# Mitigating Data Scarcity in Semantic Parsing across Languages: the Multilingual Semantic Layer and its Dataset

**Abelardo Carlos Martínez Lorenzo**[1], **Pere-Luís Huguet Cabot**[1], **Karim Ghonim**[1], **Lu Xu**[1], **Hee-Soo Choi**[2,3], **Alberte Fernández Castro**[4], **Roberto Navigli**[1]
[1]Sapienza NLP Group, Sapienza University of Rome
[2]ATILF, CNRS, Université de Lorraine [3]LORIA, Université de Lorraine [4]Roma Tre
[1]{lastname(s)}@diag.uniroma1.it
[2,3]hee-soo.choi@loria.fr [4]alb.fernandezcastro@stud.uniroma3.it

## Abstract

Data scarcity is a prevalent challenge in the era of Large Language Models (LLMs). The insatiable hunger of LLMs for large corpora becomes even more pronounced when dealing with non-English and low-resource languages. The issue is particularly exacerbated in Semantic Parsing (SP), i.e. the task of converting text into a formal representation. The complexity of semantic formalisms makes training human annotators and subsequent data annotation unfeasible on a large scale, especially across languages. To mitigate this, we first introduce the Multilingual Semantic Layer (MSL), a conceptual evolution of previous formalisms, which decouples from disambiguation and external inventories and simplifies the task. MSL provides the necessary tools to encode the meaning across languages, paving the way for developing a high-quality semantic parsing dataset across different languages in a semi-automatic strategy. Subsequently, we manually refine a portion of this dataset and fine-tune GPT-3.5 to propagate these refinements across the dataset. Then, we manually annotate 1,100 sentences in eleven languages, including low-resource ones. Finally, we assess our dataset's quality, showcasing the performance gap reduction across languages in Semantic Parsing. Our code and dataset are openly available at https://github.com/SapienzaNLP/MSL.

## 1 Introduction

One of the long-term goals of AI is to enable machines to comprehend human text in any language. At the core of Natural Language Understanding (NLU) lies the task of Semantic Parsing (SP), aiming to convert text into machine-interpretable representations. Although Large Language Models (LLMs) have advanced significantly in understanding human text, semantic representations remain crucial for various applications. These range from chatbots and Virtual Assistants like Amazon Alexa, Apple Siri, and Google Assistant, as underscored in Rongali et al. (2020), to more specific tasks, e.g., SQL queries (Dou et al., 2022) or machine-interpretable commands (Wang et al., 2021).

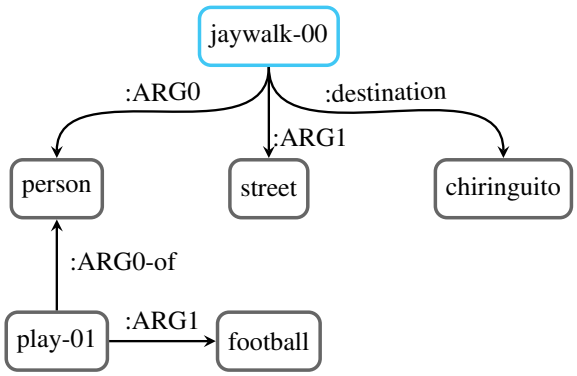



Figure 1: AMR of: *"The football player jaywalks across the street to the chiringuito."*

Achieving proficiency in SP involves enabling machines to understand semantic relations between concepts in any language. Over the years, numerous SP formalisms grounded in various linguistic theories have emerged, such as the Discourse Representation Theory (Kamp and Reyle, 1993, DRT), Prague Dependency Treebank (Nedoluzhko et al., 2016, PDT), Universal Conceptual Cognitive Annotation (Abend and Rappoport, 2013, UCCA) or Abstract Meaning Representation (Banarescu et al., 2013, AMR), among others. Graph-based representations – e.g., AMR – have attracted most of the attention since they can act as an interface that is comprehensible for humans and interpretable for machines (see Figure 1). The objective of graph-based representations is to encode the meaning of language into a directed acyclic graph, where nodes represent concepts and edges represent semantic relations between concepts. However, even if these formalisms are able to encode the meaning in English effectively, they still struggle to scale to other languages (Zhu et al., 2019).



*Findings of the Association for Computational Linguistics: ACL 2024*, pages 14056–14080

| Form. | Year | Corpus | Released | Size | Non-English languages | Non-English size | Quality | Concepts | Graph | Arguments |
|---|---|---|---|---|---|---|---|---|---|---|
| DRT | 1993 | PMB 4.0 | 2021 | 16,712 | 3 | 6,000 | Silver | WordNet | English | Semantic |
| PDT | 2003 | PDT 3.0 | 2013 | 49,431 | 1 | 49,431 | Silver | – | English | Syntactic |
| UCCA | 2013 | Wikipedia | 2020 | 7,934 | – | – | – | – | English | Syntactic |
| AMR | 2013 | AMR 3.0 | 2020 | 59,255 | – | – | – | PropBank | English | Semantic |
| UMR | 2021 | UMR 1.0 | 2023 | 2,186 | 5 | 1,993 | Gold | PropBank | Non-specific | Semantic |
| BMR | 2022 | BMR 1.0 | 2022 | 59,255 | 5 | – | Silver | BabelNet | English | Semantic |

Table 1: Main SP Formalisms. Columns: Formalism, year, large corpus available, corpus released to date, corpus size (sentences), number of non-English languages, the total number of annotations in non-English languages, quality, repositories of meaning, the language of the annotation (English/ non-specific), level of the annotation.

Over the years, multiple attempts have been proposed to encode other languages: from adaptations based on language-specific repositories – e.g., Spanish AMR (Migueles-Abraira et al., 2018) with AnCoraNet (Aparicio et al., 2008) or the Chinese AMR (Li et al., 2019) with Chinese Propbank (Xue et al., 2005) – to interlingua formalisms, such as the Uniform Meaning Representation (Gysel et al., 2021, UMR) or the BabelNet Meaning Representation (Navigli et al., 2022, BMR). However, the main challenge persists: the scarcity of annotated data keeps these formalisms in the realm of abstraction, since annotated data is crucial for training parsers. This scarcity is primarily attributable to the high costs and complexity of manual annotation. Annotators must thoroughly understand each specific language's rules, the formalism and the meaning repository, making the process highly labour-intensive. Moreover, even though some projects have generated annotated data for these formalisms – like AMR 3.0 (Knight et al., 2020), BMR 1.0 (Martínez Lorenzo et al., 2022), or UMR (Bonn et al., 2023b) – these datasets are English-oriented and hindered by paywalls or restrictive licenses. Furthermore, even though there have been efforts to annotate non-English languages – like UMR 1.0 (Bonn et al., 2023a), Turkish (Azin and Eryiğit, 2019), Persian (Tohidi et al., 2024), Portuguese (Sobrevilla Cabezudo and Pardo, 2019), or Vietnamese (Linh and Nguyen, 2019) – they have not produced more than 200 examples per language, which is inadequate for effective parser training. This situation underscores a critical bottleneck in employing these formalisms for wider linguistic applications.

This paper introduces a practical solution for tackling data scarcity, which can scale across different languages. Firstly, we propose the **Multilingual Semantic Layer (MSL)**, an evolution of previous formalisms, which acts as the necessary tool to enable data generation across languages. Secondly, by leveraging this novel representation, we develop a methodology for semi-automatically generating a vast amount of high-quality annotations across languages by making use of LLMs (under an academic budget). Lastly, we have annotated a gold standard dataset of 1,100 graphs across eleven languages, including low-resource ones, allowing us to assess the quality of our dataset.

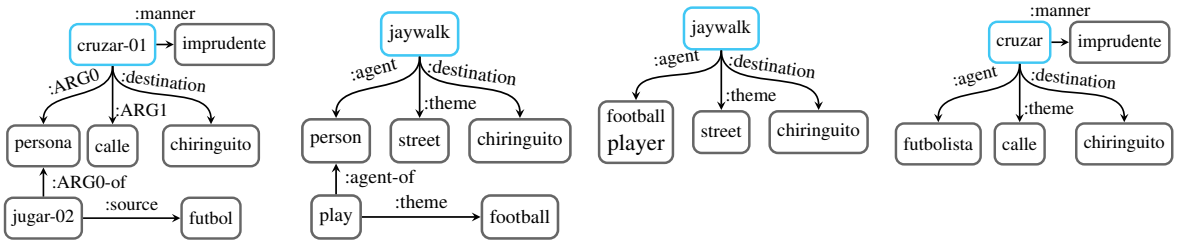
## 2 Related Work

Even if multiple formalisms have been proposed over the years, just a few of these provide annotated data. Table 1 highlights their main features of the largest dataset annotated for each formalism.

**Discourse Representation Theory (Kamp and Reyle, 1993, DRT)** provides a framework for representing the meaning of entire texts at the discourse level, which can be directly converted into logical forms. The Parallel Meaning Bank (Bos, 2013, PMB) is the most recent dataset under the DRT formalism, which incorporates word senses from WordNet (Miller, 1992), and semantic roles from VerbNet (Brown et al., 2019). However, the multilingual annotations use the English graph as the interlingua by automatically projecting the English annotations to other languages. Moreover, the number of annotations is insufficient.

**Prague Dependency Treebank (Hajič, 1998, PDT)** is a multi-layer formalism that defines a syntactic analysis for the sentence. The Tectogrammatical layer (Zeman and Hajic, 2020, PTG) covers semantic distinctions, such as the predicate-argument structures, word senses or co-references. The corpora are in Czech and English and rely on language-specific repositories.

**Universal Conceptual Cognitive Annotation (Abend and Rappoport, 2013, UCCA)** has a foundational layer focusing on the predicate-argument structure. However, it does not abstract




(a) Spanish AMR (b) Spanish PMB (c) BMR (d) MSL

Figure 2: Comparison across multilingual formalisms for the sentence in Figure 1 translated to Spanish. *"El futbolista cruza la calle de manera imprudente hacia el chiringuito."*.

away from word order and from the actual words.

