported score for French in the field. They also provide a useful framework for the error categorization of AMR parsers.

3. Evaluation Metrics

We will consider two types of evaluation metrics: translation metrics to measure translation quality and graph-comparison metrics to evaluate silver AMR graphs against a gold standard.

3.1. Translation Metrics

As our approach depends on high-quality translations, we employ the following translation metrics: (1) BLEU (1~4) (Papineni et al., 2002) to measure lexical similarity, (2) cosine similarity on sentence embeddings via `bert-base-uncased` (Devlin et al., 2019) to compare semantic content, and (3) COMET (Rei et al., 2020) to produce human-like judgements of translation quality, simultaneously taking into account the original English sentence, its translation and back-translation.

3.2. Graph Comparison Metrics

The performance of the parsers is evaluated by comparing the gold AMR graphs and the produced AMR graphs with SMATCH (Cai and Knight, 2013) – the canonical metric for comparing two AMR graphs. It functions by representing each graph as a set of triples (e.g. ("want-01", "ARG0", "boy")) and computing the precision, recall, and F1-score between those two sets.

4. Datasets

4.1. AMR3.0

The Abstract Meaning Representation (AMR) Annotation Release 3.0 (AMR3.0) is a corpus released by the Linguistic Data Consortium (LDC) under an LDC User Agreement for Non-Members license. It is the largest corpus of gold AMR annotations for English, containing 59,255 English natural language sentences paired with their associated AMR graphs (in PENMAN notation). Given that these sentences come from various sources (online forums, news articles and transcripts, English translations of Chinese texts, weblogs, literary texts, etc.) [4], the entire corpus can be divided into 13 subsets, where each source is its own subset. Also, AMR3.0 comes with predefined train, test, and dev splits, the sizes of which are listed in Table 1.

[4] For more details regarding the genre distribution, see https://catalog.ldc.upenn.edu/LDC2020T02

Split	Train	Dev	Test	Total
Count	55,635	1,722	1,898	59,255
Percent	93.9%	2.9%	3.2%	100%

Table 1: Number of sentences and Proportions of the train, dev, and test splits of the AMR3.0 corpus.

4.2. Translation of AMR3.0

The sentences in the AMR3.0 corpus were translated into Bulgarian, French, Hungarian, and Ukrainian using state-of-the-art neural machine translation (NMT) models. As not every entry in the corpus corresponds to text that requires translation (e.g., numbers, dates, times, codes, links, etc.), such cases were identified using regular expressions and were not selected. The models used for translation were `opus-mt-tc-big` models by Helsinki-NLP[5] for Bulgarian, French and Hungarian, and `ct2fast-m2m100_1.2B`[6] - a fast-inference implementation of `m2m100_1.2B` (Fan et al., 2021) for Ukrainian. The models were selected by comparing the performance of several machine translation models on a subset of the corpus, which has been detailed in Appendix A.

In order to estimate the quality of the translations, translations were back-translated into English, and then compared with the original English source sentences using BLEU score, cosine similarity of sentence embeddings and COMET score (see 3.1), which are presented in Table 2. Across all languages, BLEU scores follow the expected declining trend from BLEU-1 to BLEU-4, reflecting the increasing difficulty of matching longer n-grams. French and Bulgarian achieve the highest BLEU scores overall, with BLEU-4 scores of 0.47 for both, while Ukrainian consistently scores lowest across all BLEU variants. COMET scores are consistently high across all languages, indicating strong semantic adequacy and fluency in the translations. French achieves the highest COMET score (0.90), closely followed by Bulgarian (0.89) and Hungarian (0.88). Ukrainian again lags slightly behind with a score of 0.85. Cosine similarity scores largely mirror this trend, further confirming the preservation of semantic content for all languages. It is important to note, however, that it is still possible to have a bad quality translation which results in a good back-translation.

[5] https://huggingface.co/Helsinki-NLP

[6] https://huggingface.co/michaelfeil/ct2fast-m2m100_1.2B

Metric	BG	FR	HU	UK
BLEU-1	0.76	0.75	0.68	0.65
BLEU-2	0.66	0.64	0.56	0.51
BLEU-3	0.56	0.55	0.45	0.41
BLEU-4	0.47	0.47	0.36	0.32
COMET	0.89	0.90	0.88	0.85
Cosine Similarity	0.86	0.85	0.82	0.80

Table 2: Summary of Translation Metrics on the whole AMR3.0 corpus.