**Abstract Meaning Representation (Banarescu et al., 2013, AMR)** provides a theoretical framework that does abstracts away from the syntax level. Some adaptations exist in other languages, but no large corpora are annotated.

**Uniform Meaning Representation (Bonn et al., 2023a, UMR)** is a semantic representation based on AMR that expands the semantic range of AMR at the sentence level by adding aspect and scope, while also expanding AMR to a document-level representation by incorporating co-reference, temporal and modal dependencies through documents. UMR also accommodates the formalism to other languages, specifically low-resource ones, by incorporating new explicit argument relations that mitigate the lack of predicate repositories. However, no large annotated datasets are available to train LMs; there are only 2,186 pairs of sentence graphs across six languages (Arapaho, Chinese, Cocama-Cocamilla, English, Navajo and Sanapaná).

**BabelNet Meaning Representation (Navigli et al., 2022, BMR)** extends AMR and aims to be an interlingua representation by using BabelNet concepts (Navigli and Ponzetto, 2010) and VerbAtlas frames (Di Fabio et al., 2019). However, the cross-lingual corpus is English-oriented, and BabelNet is not open-source (extended in Appendix A).

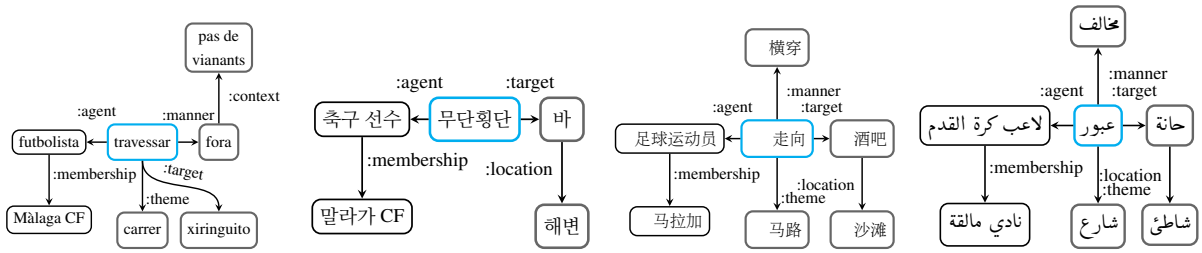
Even though certain formalisms have been developed to encode English semantics, their application to other languages often remains theoretical. These formalisms depend on external resources (e.g., WordNet, PropBank, or others), most of which are only available in English. While initiatives like Spanish AMR with AncoraNet or UMR 1.0 for low-resource languages attempt to bridge this gap, their annotations are limited (not exceeding 2,000 graphs), rendering them impractical for widespread use and requiring trained annotators on both the formalism and the external inventories. Moreover, even though multilingual formalisms like PMB and BMR exist, they essentially annotate English sentences and employ the English graph as the representation of parallel automatic translations. This leads to language divergences, as demonstrated in Figure 2 where *"jaywalk"* requires two nodes for its non-literal Spanish translation, underscoring the discrepancies and challenges in cross-lingual semantic representation.

## 3 Our Contributions

In this section, we overcome the data scarcity in SP by providing the novel Multilingual Semantic Layer (MSL) – with the necessary tools to encode the semantic relations between concepts across languages (Section 3.1) – and a multilingual dataset with millions of high-quality annotations, including manual annotations, that enables SP in different languages (Section 3.2).

### 3.1 MSL

We introduce MSL as a means for decoupling SP from the meaning behind concepts, focusing on extracting semantic relations between concepts in the sentence. MSL is built on top of previous graph-based formalism theories (such as AMR, BMR or UMR), and it employs a directed acyclic graph with nodes representing concepts and edges denoting semantic relations. However, unlike previous formalisms, MSL leverages solely explicit semantic relations without relying on external repositories (e.g., PropBank or VerbAtlas), and avoids using the English lexicon and English structures as an interlingua representation. MSL moves away from verbal-predicate-oriented representations (e.g., AMR or UMR) where nominal structures and predicates are represented with verbal




(a) Catalan MSL. (b) Korean MSL. (c) Chinese MSL. (d) Arabic MSL.

Figure 3: MSL graphs of the parallel sentences, showcasing the language divergences.
Catalan sentence: *"El futbolista del Màlaga CF va travessar el carrer fora del pas de vianants cap al xiringuito."*
Korean sentence: 말라가 CF 축구 선수가 해변 바를 향해 무단횡단하고 있습니다.
Chinese sentence: 马拉加足球运动员横穿马路向沙滩酒吧走去 。
Arabic sentence: لاعب كرة قدم في نادي مالقة يعبر الشارع عبور مخالف باتجاه حانة الشاطئ

predicates. Following BMR, MSL leverages explicit semantic relations that are consistent across languages, making them understandable without any repository and simplifying the data generation for non-English languages. For example, we establish new nominal relations – such as *:similar*, *:compared*, or *:related* – to enable nominal predicate representations without a repository.[1] Furthermore, MSL leverages the sentence's native lexicon and structures, simplifying data generation across languages by eliminating the necessity for concept disambiguation in each language. Therefore, unlike AMR or UMR, there is a direct correspondence between graph nodes and sentence spans. For instance, following BMR, the concept *"the football player"* is represented as it occurs in the sentence, not with two nodes, *"person"* and *"football"*, related by *"play"* (Figure 2).

MSL does not attempt to be a unified meaning representation across languages as BMR does, but aims to provide the necessary tools to enable parsing across languages. This is because the semantic structure of parallel sentences across languages might be different, so we cannot use only one of these representations as unification as BMR does (see Figure 3). MSL is an abstraction layer that offers a vanilla representation – not tied to any external repository – in more languages than any previous formalism. As a result, MSL does not provide the complete meaning of text, since it does not link words with their meaning. However, our intention is not to devise MSL as a new formalism, but rather as a semantic layer which can easily be integrated with other NLU layers, such as Word Sense Disambiguation (WSD), Entity Linking (EL), and Entity Typing (ET), which connect the concepts to external knowledge bases (e.g., Wikipedia, WordNet, BabelNet, etc.). Therefore, our goal with MSL is to provide not only a readily available representation and multilingual open-access dataset, but also a stepping stone towards a more flexible semantic representation upon which to build narrower or more specialized representations, improving the broader research community's utilization of Semantic Parsing, even in low-resource languages. Figure 3 illustrates an example of an MSL graph in different languages.

[1]Appendix Section B.2 explains all the relations.

### 3.2 MSL Dataset

In this section, we discuss the process of creating a corpus for MSL. First, we explain how we semi-automatically produced a preliminary silver annotation ($\text{MSL}_{AMR}$). Second, how we used this data to train a model and predict annotations over a vast amount of data across languages ($\text{MSL}_{silver}$). Third, how our annotators corrected the predicted graphs, and we projected these corrections to our corpus using LLMs ($\text{MSL}_{HQ}$). Fourth, how we used LLMs and parallel corpora to project our annotations to other languages ($\text{MSL}_{HQE}$). Finally, how we created 1,100 pairs of sentence graphs from scratch across ten languages, all within an academic budget.

**Silver Dataset Creation** Our initial goal was to obtain annotations of sentences in various languages. We began with the AMR 3.0 corpus – the largest AMR corpus, containing 59,255 English sentence-graph pairs – to generate our preliminary silver dataset, $\text{MSL}_{AMR}$ in English, French, German, Italian and Spanish. Firstly, ❶ we translated the English sentences into French, German,

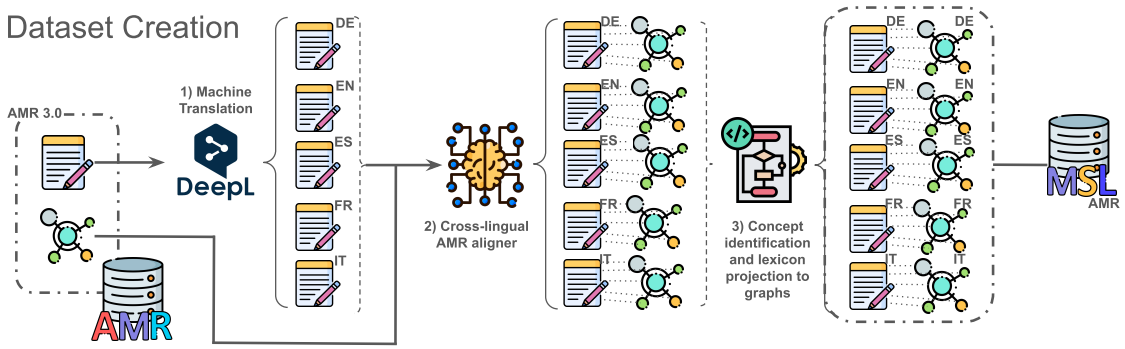



Figure 4: Description of the creation process for our multilingual version of the AMR dataset in the MSL layer. We (1) translate each sentence sample to 5 languages using DeepL, (2) align each English AMR graph to the sentence in each language (3) apply the MSL-specific modifications, producing language-specific graphs per sentence.