4.3. Pinocchio AMR Corpus

We present the Pinocchio AMR corpus - a novel parallel corpus of the first chapter of Carlo Collodi's "The Adventures of Pinocchio" in English, Bulgarian, French, Hungarian and Ukrainian. The Bulgarian and Ukrainian texts were annotated in AMR with English PropBank frames by native speakers in the authors' collective experienced in AMR annotation. Additionally, silver AMR annotations were produced with the models developed during this study for all languages. An overview of the corpus is presented in Table 3, including sentence and average word counts and the source of the texts. As not all texts are public domain, in such cases we release only the AMR annotations.

Language	#sent	#words (avg.)	Source
English	54	12.76	(Collodi, 2006)
Bulgarian	41	14.02	(Collodi, 1950)
French	49	13.41	(Collodi, 2004)
Hungarian	50	10.52	(Collodi, 2003)
Ukrainian	54	8.83	(Collodi, 1967)

Table 3: Overview of the Pinocchio AMR Corpus.

5. Methodology

Here, we introduce our pipeline for cross-lingual AMR Parsing for Bulgarian, French, Hungarian, and Ukrainian. Figure 1 presents an overview of the process: filtering out bad translations, fine-tuning and evaluation.

5.1. Filtering Out Bad Translations

The filtering step was considered necessary due to the nature of the task – we want to leverage gold AMR annotations for English sentences by translating them into a target language and use the AMR graphs as a language-agnostic representation of their meaning. As such, it is intuitive that increasing the quality of the translation would result in a better performing semantic parser. However, this comes with a trade-off between quality and quantity, since filtering inevitably reduces the amount of data, and as a result, this could lead to the opposite effect where the parser's performance stagnates or decreases.

We have $(\mathrm{S}^{\mathrm{L}}, \mathrm{AMR}_{\mathrm{G}})$ pairs, where S^{L} is a set of sentences in a given language $\mathrm{L} \in \{\mathrm{EN}, \mathrm{BG}, \mathrm{FR}, \mathrm{HU}, \mathrm{UK}\}$, and $\mathrm{AMR}_{\mathrm{G}}$ is a set of gold AMR annotations from AMR3.0. As mentioned in Section 4, the corpus comes with predefined train, dev, and test splits, which we refer to as $\mathrm{train}^{\mathrm{L}}$, $\mathrm{dev}^{\mathrm{L}}$, and $\mathrm{test}^{\mathrm{L}}$ for each language. We filtered each of them and later used both their filtered and unfiltered versions to measure the effect that filtering has on the performance of the parser.

The $\mathrm{train}^{\mathrm{L}}$ and $\mathrm{dev}^{\mathrm{L}}$ splits were filtered by first dropping all duplicate $(\mathrm{S}^{\mathrm{L}}, \mathrm{AMR}_{\mathrm{G}})$ pairs and making sure that there were also no duplicates between $\mathrm{train}^{\mathrm{L}}$ and $\mathrm{dev}^{\mathrm{L}}$, and $\mathrm{train}^{\mathrm{L}}$ and $\mathrm{test}^{\mathrm{L}}$. Following this, we kept only translations which had a COMET score (see 3) above 0.7 (a stricter value would have substantially decreased the training data for Ukrainian), resulting in $\mathrm{train}^{\mathrm{L}}_{\mathrm{df}}$ and $\mathrm{dev}^{\mathrm{L}}_{\mathrm{df}}$. As for $\mathrm{test}^{\mathrm{L}}$, we obtained a version of it just without duplicates $\mathrm{test}^{\mathrm{L}}_{\mathrm{d}}$, and a version both without duplicates and with COMET filtering applied $\mathrm{test}^{\mathrm{L}}_{\mathrm{df}}$ in order to isolate the effect filtering has on the results. Note that for $\mathrm{L} = \mathrm{EN}$, $\mathrm{test}^{\mathrm{L}}_{\mathrm{d}} = \mathrm{test}^{\mathrm{L}}_{\mathrm{df}}$, as well as that $\mathrm{train}^{\mathrm{L}}_{\mathrm{df}} = \mathrm{train}^{\mathrm{L}}_{\mathrm{d}}$ and $\mathrm{dev}^{\mathrm{L}}_{\mathrm{df}} = \mathrm{dev}^{\mathrm{L}}_{\mathrm{d}}$ as no COMET filtering is performed. Also, by convention we omit the superscript L when we are talking about the splits in general, not referring to those of a specific language. A summary of the sentence counts and remaining percentage of each split per language is shown in Table 4.

5.2. Fine-tuning

We fine-tuned a `bart-large` model (Lewis et al., 2020) using the CLAP architecture (Martinez Lorenzo and Navigli, 2024), which offers an efficient and flexible linearization method, cutting token counts by 40–50% by removing redundant syntax (e.g., slashes and parentheses), simplifying named entities and frames, and omitting unnecessary node variables.

Three types of training were conducted on $(\mathrm{S}^{\mathrm{L}}, \mathrm{AMR}_{\mathrm{G}})$ pairs using filtered/unfiltered train/dev combinations of the AMR3.0 corpus: $\mathrm{train}/\mathrm{dev}$, $\mathrm{train}_{\mathrm{df}}/\mathrm{dev}$, and $\mathrm{train}_{\mathrm{df}}/\mathrm{dev}_{\mathrm{df}}$. Then, each model was tested on test, $\mathrm{test}_{\mathrm{d}}$, and $\mathrm{test}_{\mathrm{df}}$. By comparing the results for each testing instance it is possible to grasp the effect of filtering out bad translations more clearly. Furthermore, a complementary experiment was performed on the Pinocchio AMR corpus for Bulgarian and Ukrainian to test the models' robustness on real (non-automatically translated) text from an underrepresented domain in the training

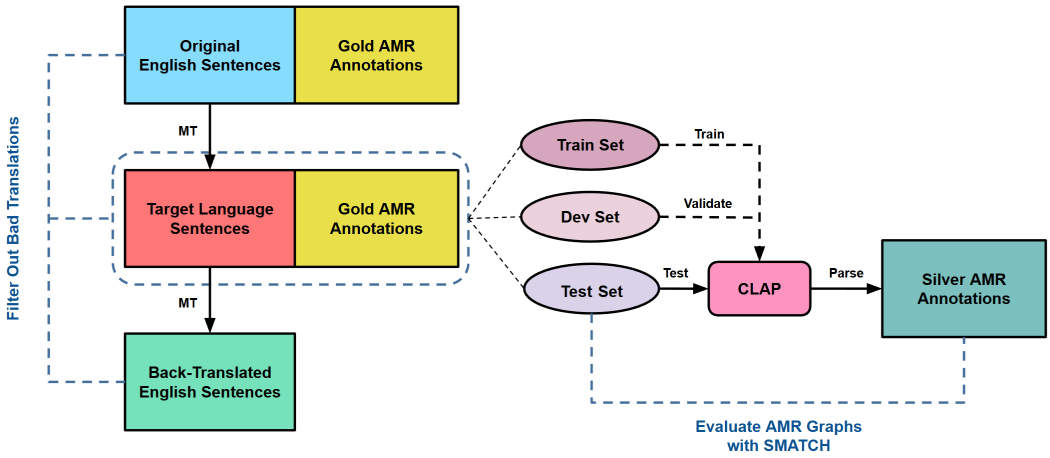


Figure 1: Cross-lingual AMR Parsing Pipeline.

Split	EN	BG	FR	HU	UK
train_{df}	54,524 (98.00%)	53,143 (95.52%)	53,278 (95.76%)	52,727 (94.77%)	49,639 (89.21%)
dev_{df}	1,694 (98.37%)	1,679 (97.50%)	1,677 (97.39%)	1,681 (97.62%)	1,629 (94.60%)
test_{d}	1,830 (96.42%)	1,826 (96.21%)	1,829 (96.36%)	1,829 (96.36%)	1,828 (96.31%)
test_{df}	-	1,799 (94.78%)	1,807 (95.02%)	1,799 (94.78%)	1,730 (91.14%)

Table 4: Sentence counts and remaining percentages of the train, dev, and test splits after filtering for all languages.

data – literary. Each model outputs silver AMR annotations in English for the input language ($\text{AMR}^{\text{L}}_{\text{S}}$).

Fine-tuning was conducted using NVIDIA A40 (300 W TDP), each model running for 20 epochs. Batch sizes were 10000 tokens for English and French, and 6000 for the other languages to avoid memory issues. The cumulative training time was ≈120 h, corresponding to an estimated electrical energy use of ≈58 kWh.

6. Results

Here, we detail the results of the models on the filtered/unfiltered variations of the test set of the translated AMR3.0 corpus for all languages and on the Pinocchio AMR corpus for Bulgarian and Ukrainian.

6.1. Performance on the Translated AMR3.0 Corpus

The SMATCH F1 scores for the parser for all languages are shown in Table 5. Scores on the train/dev/test splits are presented for comparison with similar works but are not considered for best scores as they contain duplicates.

The highest SMATCH score is obtained on the English parser (82.7), followed by French (74.4) on the $\text{train/dev/test}_{\text{df}}$ splits. The third and fourth best performing models are only marginally different with Bulgarian and Hungarian scoring 69.5 and 69.4 respectively on the $\text{train}_{\text{df}}/\text{dev/test}_{\text{df}}$ splits. Finally, the lowest score was obtained by the Ukrainian one – 67.6 on the same splits.