Italian and Spanish using DeepL (DeepL GmbH, 2024). Then, ❷ in order to align each language's sentence concepts with their respective English graph nodes, we employed the method proposed by Martínez Lorenzo et al. (2023), training a cross-lingual AMR parser using the DeepL translations and English graphs, and leveraging the model's cross-attention to extract alignments. We transitioned from PropBank argument relations to our cross-lingual and non-frame-dependent relations through manual mapping, integrating MSL's conceptual alterations with AMR. This process involved creating mappings and specific heuristics for modifying nominal structures.[2] MSL has a direct correspondence between sentence concepts and graph nodes, unlike AMR or UMR. Therefore, we have to identify multi-word and idiomatic expressions in the sentence, and – leveraging the previously extracted alignment – collapse their related graph nodes to represent them as single concepts. However, linguistic divergences often lead to variations in conceptual representation across languages. For example, Figure 2 shows how *"jaywalk"* is translated into *"cruzar de manera imprudente"* in Spanish, necessitating different graph representations. To address this, we devised a heuristic based on the alignment, the graph structure and a set of multi-word expressions in each language, that was capable of splitting or merging graph nodes when required.[3] These modifications comprise the last step ❸ in creating $\text{MSL}_{AMR}$.

**Generating the Large Corpus $\text{MSL}_{silver}$** Our $\text{MSL}_{AMR}$ is still under the AMR's license and uses silver translations. Consequently, our aim is to annotate a substantial corpus using non-AMR licenced data, $\text{MSL}_{silver}$. We trained a seq2seq SP parser in a multilingual fashion using mT5-large (Xue et al., 2021) with $\text{MSL}_{AMR}$.[4] Subsequently, ❹ we generated $\text{MSL}_{silver}$ in German, English, Spanish, French, and Italian gold sentences by predicting from two sources: i) a parallel corpus from various OPUS datasets (Tiedemann, 2009) such as TED, OpenSubtitles, Ubuntu, Bible, and Books; and ii) a non-parallel corpus, from Wikipedia.

**Manual Validation** ❺ At this stage, we have millions of multilingual annotations that are free from licensing restrictions, albeit still silver-generated. To enhance their quality, we engaged proficient annotators in AMR and trained in MSL to manually validate a subset of the graphs in each language.[5] Throughout this process, these annotators rectified numerous errors within $\text{MSL}_{silver}$, integrated previously absent elements from the original AMR corpus (such as tense, aspect, mood, and modality for verbs), and revised structural representations. This process ended up correcting around 2,000 graphs, 400 per language.[6]

**Replicating Changes** ❻ After rectifying errors and updating structures in the manual validation phase, we had to replicate all our corrections and changes across the five languages to the rest of $\text{MSL}_{silver}$. Previous studies have demonstrated the efficacy of LLMs as zero-shot annotators (Zhang et al., 2023), offering human-comparable labelling at lower costs (Gilardi et al., 2023; Zhu et al., 2023). Leveraging this and given the task's sim-

[2] Appendix B provides all the details.
[3] Appendix C explains all details of this process.
[4] See Appendix G for more details.
[5] Each annotator was a native-speaker of their language.
[6] Appendix D explains this process and the annotators guidelines in detail.

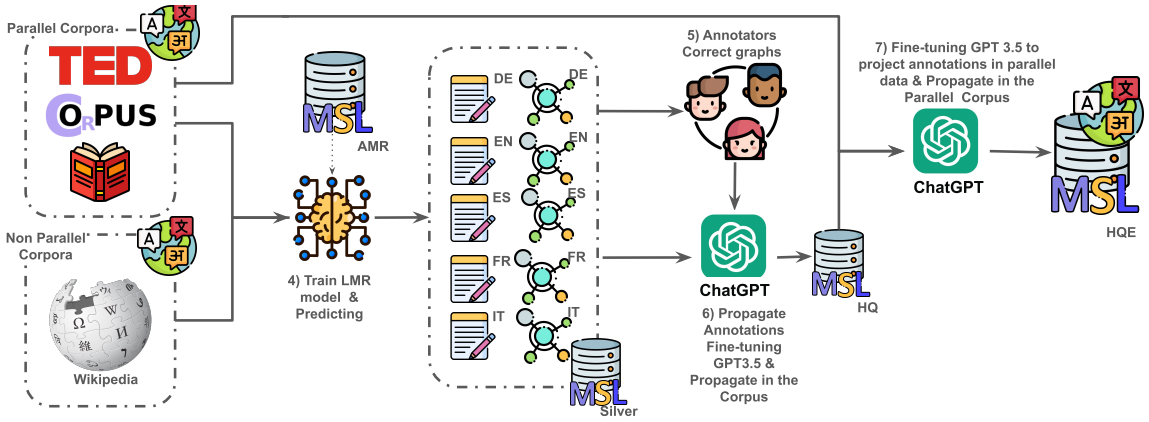



Figure 5: Pipeline showing how to create a license-free automatically annotated MSL dataset and how to expand it to new languages. From the creation of $\text{MSL}_{AMR}$ we (4) train a parser in the 5 available languages and use it to parse license-free text from diverse sources, creating $\text{MSL}_{Silver}$ (5) manually revise a subset of those predictions with expert annotators (6) propagate those corrections by teaching an LLM to apply the same fixes and create $\text{MSL}_{HQ}$ (7) train another LLM to create a sentence and its graph for a target language when given a pair in a source language from $\text{MSL}_{HQ}$, and use this pair to expand it to new languages and create $\text{MSL}_{HQE}$.

plicity compared to parsing from scratch, we used GPT-3.5 to apply our corrections corpus-wide. The model was tasked with generating accurate graphs given as input sentences and their corresponding silver graphs. Initially, a few-shot approach with example prompts proved ineffective due to inconsistency with our MSL structure and high operational costs due to long prompts. Consequently, we fine-tuned the model with our comprehensive set of corrections in German, English, Spanish, Italian and French. The trained model is given the original sentence and silver graph before corrections, and tasked to output the corrected graph. We validated this approach on a corrections subset, assessing performance using perplexity.[7] With the model fine-tuned, we efficiently replicated all corrections across the corpus, producing the $\text{MSL}_{HQ}$ in German, English, Spanish, Italian and French. The effectiveness of this approach will be further validated by our ablation experiments.

**Projecting to Other Languages** (7) Having developed a high-quality dataset in five languages, we address the data scarcity in other languages by generating high-quality annotations in Arabic, Catalan, Korean, Galician, Portuguese and Chinese (Table 2 shows stats). Under the premise that projecting annotations across parallel languages is simpler than starting from scratch (Barba et al., 2020; Daza and Frank, 2020; Blloshmi et al., 2020), we leverage our dataset's parallel nature to project them. As depicted in Figure 3, language divergences across translations for the same sentence should be reflected in the graph. Therefore, we again fine-tune GPT-3.5 using a subset of $\text{MSL}_{HQ}$ that has parallel sentences in German, English, Spanish, Italian and French. Given a sentence, its semantic graph and the parallel sentence in another language, the model has to generate the corresponding graph. We use English as the source language with Arabic, Chinese and Korean as the targets during inference. Then, Spanish served as a pivotal source language for Catalan, Galician, and Portuguese, due to their linguistic proximity.[8] The total cost for training and inference to apply our corrections and project to other languages was around 400$, 0.0013$ per graph. For comparison, the license to use AMR 3.0 alone is 300$, 0.005 per graph.

### 3.3 Manually Annotated Test Set

To evaluate our dataset's quality and establish a gold standard benchmark, we tasked annotators to manually annotate 100 parallel sentences from OPUS in Arabic, Catalan, Chinese, English, French, Galician, German, Italian, Korean, Portuguese, and Spanish.[9] Recognizing the high cost of manual annotation, we reserved this task exclusively for benchmark creation, ensuring our annotators were proficient in generating MSL graphs for all the listed languages.[10] Moreover, we could

[7] Appendix E shows examples and performance details.

[8] Appendix F shows some data examples and more detail.

[9] The annotators are native speakers in each language.

[10] Appendix D explains the annotation guidelines.

| Dataset | Arabic | Catalan | German | English | Spanish | Korean | French | Galician | Italian | Portuguese | Chinese | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $AMR_{3.0}$ | – | – | – | 59,255 | – | – | - | – | - | – | – | 59,255 |
| $MSL_{AMR}$ | – | – | 59,255 | 59,255 | 59,255 | – | 59,255 | – | 59,255 | – | – | 296,275 |
| $MSL_{Silver}$ | – | – | 2,574,529 | 3,029,254 | 1,115,822 | – | 1,711,337 | – | 1,603,467 | – | – | 12,655,303 |
| $MSL_{HQ}$ | – | – | 23,951 | 38,505 | 47,665 | – | 26,784 | – | 31.410 | – | – | 168,315 |
| $MSL_{HQE}$ | 17,529 | 17,550 | 23,951 | 38,505 | 47,665 | 7,826 | 17,531 | 20,000 | 31.410 | 17,551 | 17,550 | 257,068 |

Table 2: Number of annotated Graphs per language in the different datasets.

calculate the inter-annotator agreement in Spanish – 92.34 in SMATCH – as it was the only language with native speaker overlap in the graphs (Galician and Catalan annotators are also native Spanish speakers). We also reserved 1700 sentences per language from OPUS to test our systems for back-translation, and these sentences are not present in any of our training sets.

## 4 Experiments

### 4.1 Experimental Setup

**Tasks** Our evaluation framework consists of two parts: i) **Parsing** and **Generation** to ablate each annotation step, and ii) **Back Translation Experiment** to compare MSL against AMR and BMR.[11] The objective of **Parsing** – transforming the text into a graph representation according to a specific formalism – is to assess the complexity of generating MSL graphs due to the nuanced characteristics of MSL. Then, the **Generation** task – producing the original text from the graph representation – has the goal of appraising the effectiveness of MSL for preserving information. Finally, we perform a **Back Translation Experiment** – first parse and then generation – to compare the suitability of SP datasets, such as AMR 3.0, BMR 1.0, and MSL dataset, for training semantic parsers across languages.