A noticeable trend exists when comparing the results for the same train/dev combination on the three test sets - test, test_{d} and test_{df}. The score consistently drops after dropping duplicates and there is always an increase after filtering out bad translations from the test data. Furthermore, when comparing the difference between the scores on test and those on test_{df} for most models we observe a marginal increase of around 0.02 points. However, for Ukrainian there is a relatively larger increase of 0.06 to 0.08 points. Overall, the effects of filtering is minimal as scores for a given language tend to remain consistent. Still, leaving the dev split unfiltered on average scores higher than when it is filtered.

6.2. Performance on the Pinocchio AMR Corpus

The experiment on the Pinocchio AMR corpus was performed using the $\text{train}_{\text{df}}/\text{dev}_{\text{df}}$ configuration to maintain a homogenous and duplicate-free setup,

Splits	EN	BG	FR	HU	UK
$\mathrm{train/dev/test}$	82.9	69.2	74.2	69.0	67.1
$\mathrm{train/dev/test_d}$	**82.7**	69.1	74.0	68.8	66.7
$\mathrm{train/dev/test_{df}}$	-	69.2	**74.4**	69.1	67.5
$\mathrm{train_{df}/dev/test}$	82.3	69.3	73.6	69.2	67.0
$\mathrm{train_{df}/dev/test_d}$	82.2	69.2	73.5	69.0	66.5
$\mathrm{train_{df}/dev/test_{df}}$	-	**69.5**	74.0	**69.4**	**67.6**
$\mathrm{train_{df}/dev_{df}/test}$	**82.7**	69.2	74.0	68.8	65.2
$\mathrm{train_{df}/dev_{df}/test_d}$	**82.7**	69.1	73.7	68.6	64.9
$\mathrm{train_{df}/dev_{df}/test_{df}}$	-	69.4	74.1	69.0	66.0

Table 5: SMATCH F1 scores for every language and all split combinations. *d*: no duplicates; *df*: no duplicates and COMET > 0.7 filtering. Best scores in **bold**.

at the cost of only a marginal decrease in performance. When evaluated against the gold annotations of the Bulgarian and Ukrainian versions of the corpus, the parser's scores decline substantially – from 69.4 (on $\mathrm{test_{df}}$) to 47.0 for Bulgarian and from 66.0 (on $\mathrm{test_{df}}$) to 43.1 for Ukrainian. A summary of the metrics is presented in Table 6.

Language	SMATCH F1
Bulgarian	47.0
Ukrainian	43.1

Table 6: SMATCH scores of the $\mathrm{train_{df}/dev_{df}}$ configuration of the semantic parser outputs for the Bulgarian and Ukrainian subsets of the Pinocchio AMR corpus compared to gold annotations.

7. Discussion

7.1. Effect of Translation Quality and Filtering

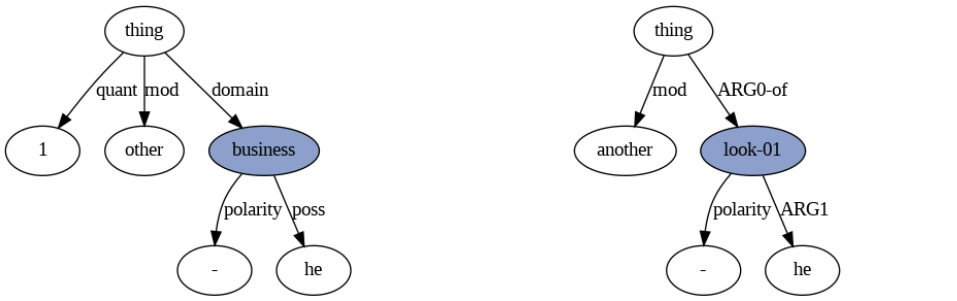
Based on the results obtained on the translated AMR3.0 test set, the ranking of parsers per language is consistent with the ranking by translation quality (Table 2), indicating that translation quality does influence the overall performance of the parser. However, while French and Bulgarian scored similarly in terms of BLEU, cosine similarity, and COMET, the French AMR graphs were better. This could be because of the lexical similarities between French and English, enabling more accurate English AMR annotations for French sentences.