**Datasets** For the Parsing and Generation study, we use the four training sets described in the paper ($MSL_{AMR}$, $MSL_{Silver}$, $MSL_{HQ}$ and $MSL_{HQE}$), for the test we use the gold test set annotated in Section 3.3, that comprises 1,100 sentence graph pairs, 100 sentences in each language (Arabic, Catalan, Chinese, English, French, Galician, German, Italian, Korean, Portuguese, Spanish). Then, for back-translation, we use our $MSL_{HQE}$ and, to train the other models, we use AMR 3.0 (Knight et al., 2020)[12] and the BMR 1.0 (Martínez Lorenzo et al., 2022). However, since AMR 3.0 provides only English sentences, we use the DeepL translations from ❶. As test data we use the parallel sentences from Abstract Meaning Representation 2.0 - Four Translations,[13] that translated the AMR 3.0 test set into German, Italian and Spanish (consisting of $1,371$ sentence in each language). Furthermore, we use parallel sentences from the OPUS corpus to test the performance in non-AMR data (Out-of-domain), which was not included in MSL. To maintain a consistent number of training samples across datasets, we reduce $MSL_{Silver}$ to $MSL^{HQ}_{Silver}$ by selecting only those sentence graphs that match the sentences in $MSL_{HQ}$ (168,315).

**Models** For training the parsers across formalisms and languages, we leverage CLAP (Martinez Lorenzo and Navigli, 2024), where the parsing task is framed as a seq2seq task for an Encoder-Decoder system, where the model is trained to generate a linearized version of a graph from a sentence and, vice versa, produce a sentence from a linearized graph. We adapt the model's vocabulary to include our set of relations and use mT5-large as the underlying language model, which supports all our languages.[14] First, given the datasets described in Section 4.1, we train a model for each dataset, AMR 3.0 and BMR 1.0 for each language (DE, EN, ES, FR, and IT). Then, for MSL we train four models on, respectively: i) $MSL_{AMR}$ (DE, EN, ES, FR, and IT); ii) $MSL^{HQ}_{Silver}$ (DE, EN, ES, FR, and IT); iii) $MSL_{HQ}$ (DE, EN, ES, FR, and IT); and iv) $MSL_{HQE}$ (AR, CA, DE, EN, ES, KO, FR, GL, IT, PT and ZH).

**Evaluation Measures** We evaluate Generation and Back-translation by using the standard NLU metric BLEU (Papineni et al., 2002). For Parsing, we employ the SMATCH measure (Cai and Knight, 2013), which is the most famous metric for AMR parsing. The SMATCH calculates the maximum overlap between the predicted and reference graphs. The SMATCH also works with MSL, since it follows AMR theory.

[11] The only formalisms with available parsers to train.

[12] AMR 3.0 is licensed by LDC at https://catalog.ldc.upenn.edu/LDC2020T02

[13] https://catalog.ldc.upenn.edu/LDC2020T07

[14] Appendix G provides more details and hyperparameters.

| Parsing | Arabic | Catalan | German | English | Spanish | Korean | French | Galician | Italian | Portuguese | Chinese | AVG | SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $MSL_{AMR}$ | 19.40 | 38.37 | 48.91 | 54.28 | 49.73 | 26.46 | 49.01 | 42.44 | 46.52 | 40.73 | 19.57 | 39.58 | 11.81 |
| $MSL_{Silver}$ | 20.12 | 37.41 | 48.82 | 55.12 | 49.34 | 27.04 | 51.52 | 41.12 | 47.32 | 41.23 | 20.12 | 39.92 | 11.86 |
| $MSL_{HQ}$ | 19.19 | 56.29 | **67.21** | **71.98** | 71.89 | 35.01 | **72.31** | 57.54 | 71.47 | 58.37 | 30.02 | 55.57 | 18.13 |
| $MSL_{HQE}$ | **56.36** | **72.38** | 66.94 | 71.35 | **72.86** | **56.38** | 71.91 | **69.31** | **71.83** | **72.26** | **58.36** | **67.26** | 6.48 |
| Generation | Arabic | Catalan | German | English | Spanish | Korean | French | Galician | Italian | Portuguese | Chinese | AVG | SD |
| $MSL_{AMR}$ | 0.00 | 20.21 | 33.48 | 40.60 | 44.47 | 2.80 | 37.36 | 15.00 | 38.28 | 25.50 | 1.27 | 23.54 | 15.97 |
| $MSL_{Silver}$ | 0.00 | 18.20 | 32.47 | 44.24 | 48.02 | 1.08 | 39.82 | 16.04 | 41.23 | 25.89 | 0.98 | 24.36 | 17.41 |
| $MSL_{HQ}$ | 0.00 | 16.37 | 38.00 | 55.19 | 57.00 | 1.15 | 51.24 | 16.08 | **51.80** | 25.08 | 1.07 | 24.63 | 17.91 |
| $MSL_{HQE}$ | **30.84** | **57.30** | **40.46** | **57.26** | **57.23** | **22.62** | **51.62** | **54.31** | 51.31 | **49.57** | **48.48** | **47.36** | 10.93 |

Table 3: SMATCH score for Parsing (Top) and BLEU for Generation (bottom) per language for our ablation of annotation steps. Columns: Results per language, Average and Standard Deviation.

### 4.2 Results

**Parsing and Generation** In Table 3, we observe that $MSL_{Silver}$ achieves results comparable to $MSL_{AMR}$, underscoring the success of distilling $MSL_{AMR}$. $MSL_{HQ}$ obtains better results across all languages compared to $MSL_{silver}$ in both directions (parsing and generation), showcasing the benefits of cleaning the data thanks to our automatic corrections. Furthermore, the benefits of $MSL_{HQE}$ are clear; expanding the dataset with additional languages enables semantic parsing capabilities in these languages. Notably, $MSL_{HQE}$ demonstrates a much smaller variance compared to other models. For comparison, the current multilingual state-of-the-art AMR parser by Cai et al. (2021) has a 9 SMATCH points gap between English (83.9) and Spanish (75.9), German (73.1) or Italian (75.4), while in MSL the gap is significantly smaller – considering that SMATCH values are not comparable across formalisms.[15] The text generation results show even more dramatic improvements compared to Martínez Lorenzo et al. (2022): AMR or BMR never reached more than 37 BLEU on non-English languages, while English attained 50 BLEU points. This points to MSL being a more expressive semantic layer, which benefits text generation.

**Back Translation Experiment** In Table 4, we can observe the effectiveness of different parsers in preserving information when back-translating from and into the same language passing through the graph using a cross-lingual model. We have to clarify that Martínez Lorenzo et al. (2022) reported BLEU scores of 45.3 for AMR and 50.1 for BMR in English using a monolingual model, unlike our experiments, which use a shared cross-lingual model for all languages. We can observe how the model trained with the MSL dataset outperforms not only in the AMR benchmark against the model trained with this specific data by 19 and 15 BLEU points, respectively, but also maintains the same performance across the Out-of-Domain, where models trained in AMR 3.0 and BMR 1.0 drop around 10 points in average. Moreover, the gap in non-English languages is more pronounced, since AMR and BMR's dependence on an English graph structure as the interlingua introduces semantic parsing inaccuracies similar to those encountered in machine translation. This attests to the quality, diversity, and size of the MSL dataset across languages compared to the previous dataset. Additional metrics are presented in Appendix H.

### 4.3 A Case Study

To show the limitations of previous formalisms in encoding the nuances of each language, consider the next example from the TED parallel corpus:

English: *We don't really stop to think about a raindrop the size of an actual cat or dog when we hear 'it's raining cats and dogs', but as soon as I do, I realize that I'm quite certain the dog has to be a small one – a cocker spaniel, or a dachshund – and not a golden Lab or Newfoundland.*

Spanish: *No pensamos en gotas de lluvia del tamaño de un cántaro cuando escuchamos 'llueve a cantaros', pero al hacerlo, nos damos cuenta que el cántaro debe ser uno muy pequeño; un botijo, un tarro, y no ollas con asas laterales.*

This example highlights how formalisms like AMR and UMR will struggle with idiomatic expressions, often leading to literal interpretations (e.g., *rain-01 :theme (and :op1 (cat) :op2 (dog))*). Additionally, formalisms like BMR work under the premise that parallel sentences in different languages should have the same representations and

[15] Explained in Appendix I.

| | AMR | | | | | | Out Of Domain - ∩ | | | | | | Out Of Domain - ∪ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | DE | EN | ES | IT | AVG | SD | DE | EN | ES | IT | AVG | STD | AR | CA | KO | FR | GL | PT | ZH | AVG |
| AMR | 21.55 | 32.40 | 31.02 | 29.00 | 28.49 | 4.83 | 15.87 | 25.49 | 18.93 | 15.43 | 18.83 | 4.02 | 0.00 | 1.52 | 1.73 | 0.45 | 2.97 | 1.68 | 0.00 | 1.32 |
| BMR | 27.16 | 38.97 | 36.67 | 29.30 | 33.03 | 5.68 | 22.01 | 36.29 | 26.41 | 21.89 | 25.97 | 6.05 | 0.00 | 1.34 | 1.82 | 0.24 | 3.31 | 1.41 | 0.25 | 1.36 |
| MSL | **41.82** | **51.36** | **52.77** | **42.56** | **47.13** | 5.73 | **43.52** | **52.81** | **53.48** | **44.88** | 48.82 | 4.52 | **24.33** | **46.34** | **27.56** | **49.41** | **48.41** | **48.02** | **51.35** | 41.00 |

Table 4: BLEU result for the Back-translation experiment. Rows (Formalisms): AMR, BMR and MSL. Columns (Datasets per language): AMR 4 translations, Out of Domain, Average and Standard Deviation. ∩ means languages seen by all models at train time, ∪ rest of languages in MSL.

consequently fall short in practice. A major flaw in BMR 1.0 is its reliance on English as the interlingua graph, failing to capture nuanced semantics across non-English languages effectively. Although BabelNet links the English lemmatization of *"raining cats and dogs"* with the Spanish *"llueve a cántaros"*, the English sentence associates the phrase with specific dog breeds, while the Spanish refers to types of jars (*"botijo"*), leading to fundamentally different conceptual graphs even though the idioms are equivalent across languages. This underscores the limitations of using a single language-based representation as a universal interlingua for comprehensive multilingual semantic understanding. It reveals how language-specific divergences cause different languages to represent ideas or concepts uniquely – highlighting that *human languages are designed to model ideas and concepts, and not the other way around.*

This underscores our approach: (i) viewing MSL not as an interlingua but as a layer for semantic representation across languages, not focused on representing the same idea equally across languages, and (ii) refining graph structures using an LLM, acknowledging that parallel annotations can vary significantly, thus requiring more than just algorithmic adjustments as seen in the BMR 1.0 dataset.