Next, it is hard to grasp the effect that filtering out bad translations had on parser performance: the results for $\mathrm{train_{df}}$ are only marginally better than those for train. We assume that this is partly due to the low COMET score threshold (0.7) that we used. Whereas the Bulgarian, French, and Hungarian splits did not change significantly after filtering, the results for Ukrainian parsing seemed to indicate that filtering had a positive effect on parser performance, based on comparing $\mathrm{test^{UK}}$, $\mathrm{test_d^{UK}}$ and $\mathrm{test_{df}^{UK}}$. Interestingly, however, there is a noticeable difference between using dev and $\mathrm{dev_{df}}$ for Ukrainian: the former has better results, possibly indicating that lower quality translations in the dev set enable the model to generalize better.

7.2. Manual Inspection of Outputs on Translated AMR3.0

We studied graphs for each language to put forth preliminary observations. To remind our reader(s), our pipeline involves both a translation and a parsing component, and so the errors we observed in the final graphs could originate from either step. As such, we identify four error scenarios: (1) accurate translation with an inaccurate AMR graph, (2) inaccurate translation with an accurate AMR graph w.r.t the original meaning, (3) accurate translation with an accurate AMR graph (with synonymous nodes, however), and (4) inaccurate translation with an inaccurate AMR graph w.r.t original meaning. Regardless of the scenario (and target language), named entities, wikification, less frequent words (e.g., "baffle"), and abbreviations (e.g., "SOS") were often badly translated or misrepresented in graphs. Also, the AMR graphs for longer sentences were sometimes incomplete, and the parser often hallucinated predicates that do not exist in the PropBank frames.

Figure 2 shows a **Scenario (1)** example, where the English sentence *"...one other thing it is none of his business..."* is more or less accurately translated to French as *"...une autre chose qui ne *lui regarde pas..."* The resulting AMR graph, however, represents a literal, inaccurate interpretation (roughly, *"one other thing that does not look at him/her/it"*).



*"...one other thing it is **none of his business...**"* *"...une autre chose qui **ne** * **lui regarde pas...**"*

Figure 2: Example of Scenario (1) - accurate translation with an inaccurate AMR graph (French)

For **Scenario (2)**, we have the sentence, *"Are you on any drugs or getting any therapy?"*; *"get therapy"* is literally translated to Bulgarian as "получавам терапия" ('receive therapy'), when the correct expression is "ходя на терапия" ('go to therapy') (Figure 3). Its AMR graph, however, is correct, revealing that, although some sentences are direct (inaccurate) translations from English, they enable the parser to more easily produce their correct, English-labelled AMR graphs. As such, parser performance should also be tested with gold translations as input.

For **Scenario (3)**, the English sentence *"We're hungry?"* is accurately translated into Hungarian as *"Éhesek vagyunk?"*, but the output graph has the synonymous predicate `starve-01`, instead of `hunger-01` (with "*starve*" being but a more emphatic version of "*hunger*").

Finally, for **Scenario (4)** we have the sentence *"You all are dead wrong"*, translated to Ukrainian as "Ви всі мертві помиляєтеся" (roughly, *'All you dead are wrong'*). We can infer that the translation model did not recognize *"dead"* as a modifier of *"wrong"* and instead gave a literal translation. The resulting graph, however, is also inconsistent with the input translation: we obtain a semantic representation roughly equivalent to *"You all are wrong about death"*.

We also note that there are multiple errors stemming from **typological differences** between English and the target languages. For instance, though English maintains grammatical gender for its personal and possessive pronouns, it is far more ubiquitous in languages like French, Bulgarian, and Ukrainian, where adjectives and participles, for example, retain explicit grammatical gender marking for correctness. As such, translations often default to a specific gender that was not explicitly stated in the original English input sentence. Hungarian, on the other hand, does not exhibit any grammatical gender, which has the adverse effect: information regarding gender is often lost during the translation step.

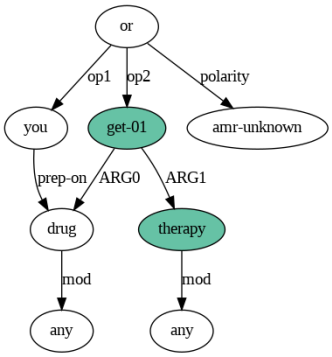
Another case of losing gender information is linked with pronoun dropping in Bulgarian, Hungarian, and Ukrainian. While English and French state the subject of a predicate (to avoid verb inflection ambiguity in speech), when translated to a *pro-drop* language, this could lead to gender ambiguity when there are no gender-marked adjectives or participles, for example. Further, in French, there is syncretism with the singular dative personal pronoun, "*lui*", where the feminine and masculine forms are identical, and so translation can cause gender information to be lost.

Overall, despite some issues stemming from the translation and parsing steps, as well as from typological differences between English and the target languages, we maintain that there are plenty of satisfying examples, reflecting the above-average SMATCH F1 scores.