## 5 Potential and Future of MSL

In the era of LLMs, we aimed to revamp the dream of structuring natural text across multiple languages through the use of MSL. This work opens up new possibilities for applying graph-based and big-data analyses to multilingual content. We envision the potential of MSL as follows:

- **NLP Task Integration:** MSL can be integrated with other resources like BabelNet to generate BMR-like graphs and with predicate-argument structures to yield AMR and UMR representations. Furthermore, MSL could be integrated with Entity Linking, Entity typing, or Word Sense Disambiguation, among others, to create a full semantic representation.

- **Addressing Data Scarcity:** The semantics of sentences remain consistent across formalisms, though they are represented differently. Consequently, the MSL dataset could be utilized for transfer learning. This involves pre-training a model on MSL and subsequently fine-tuning parsers for other formalisms. This approach reduces the need for extensive data development by allowing parsers to be fine-tuned using smaller, more specific datasets.

- **Linguistic Expansion:** With current LLMs, we can expand MSL to include more languages, thereby continuing to bridge the semantic parsing gap across different linguistic contexts.

- **Parsing Efficiency:** MSL enables the shift from encoder-decoder to encoder-focused architectures, thanks to the one-to-one correspondence between sentence concept and graphs nodes, reducing computational demands and accelerating processing speeds.

## 6 Conclusion

This paper addresses the challenge of data scarcity in multilingual SP. Given the complexity and cost of training annotators for large-scale tasks, we introduce i) the MSL layer and ii) provide a large, high-quality corpus across languages. The layer decouples SP from Word Sense Disambiguation and Entity Linking, focusing on the extraction of sentence-level semantic relations. Then, we automatically develop a preliminary silver dataset across languages, refining it through manual corrections and propagating these annotations using fine-tuned GPT LLMs. We also manually annotated 1,100 sentences in 11 languages, including low-resource ones, to demonstrate the flexibility of our layer and dataset over previous ones. Our code and dataset are available at https://github.com/SapienzaNLP/MSL.

## 7 Limitations

While we lower the cost burden of manually annotating training corpora for SP by leveraging LLMs, expanding to further languages would still require manual annotation to create new evaluation sets. Furthermore, while we show the effectiveness of relying on LM and LLMs to aid in annotation and propagating manual corrections, these still rely on access to either local computing or remote computing via third-party APIs. Nevertheless, as discussed in Section 3.1 our approach is still more affordable than previous SP annotation attempts. We default to external APIs for LLMs in only two steps of our dataset creation, a resource that we will release publicly, and therefore a step that does not require to be replicated. All Parsing and Generation models are trained with open-access models and therefore can be trained and used in inference by anyone with enough computing to do so. Therefore, this may pose a limitation if there are changes to OpenAI API and someone wants to replicate or expand to new languages adopting the same approach. We were limited to single-GPU experiments with 24GB of memory, so we opted for an external API rather than compromising quality with smaller models. However, we are confident a similar approach could be achieved using locally available LLMs like LLaMA (Touvron et al., 2023) or Mistral (Jiang et al., 2023).

## Acknowledgments

The authors gratefully acknowledge the support of:

- CREATIVE project (CRoss-modal understanding and gEnerATIon of Visual and tExtual content) funded by the MUR Progetti di Ricerca di Rilevante Interesse Nazionale programme (PRIN 2020).
- FAIR project (Future Artificial Intelligence Research), the PNRR MUR project PE0000013-FAIR.
- Marie Skłodowska-Curie project *Knowledge Graphs at Scale* (KnowGraphs) No. 860801 under the European Union's Horizon 2020 research and innovation programme.
- French Agence Nationale pour la Recherche, through the SELEXINI project (ANR-21-CE23-0033-01).
- DrEAM (Doctor, Explore and Achieve More) mobility program by the Lorraine Université d'Excellence (LUE).

**Special Abelardo Carlos Acknowledgments**

As I conclude my doctoral journey, I am immensely grateful to everyone who has supported me throughout this period. Foremost, I extend my deepest appreciation to my family: my parents, José Abelardo Martínez Férriz and Matilde Lorenzo López; my sister, Miriam Martínez Lorenzo; and my relatives, Irene Pérez Lorenzo, Raquel Pérez Lorenzo, Maricarmen Lorenzo López, Matilde López Aguilar, Manuel Lorenzo Castilla, Rafael Ortega Ríos, María Consuelo Martínez Férriz, Rafa Ortega Martínez, and Consuelo Férriz.

I am also profoundly thankful to the Sapienza NLP Group for their invaluable guidance and support, with a special mention to Pere-Lluís Huget Cabot for his help and friendship.

My gratitude extends to all my friends who supported me during this chapter of my life—especially those I met in this period, DesBariaos and Napoleone (Matheus, Nico, Gonzalo, Eva, Alain, Ivan and Anna), who have been pillars of support throughout my PhD journey.

Lastly, I would like to give a special acknowledgement to Marcela Soares Reed for her enduring support and encouragement in this process.

## References

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Juan Aparicio, Mariona Taulé, and M. Antònia Martí. 2008. AnCora-verb: A lexical resource for the semantic annotation of corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Zahra Azin and Gülşen Eryiğit. 2019. Towards Turkish Abstract Meaning Representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 43–47, Florence, Italy. Association for Computational Linguistics.

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.

Edoardo Barba, Luigi Procopio, Niccolò Campolungo, Tommaso Pasini, and Roberto Navigli. 2020. Mulan: Multilingual label propagation for word sense disambiguation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3837–3844. International Joint Conferences on Artificial Intelligence Organization. Main track.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12564–12573.

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, Online. Association for Computational Linguistics.

Claire Bonial, Lucia Donatelli, Stephanie M. Lukin, Stephen Tratz, Ron Artstein, David Traum, and Clare Voss. 2019. Augmenting Abstract Meaning Representation for human-robot dialogue. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 199–210, Florence, Italy. Association for Computational Linguistics.

Julia Bonn, Chen Ching-wen, James Andrew Cowell, William Croft, Lukas Denk, Jan Hajič, Kenneth Lai, Martha Palmer, Alexis Palmer, James Pustejovsky, Haibo Sun, Rosa Vallejos Yopán, Jens Van Gysel, Meagan Vigus, Nianwen Xue, and Jin Zhao. 2023a. Uniform meaning representation. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Julia Bonn, Skatje Myers, Jens E. L. Van Gysel, Lukas Denk, Meagan Vigus, Jin Zhao, Andrew Cowell, William Croft, Jan Hajič, James H. Martin, Alexis Palmer, Martha Palmer, James Pustejovsky, Zdenka Urešová, Rosa Vallejos, and Nianwen Xue. 2023b. Mapping AMR to UMR: Resources for adapting existing corpora for cross-lingual compatibility. In *Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023)*, pages 74–95, Washington, D.C. Association for Computational Linguistics.

Johan Bos. 2013. The Groningen meaning bank. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*, page 2, Trento, Italy.

Houda Bouamor, Hanan Alshikhabobakr, Behrang Mohit, and Kemal Oflazer. 2014. A human judgement corpus and a metric for Arabic MT evaluation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 207–213, Doha, Qatar. Association for Computational Linguistics.

Susan Windisch Brown, Julia Bonn, James Gung, Annie Zaenen, James Pustejovsky, and Martha Palmer. 2019. VerbNet representations: Subevent semantics for transfer verbs. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 154–163, Florence, Italy. Association for Computational Linguistics.

Deng Cai, Xin Li, Jackie Chun-Sing Ho, Lidong Bing, and Wai Lam. 2021. Multilingual AMR parsing with noisy knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2778–2789, Punta Cana, Dominican Republic. 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.

Angel Daza and Anette Frank. 2020. X-SRL: A parallel cross-lingual semantic role labeling dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3904–3914, Online. Association for Computational Linguistics.

DeepL GmbH. 2024. Deepl translator. `https://www.deepl.com/translator`.

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

Longxu Dou, Yan Gao, Xuqi Liu, Mingyang Pan, Dingzirui Wang, Wanxiang Che, Dechen Zhan, Min-Yen Kan, and Jian-Guang Lou. 2022. Towards knowledge-intensive text-to-SQL semantic parsing with formulaic knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5240–5253, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Jens Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Timothy J. O'Gorman, Andrew Cowell, W. Bruce Croft, Chu-Ren Huang,

Jan Hajic, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a uniform meaning representation for natural language processing. *KI - Künstliche Intelligenz*, 35:343 – 360.

Jan Hajič. 1998. Building a syntactically annotated corpus: The prague dependency treebank. *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, pages 106–132.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, volume 42 of *Studies in Linguistics and Philosophy*. Springer, Dordrecht.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Kevin Knight, Bianca Badarau, Laura Baranescu, Claire Bonial, Madalina Bardocz, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Martha Palmer, Tim O'Gorman, et al. 2020. Abstract Meaning Representation (AMR) Annotation Release 3.0 (LDC2020T02). Philadelphia. Linguistic Data Consortium.

Valentina Leone, Giovanni Siragusa, Luigi Di Caro, and Roberto Navigli. 2020. Building semantic grams of human knowledge. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2991–3000, Marseille, France. European Language Resources Association.

Bin Li, Yuan Wen, Li Song, Weiguang Qu, and Nianwen Xue. 2019. Building a Chinese AMR bank with concept and relation alignments. *Linguistic Issues in Language Technology*, 18.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ha Linh and Huyen Nguyen. 2019. A case study on meaning representation for Vietnamese. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 148–153, Florence, Italy. Association for Computational Linguistics.

Abelardo Carlos Martínez Lorenzo, Pere Lluís Huguet Cabot, and Roberto Navigli. 2023. Cross-lingual AMR aligner: Paying attention to cross-attention. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1726–1742, Toronto, Canada. Association for Computational Linguistics.

Abelardo Carlos Martínez Lorenzo, Marco Maru, and Roberto Navigli. 2022. Fully-Semantic Parsing and Generation: the BabelNet Meaning Representation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1727–1741, Dublin, Ireland. 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, Turin, Italy. ELRA and ICCL.

Noelia Migueles-Abraira, Rodrigo Agerri, and Arantza Diaz de Ilarraza. 2018. Annotating Abstract Meaning Representations for Spanish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Roberto Navigli, Rexhina Blloshmi, and Abelardo Carlos Martinez Lorenzo. 2022. BabelNet Meaning Representation: A Fully Semantic Formalism to Overcome Language Barriers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. 2016. Coreference in Prague Czech-English Dependency Treebank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 169–176, Portorož, Slovenia. European Language Resources Association (ELRA).

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Subendhu Rongali, Luca Soldaini, Emilio Monti, and Wael Hamza. 2020. Don't parse, generate! a sequence to sequence architecture for task-oriented semantic parsing. In *Proceedings of The Web Conference 2020*, WWW '20, page 2962–2968, New York, NY, USA. Association for Computing Machinery.

Marco Antonio Sobrevilla Cabezudo and Thiago Pardo. 2019. Towards a general Abstract Meaning Representation corpus for Brazilian Portuguese. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 236–244, Florence, Italy. Association for Computational Linguistics.

Jörg Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248.

Nasim Tohidi, Chitra Dadkhah, Reza Nouralizadeh Ganji, Ehsan Ghaffari Sadr, and Hoda Elmi. 2024. Pamr: Persian abstract meaning representation corpus. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Bailin Wang, Mirella Lapata, and Ivan Titov. 2021. Learning from executions for semantic parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2747–2759, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.

Daniel Zeman and Jan Hajic. 2020. FGD at MRP 2020: Prague tectogrammatical graphs. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 33–39, Online. Association for Computational Linguistics.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Huaiyu Zhu, Yunyao Li, and Laura Chiticariu. 2019. Towards universal semantic representation. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 177–181, Florence, Italy. Association for Computational Linguistics.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks.

## A BMR Issues

While BMR represents a significant advancement in cross-lingual representation, it encounters several challenges that limit its effectiveness and scope:

- **Dependency on BabelNet**: BMR relies on the BabelNet repository, which is not open-source. This reliance on a proprietary resource limits accessibility and modifiability for researchers and developers.
- **Limited scope of concepts and languages**: The number of concepts and languages covered by BabelNet is restricted. This limitation constrains the representational breadth of BMR, potentially leaving out diverse linguistic and cultural nuances.
- **English-centric development**: The corpus for BMR is derived automatically from the AMR (Abstract Meaning Representation) corpus, which is fundamentally English-centric. Moreover, the AMR corpus itself is not publicly available due to licensing restrictions, which further limits the adaptability and usability of BMR in diverse linguistic contexts.
- **Challenges with LM-based parsers**: Language Model (LM) based parsers used in BMR have limited capabilities in integrating new BabelNet concepts. To compensate for unavailable concepts, these parsers revert to using an English lexicon, which may not accurately represent meanings across different languages. abelNet repository, which is not open-source. This reliance on a proprietary resource limits accessibility and modifiability for researchers and developers.
- **Complex annotation process**: The annotation process for BMR is significantly more complex than that for AMR, posing a barrier for scalability and ease of use in linguistic research and applications.

## B MSL Relations

As noted earlier, AMR leverages coarse-grained frames and argument structures from the English PropBank within OntoNotes, a resource-limited to English. These frames denote predicate-specific semantic relations and are often unclear without a gloss. For instance, the subgraph in Figure 6

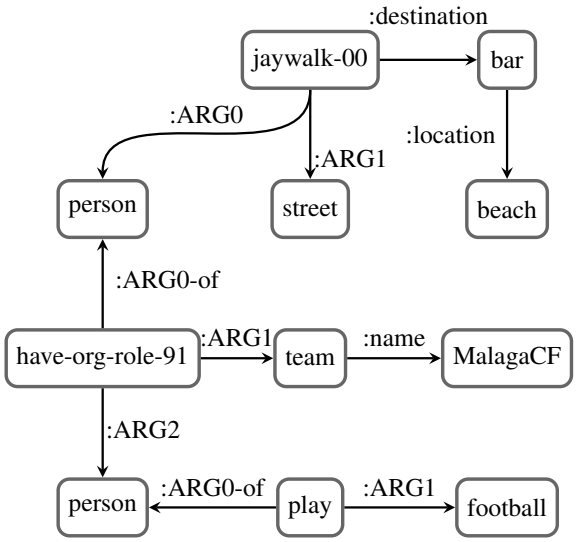



Figure 6: AMR of: *"The Malaga CF football player jaywalks across the street to the beach bar."*. The named-entity was simplified to enhance interpretability. In AMR they annotate a frame with -00 when it is not presented in Ontonotes.

representing "Malaga CF football player" centers around the frame `have-org-role-91`, with relations `:ARG0`, `:ARG1`, and `:ARG2` indicating the person, the organization (Malaga CF), and the role (football player), respectively, tying the argument's meaning directly to the predicate. However, language-specific repositories akin to PropBank, used for annotating non-English sentences, lack a precise one-to-one frame correspondence, complicating direct mapping. We introduce our explicit relations for universally applicable semantics in order to address language specificity.

For constructing the MSL dataset, a linguist[16] manually mapped PropBank arguments to our relations, replacing original AMR frames and roles (e.g., mapping `write-01`'s `ARG0` to `agent` and `ARG1` to `theme`). For non-verbal and special predicates in PropBank that AMR employs for unique semantic structures (like `have-org-role-91`), we mapped these to our new nominal structures, enhancing their argument representation (Section B.1 lists the MSL semantic roles, and Section B.2 provides mapping examples. Figure 7 illustrates the modified AMR graph from Figure 6.

### B.1 Semantic Roles

Our new set of relations (Table 5) is adapted to include non-verbal entities, drawing on property

[16]Annotators possess proficient English skills and were compensated according to their local standards.

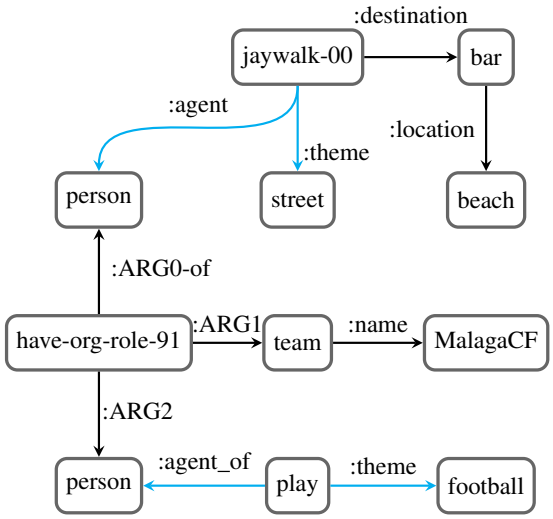



Figure 7: Graph moving from PropBank to our MSL set of relations.

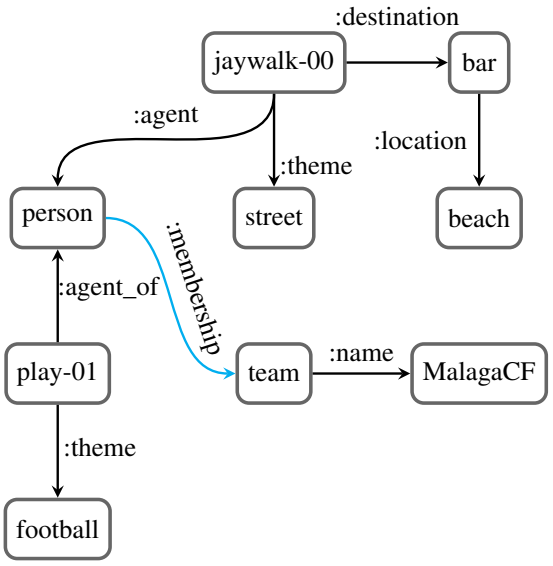



Figure 8: Graph moving from AMR non-predicate structures to our MSL set of relations.

lists from existing literature (Leone et al., 2020). Like AMR, each relation in our framework has an inverse, indicated by appending *_of* (e.g., *:purpose* becomes *:purpose_of*). Roles from AMR not covered in Table 5 are retained in MSL, including *:degree*, *:frequency*, and *:manner*. Figure 7 shows the AMR of Figure 6 after moving to our relations.

### B.2 Nominal Structures

Non-verbal predicates and special predicates found within AMR 3.0 have been mapped to the set of semantic relations described in Section B.1 by means of an in-house annotation interface. See Table 6 for an AMR 3.0 to MSL sample. Figure 8 shows the AMR of Figure 7 after adapting to our new nominal structures.