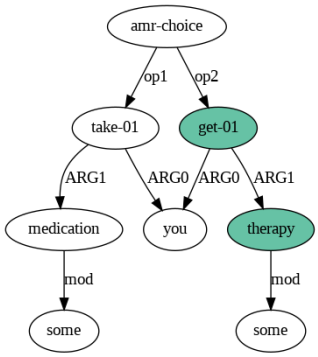
7.3. Manual Inspection of Outputs on The Pinocchio AMR Corpus

The SMATCH scores calculated between the output graphs and the gold annotations of the Bulgarian and Ukrainian version of the Pinocchio AMR corpus revealed that there was a significant drop in parsing quality. Among the errors, we find that the models tend to struggle most when faced with uncommon words or expressions, which are prevalent in the corpus. For instance, both the Bulgarian and the Ukrainian texts often contain diminutive forms of nouns, specialized terms related to woodwork and old-fashioned words which now would be considered obsolete. In many cases, this caused the models to output what seemed to be random graphs which had little to no connection to the meaning of the input text.

A Ukrainian example is presented in Figure 4, where the words "трубі" (*hrubi*, 'hearth', N.FEM.SING.LOC) and "печі" (*pechi*, 'woodstove', N.FEM.SING.LOC) are incorrectly represented in the AMR graph as `chest` and `shoulder`, respectively. We infer that this is due to

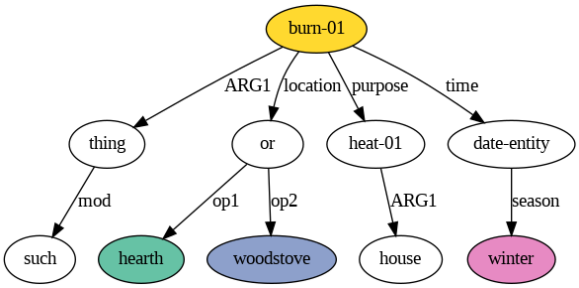


*"Are you on any drugs or **getting any therapy?**"*



"Взимаш ли някакви лекарства или получаваш някаква терапия?"

Figure 3: Example of Scenario (2) - inaccurate translation with an accurate AMR graph (Bulgarian)



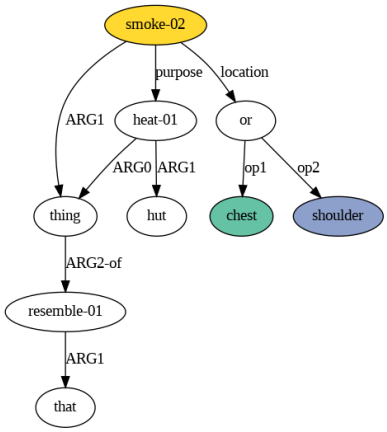


Figure 4: Gold (left) and Silver (right) AMR graph of the Ukrainian sentence "Такими взимку палять у грубі або в печі, щоб нагріти хату." ("*Such things are burned in the hearth or woodstove to heat the house in winter.*") from the Pinocchio AMR corpus.

their surface-level resemblance to "груди" (*hrudy*, 'chest', N.PL.NOM) and "плечі" (*plechi*, 'shoulder', N.PL.NOM). In addition to failing to use case information to disambiguate the meaning, the Ukrainian word for "hearth" is rarely used in contemporary speech. Similarly, "взимку" (*vzymku*, 'in winter', ADV) is more commonly used in literary texts, and as such could be the reason why the model omitted it from the graph altogether. Finally, "палять" (*palyat*, 'to burn/to smoke', V.3PL.PRS.INDIC.IPFV) is polysemous, hence the incorrect predicate `smoke-02`. Combined, these errors result in a meaning representation roughly equivalent to "*Such things are smoked in the shoulder or chest to heat the hut.*"

From the experiment we can conclude that the cross-lingual approach of this study is not robust to domain shifting. Furthermore, it highlights a limitation of using silver translations which retain a high proximity to the original English text, instead of producing natural sentences that are adequate for the domain in question.

8. Conclusion

In this paper, we explored cross-lingual AMR parsing for Bulgarian, French, Hungarian and Ukrainian by leveraging machine translation on the AMR3.0 corpus. The models achieve near state-of-the-art performance for French and Hungarian, and represent the first to our knowledge AMR parsers for Bulgarian and Ukrainian. Also, we conducted a complementary experiment for Bulgarian and Ukrainian on our novel Pinocchio AMR corpus where we found that changing the domain and shifting from silver translations to real text shows difficulties for correct AMR graph production. Our study highlights the importance of manually curated benchmarks for cross-lingual AMR parsing spanning a diverse range of domains in order to have a realistic evaluation of parsers' capabilities. To support this, we contribute to the research community the Pinocchio AMR corpus with gold AMR annotations in Bulgarian and Ukrainian, and silver annotations in English, French and Hungarian.

Our manual inspection of the outputs revealed that there are many challenges for cross-lingual semantic parsing related to both the pipeline (i.e. translation quality, models' parametric knowledge) and to typological language differences (e.g. grammatical gender, pronoun omission, case). We also acknowledged the tendency for models to be swayed by surface-level token similarities and to incorrectly identify word senses. As such, this puts into question the ability of language models to accurately capture meaning at the sentence level and to produce accurate semantic structures.

Future work on this topic can experiment with different pre-trained language models for fine-tuning. Following Barta et al. (2025), where `mT5` consistently outperformed `Llama 3.2 1B` for Hungarian semantic parsing, as well as Boritchev and Heinecke (2023) where the underlying language model for French AMR parsing was substituted for `mT5`, we would be interested to see if this model improves Bulgarian and/or Ukrainian parser performance. Furthermore, exploring a more diverse set of graph comparison metrics that capture meaning similarity, rather than structural similarity, would allow for a more accurate evaluation of the output graphs (Opitz, 2023; de Vergnette et al., 2025).

9. Ethics Statement / Broader Impact

Cross-lingual AMR parsing raises ethical questions related to linguistic equity and representation. The creation and deployment of models trained primarily on high-resource languages such as English can reinforce global language hierarchies if the resulting systems underperform for underrepresented languages. This work attempts to mitigate such inequities by extending semantic parsing to Bulgarian, Hungarian, and Ukrainian—languages that are often underrepresented in computational linguistics. However, this approach still relies on English-centric resources and translation systems, which can introduce cultural and semantic biases that reflect the dominant language's worldview rather than the linguistic and cognitive structures of the target communities.

10. Limitations

A primary limitation of this study lies in its dependence on machine translation quality. The cross-lingual AMR pipeline relies on translating English inputs into other languages before parsing, which introduces compounding errors—especially for morphologically rich or typologically distant languages such as Ukrainian and Bulgarian. As shown in the evaluation on the Pinocchio corpus, domain shift and idiomatic language substantially degrade parsing accuracy. This indicates that the models may capture translation artifacts rather than genuine semantic equivalence, which limits the generalizability of the findings to real-world linguistic data.

Another limitation concerns the scope and diversity of training and evaluation data. While AMR3.0 provides a robust English benchmark, it does not cover the linguistic diversity necessary for fully evaluating multilingual semantic parsing. The new Pinocchio corpus is a valuable contribution, but its size and domain specificity (literary text) constrain its utility for broad performance assessment. Additionally, the evaluation relies heavily on the SMATCH metric, which does not fully capture nuanced differences in meaning or structure across languages.

11. Bibliographical References

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. Association for Computational Linguistics.

Botond Barta, Endre Hamerlik, Milán Konor Nyist, and Judit Ács. 2025. HuAMR: A hungarian AMR parser and dataset.

Rexhina Blloshmi, Rocco Tripodi, and Roberto Navigli. 2020. XL-AMR: Enabling cross-lingual AMR parsing with transfer learning techniques. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2487–2500. Association for Computational Linguistics.

Maria Boritchev and Johannes Heinecke. 2023. Error exploration for automatic abstract meaning representation parsing. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 246–251. 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.

Simone Conia, Edoardo Barba, Abelardo Carlos Martinez Lorenzo, Pere-Lluís Huguet Cabot, Riccardo Orlando, Luigi Procopio, and Roberto Navigli. 2024. MOSAICo: a multilingual open-text semantically annotated interlinked corpus. 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 7990–8004. Association for Computational Linguistics.

Marco Damonte and Shay B. Cohen. 2018. Cross-lingual abstract meaning representation parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1146–1155. Association for Computational Linguistics.

Rémi de Vergnette, Maxime Amblard, and Bruno Guillaume. 2025. Evaluation Framework for Layered Meaning Representation. In *Proceedings of The Sixth International Workshop in Designing Meaning Representation*, Prague, Czech Republic.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. 22(1):107:4839–107:4886.

Jeongwoo Kang, Maximin Coavoux, Cédric Lopez, and Didier Schwab. 2024. Should cross-lingual AMR parsing go meta? an empirical assessment of meta-learning and joint learning AMR parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 43–51. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Abelardo Carlos Martinez Lorenzo and Roberto Navigli. 2024. Efficient AMR parsing with CLAP: Compact linearization with an adaptable parser. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5578–5584. ELRA and ICCL.