## C MSL Concept Representation

As mentioned, unlike AMR or UMR, which decompose concepts into multiple nodes using OntoNotes for semantic encoding (e.g., representing a "Football player" as a "person who plays football"), MSL employs single nodes for concept representation. We detail the transition from AMR 3.0 to our dataset, focusing on node merging in Section C.1. The language evolves to model the language, so each language models concepts differently, so rather than conforming to a single language's structure for knowledge representation across languages – such as BMR with the English structures to encode meanings universally (e.g., using a single node to represent the concept of "jaywalk" across languages, even when a direct equivalent does not exist) – MSL adapts to language divergences. When a concept in one language is expressed through multiple concepts in another, we divide the nodes accordingly. For details on this process, refer to Section C.2, where we explain our methodology for adapting AMR 3.0 to our dataset.

### C.1 MSL Merging Concepts

In AMR, single concepts and multiword expressions, including idioms, are often decomposed into multiple nodes through node composition. However, two main issues arise: firstly, single concepts are split to utilize external repositories for encoding relations. For instance, the expression "football player" in AMR 3.0 is depicted using three nodes as:

$$(p \ / \ person :ARG0-of \ (p2 \ / \ play \\ :ARG1 \ (f \ / \ football)))$$

This representation implies "a person associated with Malaga CF who plays football," where (i) the football concept is not directly linked to Malaga, and (ii) merely playing football does not equate to being a football player. MSL addresses these issues by merging such nodes into a single concept, "football player," directly associated with Malaga CF, thereby clarifying that (i) football is related to Malaga and (ii) the individual is a football player rather than just someone who plays football. Additionally, the reliance on external repositories makes

| MSL | AMR | Sentence | Examples |
|---|---|---|---|
| against | – | They play against Real Madrid | play$_V$>Real Madrid$_N$ |
| age | – | Perepli is 30 yo | be$_V$>30$_N$ |
| agent | – | he is studying | study$_V$>he$_N$ |
| attribute | mod | They found it necessary | find$_V$>necessary$_A$ |
| cause | cause | I arrived late since the traffic jam | traffic$_jam_N$>arriveV |
| compared | – | He is faster than a horse | faster$_A$>horse$_N$ |
| composition | consist-of | a 900-page book | book$_N$>page$_N$ |
| concession | – | I studied but I've failed | study$_V$ >fail$_V$ |
| contrast | – | I love fish, not meat | fish$_N$>meat$_N$ |
| context | location, topic | They are good in sport | they$_N$>sport$_N$ |
| coref | – | Malaga (...) the city | city$_N$>Malaga$_N$ |
| cost | cost | This cost 5 euros | cost$_N$>euro$_N$ |
| example | – | Imagine a dog, like a Bulldog | dog$_V$>Bulldog$_N$ |
| experiencer | – | I see you | see$_V$>you$_N$ |
| extent | duration, extent | He works during 5 days | work$_V$>day$N$ |
| identity | domain/meaning/role | Abelardo, my father is (...) | Abelardo$_N$>father$_N$ |
| instrument | instrument | I cook with the pan | cook$_V$>pan$_N$ |
| location | location | the pen in the table | pen$_N$>table$_N$ |
| membership | employed-by/have-org-role-91 | The president of the company | president$_N$>company$_N$ |
| part | part/subset/superset | The finger of his hand | finger$_N$>hand$_N$ |
| participant | – | the dinner with my parents | dinner$_N$>parent$_N$ |
| patient | – | Kick the ball | kick$_V$>ball$_N$ |
| possession | poss | I have a house | have$_V$>house$_N$ |
| purpose | purpose | I went there to study | go$_V$>study$_N$ |
| quality | mod | The red book | book$_N$>red$_A$ |
| quantity | quant | Four days | day$_N$>four$_n$ |
| related | have-rel-role-91 | the mother of the guy | mother$_N$>guy$_N$ |
| result | – | I've ended up going there | end$_up_V$>go$_V$ |
| product | – | He makes shoes | make$_V$>shoe$_N$ |
| scale | source | He got 8 out of 10 | 8$_N$>10$_N$ |
| similar | – | He acts like the leader | act$_V$>leader$_N$ |
| source | source | I got funds from the university | get$_V$>university$_N$ |
| target | beneficiary/destination/direction | I gave it to my students | give$_V$>student$_N$ |
| theme | – | I read the book | read$_V$>book$_N$ |
| time | time | I went yesterday | go$_V$>yesterday$_R$ |
| url | hyperlink-91 | The website https://time.is/ | website$_N$>https://time.is/ |

Table 5: Semantic roles in MSL. Left to right: MSL role names (MSL), AMR role(s) equivalent (AMR), example sentence and role usage example (s). Examples read as follows: father node$_{PoS}$>child node$_{PoS}$.

| |
|---|
| **have_org_role.91** |
| AMR: ARG0 (entity); ARG1 (organization), ARG2 (role) |
| MSL: ARG2 :membership ARG1 |
| AMR: person :ARG0-of (have_org_role.91 :ARG1 (university) :ARG2 (professor)) |
| MSL: professor :membership (university) |
| **be_located_at.91 (reification of :time)** |
| AMR: ARG1 (entity); ARG2 (location) |
| MSL: be :theme_of (ARG1) :location (ARG2) |
| Example: I am in the cinema |
| AMR: be_located_at.91 :ARG1 (I) :ARG2 (cinema) |
| MSL: be :theme (ARG1) :location (ARG2) |
| **good.02 (generally positive: morally good, pleasing)** |
| AMR: ARG1 (generally positive/pleasing entity); ARG2 (recipient/target of good behavior) |
| MSL: be :theme (ARG1) :attribute (good) :target (ARG2) |
| Example: My mother is good to me. |
| AMR: good-02 :ARG0 (mother) :ARG1 (I) |
| MSL: be :theme (mother) :attribute (good) :target (I) |

Table 6: Mapping examples from AMR 3.0 to MSL. Each row block lists (top to bottom) original OntoNotes predicate names and glosses, original glosses for the predicate arguments, and predicate rendering in MSL. Then, an example sentence with the AMR and MSL.

these relations non-deterministic, varying with the availability and nature of the repository. For example:

$$(r \ / \ footbal\_player)$$

Furthermore, AMR's approach to splitting multiword or idiomatic expressions into multiple nodes can lead to a literal, rather than meaningful, representation. For example, "rains cats and dogs" in AMR might be represented as:

$$(r \ / \ rain-01 \ :ARG2 \ (a \ / \ and \ :op1 \ (c \ / \ cat) :op2 \ (c \ / \ dog)))$$

This fails to capture the idiom's intended meaning. Like BMR, MSL consolidates such expressions into a single concept within the graph, ensuring a more accurate representation of idiomatic or complex expressions, exemplified by treating "football player from Malaga CF" as a unified concept. Figure 9 shows the AMR of Figure 8 after adapting to our new nomical structures.

This section details how we transformed AMR 3.0 into $\text{MSL}_{AMR}$ by merging nodes.

**Multiword Expression Identification** We begin by identifying single words or multiword expressions within sentences across English and its parallel translations in German, Spanish, French, and Italian, which are represented by multiple nodes in the AMR graph. We compile a list of multiword and idiomatic expressions from open websites in these languages. Using SpaCy 3.1, we lemmatize sentences in each language.

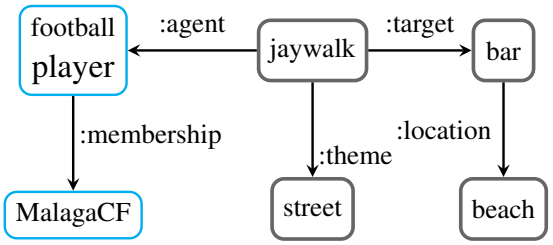



Figure 9: Graph moving from AMR non-predicate structures to our MSL set of relations.

**Manual Validation** Automatic detection of multiwords can lead to errors, such as incorrect node mergers and sense assignments. To address this, an expert linguist manually inspects all multiword instances identified in AMR 3.0, deciding on their retention, modification, or removal.

**Lemmas projections** After finalizing the multiword list and alignments, we segmented the sentence spans having clusters of single concepts, then we projected the lemmas of the cluster to their nodes and merged nodes in AMR graphs that refer to the same word or multiword expression. This bottom-up approach starts with the leaves, moving towards the root, to represent phrases like "football

player" or "rains cats and dogs" as single concepts.

### C.2 MSL Splitting Nodes

Unlike previous formalisms such as AMR, UMR, BMR, and PMB, our approach does not rely on English as an interlingua, thereby addressing language divergences more effectively. The challenge arises because existing annotations are primarily based on English, using the language to model knowledge rather than the reverse. Consequently, we cannot apply a single language's structure to represent knowledge across multiple languages. For instance, BMR encodes meanings using English structures, such as representing the concept of "jaywalking" with a single node, even though a single concept representation does not exist. In contrast, MSL utilizes parallel translations, splitting nodes when a concept is represented by multiple concepts in other languages, ensuring a more accurate cross-linguistic representation.

Starting with sentences segmented by concepts in each language, we identify instances where multiple clusters correspond to a single node and, where this is the case, proceed to split the nodes. Our initial step involves creating a heuristic that examines the part-of-speech (POS) tag of each element, alongside the arguments already utilized by the node, followed by a manual classification to determine the node encoding for such cases. For instance, the English term "jaywalk" can have several translations in Spanish, such as "cruzar la calle imprudentemente" or "cruzar la calle de manera imprudente". Here, the "jaywalk" node aligns with "cruzar" and either "imprudentemente" or "imprudente", representing two distinct clusters in Spanish. The node is split based on the POS tags (verb "cruzar" and adverb "imprudentemente"), considering existing relations like ":agent", ":theme", and ":destination". In cases where the main node is a verb and the secondary is an adverb or adjective, with these relations present, we use the ":manner" relation to specifically modify the verb action.

$$(c \ / \ cruzar \ :manner \ (i \ / \ imprudentemente))$$

Figure 2 shows an example of the final representation after doing all the heuristics. We acknowledge that while this heuristic works in many cases, it may introduce errors due to the inability to capture all linguistic nuances, which are addressed during the manual validation process.