Maja Mitreska, Tashko Pavlov, Kostadin Mishev, and Monika Simjanoska. 2022. xAMR: Cross-lingual AMR end-to-end pipeline:. In *Proceedings of the 3rd International Conference on Deep Learning Theory and Applications*, pages 132–139. SCITEPRESS - Science and Technology Publications.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. Association for Computational Linguistics.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480. European Association for Machine Translation.

Sarah Uhrig, Yoalli Garcia, Juri Opitz, and Anette Frank. 2021. Translate, then parse! a strong baseline for cross-lingual AMR parsing. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 58–64. Association for Computational Linguistics.

Bram Vanroy and Tim Van de Cruys. 2024. Less is enough: Less-resourced multilingual AMR parsing. In *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*, pages 82–92. ELRA and ICCL.

12. Language Resource References

Carlo Collodi. 1950. Приключенията на Пинокио. ИК Софи - Р. Translated by Petăr Dragoev. Originally published 1881.

Carlo Collodi. 1967. Пригоди Піноккіо. ЦК ЛКСМУ "Молодь". Translated by Yuriy Avdeev. Originally published 1881.

Carlo Collodi. 2003. *Pinokkió*. Aeternitas Irodalmi Műhely. Translated by Rónay György. Originally published 1881.

Carlo Collodi. 2004. *Les aventures de Pinocchio*. Ebooks libres et gratuits. Translated by Claude Sartirano. Originally published 1881.

Carlo Collodi. 2006. *The Adventures of Pinocchio*. Project Gutenberg. Translated by Carol Della Chiesa. Originally published 1881.

Daniel Gildea and Martha Palmer. 2002. The Necessity of Parsing for Predicate Argument Recognition. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 239–246, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. European Language Resources Association (ELRA).

Kevin Knight et al. 2020. Abstract meaning representation (amr) annotation release 3.0. Web Download. LDC2020T02.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. 31(1):71–106.

Appendix A. Comparison of Machine Translation models

We tested five neural machine translation models on a 1000 sentence sample of AMR3.0. The sample was designed to keep the distribution of the 13 subsets in the corpus, as well as the token count of the original English sentences for each subset. This was done to obtain meaningful results about each model's performances on sentences of varying lengths and semantic contents. Table 1 shows a summary of BLEU1$\tilde{4}$, cosine similarity (based on `bert-base-uncased`), and COMET scores per model and language. We concluded that `opus-mt-tc-big` is most suitable for Bulgarian, French and Hungarian, while `ct2fast-m2m100_1.2B` performs best for Ukrainian.

MT Model	Metric	BG	FR	HU	UK
opus-mt-tc-big	BLEU1	**0.79**	**0.78**	**0.72**	-
	BLEU2	**0.69**	**0.69**	**0.60**	-
	BLEU3	**0.61**	**0.61**	**0.51**	-
	BLEU4	**0.55**	**0.55**	**0.44**	-
	Cos. Sim.	**0.86**	**0.86**	**0.82**	-
	COMET	**0.90**	**0.90**	**0.88**	-
opus-mt	BLEU1	-	-	-	0.63
	BLEU2	-	-	-	0.49
	BLEU3	-	-	-	0.39
	BLEU4	-	-	-	0.31
	Cos. Sim.	-	-	-	0.79
	COMET	-	-	-	0.80
google translate	BLEU1	0.73	0.74	0.65	0.69
	BLEU2	0.61	0.62	0.51	0.56
	BLEU3	0.51	0.53	0.41	0.46
	BLEU4	0.44	0.46	0.39	0.39
	Cos. Sim.	0.85	0.86	0.81	**0.82**
	COMET	0.85	0.85	0.83	0.82
ct2fast-m2m100_1.2B	BLEU1	0.72	0.76	0.68	**0.71**
	BLEU2	0.61	0.65	0.56	**0.60**
	BLEU3	0.52	0.57	0.46	**0.50**
	BLEU4	0.45	0.51	0.39	**0.43**
	Cos. Sim.	0.81	0.82	0.79	0.80
	COMET	0.86	0.88	0.86	**0.85**
translation-bart	BLEU1	-	-	0.68	-
	BLEU2	-	-	0.55	-
	BLEU3	-	-	0.46	-
	BLEU4	-	-	0.38	-
	Cos. Sim.	-	-	0.81	-
	COMET	-	-	0.81	-

Table 1: Summary of translation evaluation metrics for five MT models on a stratified 1000 sentence sample of AMR3.0. Best results in **bold**.