## D Manual Validation

After employing our model to produce a substantial dataset with our parser trained on $\text{MSL}_{AMR}$, we acknowledge that these silver-generated graphs, despite being effective in many cases, may not fully capture the linguistic nuances across languages due to heuristic limitations in $\text{MSL}_{AMR}$'s creation. Consequently, we extracted a subset of these multilingual graphs for manual validation, aiming to refine them further before using LLMs to apply corrections across the larger dataset. This validation process focused on three key areas: i) refining the graph structure, ii) adding missing information from AMR, and iii) modifying specific encodings and introducing new relations based on insights from real examples.

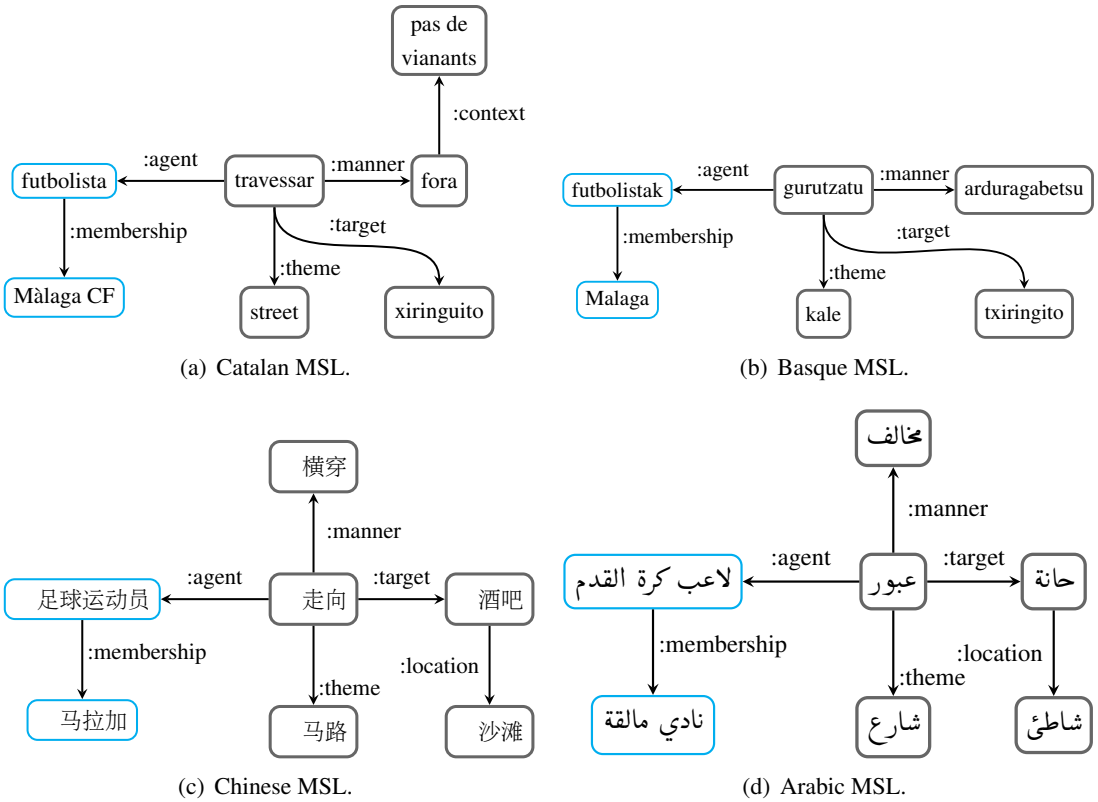
### D.1 Refining Graph Structure

Since we have silver-generated graphs created from a semi-automatically process (starting from a gold graph, we transform this), there are some errors in our silver data. Therefore, we analyzed a subportion of this data to find graphs with structural errors due to the fact that our model did not learn to properly represent this knowledge. Therefore, we ask proficient annotators to manually validate this subportion, restructuring the graph if necessary. The main errors found were i) the incorrect identification of multi-word expressions, ii) the use of incorrect relations, and iii) generating corrupted structures.

### D.2 Incorporating Missing Information

Although AMR effectively encodes textual information within its semantic structures, it overlooks essential word components like number, tense, aspect, and mood, crucial for fully grasping meaning—a gap highlighted in existing literature (Bonial et al., 2019). Recognizing the significance of these grammatical categories, we integrate these features to bolster the representational capability of our formalism.

During the annotation validation process, annotators across five languages are tasked with identifying and manually incorporating tense, aspect, number, and mood. Acknowledging the diverse mechanisms languages employ to convey this information, we provide tools for each language so as to facilitate encoding, with subsequent mappings between representations (as detailed in Table 8). For instance, we enhance verb nodes with a




(a) Catalan MSL. (b) Basque MSL.

(c) Chinese MSL. (d) Arabic MSL.

Figure 10: MSL graphs of the parallel sentences, showcasing the language divergences.
Catalan sentence: *"El futbolista del Màlaga CF va travessar el carrer fora del pas de vianants cap al xiringuito."*
Basque sentence: Malagako futbolistak kalea modu arduragabetsu batean gurutzatu du txiringitorantz
Chinese sentence: 马拉加足球运动员横穿马路向沙滩酒吧走去。
Arabic sentence: لاعب كرة قدم في نادي مالقة يعبر الشارع عبور مخالف باتجاه حانة الشاطئ

`:tense` role indicating $future$ or $past$ (defaulting to present for its absence) and an `:aspect` relation marking $imperfect$, $continuous$, $perfect$, or $perfectcontinuous$ aspects. The `:mood` relation captures the verb's mood—$subjunctive$, $conditional$, $imperative$, or by default, the indicative. Additionally, the plurality of nouns is indicated by adding a `:quantity` relation with a $+$ value for plurality.

### D.3 Modifing Encoding Structures

The simplicity of our layer, allows new forms to be easily evolved just by incorporating new relations. Since it is impossible to understand all the nuances of a language without going to the data, we allow the annotators to incorporate new annotations only when strictly necessary, such as when the previous forms do not fit properly or because it would allow easier use in the future.

**Verb To Be** In many formalisms, auxiliary verbs like "to be" and "to have" are not directly encoded but represented through relations (e.g., *:domain*in AMR or *:identity* in BMR). However, such annotation methods struggle to encapsulate negations or probabilities within the node itself (e.g., "he is not here"), leading to the creation of new nodes for situations necessitated by negations, among other reasons (e.g., the inclusion of nodes like be-located-at-91 in AMR). This approach to reification results in inconsistent annotation practices, as it allows for multiple representations of similar concepts. To address this inconsistency, we have opted to explicitly include the verb "to be" as a distinct node in our formalism, requiring annotators to manually adjust this structure when encountered. However, in instances of passive voice, we do not explicitly separate the nodes, treating sentences in direct voice instead. We maintain that at our level of abstraction, there is no semantic distinction that necessitates differentiating between passive and active voices, as "to be" serves merely as an auxiliary verb.

**Mode Verb Structure** In traditional formalisms, the possibility of an action is represented using a specific node, such as the possible-01 predicate in

| Relation | Values | | | | |
|---|---|---|---|---|---|
| :tense | (Present) | Past | Future | | |
| :aspect | (Simple) | Perfect | Imperfect | Continuous | Perfect Continuous |
| :mood | (Indicative) | Imperative | Conditional | Subjunctive | |

Table 7: Table showing how we encode the Tense, Aspect and Mood. First column relation, next possible values. Inside parentheses means it is empty since it is the base form.

AMR. For instance, "I can read" in AMR might be depicted as:

$$(p \ / \ possible-01 \ :ARG1 \ (r \ / \ read-02 \ :ARG0 \ (i \ / \ I)))$$

In our approach, aiming to align concepts more closely with sentence lemmas without relying on an external repository, "I can read" is represented as:

$$(p \ / \ can \ :theme \ (r \ / \ read{:}agent \ (i \ / \ I)))$$

However, "can" serves as an auxiliary verb modifying the main verb, leading to various potential representations, such as "may" or "might". To address this, annotators refined the encoding to specifically denote the modal aspect of the verb, as in:

$$(r \ / \ read \ :mode \ possible \ :agent \ (i \ / \ I)))$$

This approach has been extended to other modalities, including Necessity, Obligation, Recommendation, Possibility, and Permission, enhancing clarity and consistency in representation.

Once all graph subsets have been validated, we retain the incorrect silver instances, as they will be used to train a large language model. This model will then apply our validations across the entire silver corpus.

## E Projecting Corrections

After completing the manual validation phase rectifying errors and updating structures, our next step, ❻ of Section 3.2, involved replicating these corrections and changes across all five languages within the $\text{MSL}_{silver}$ dataset. As outlined in Appendix D, tasks such as incorporating missing information and modifying encoding structures are relatively straightforward for humans, requiring only identification of the corresponding sentence part in order to integrate the concept into the graph. However, the linguistic nuances and the need to apply these changes across languages render the process impractical for algorithmic execution. To overcome this, we leveraged Large Language Models (LLMs), which have proven effective as zero-shot annotators for tasks that are conceptually simple and well-supported by sufficient examples.

Initially, we utilized a few-shot approach with GPT-3.5, providing the model with instructions and examples of the modifications needed. While this method sometimes led to accurate captures of meaning and appropriate changes, it often resulted in structural disruptions within the graph, deviating from our MSL framework and generating incorrect graphs. Furthermore, the necessity for a large number of examples in order for effective learning to take place led to unwieldy prompts and the inability to encompass all modification examples in a single prompt, thereby necessitating extensive examples and, in turn, leading to significant costs.

Therefore, we shifted to fine-tuning GPT-3.5 using the silver graphs, the corresponding sentences, and our corrections across German, English, Spanish, French, and Italian. This approach does not require a high volume of examples in order for GPT-3.5 to grasp the task, making it an efficient method for replicating human-like corrections. Subsequently, we validated this refined method on a subset of corrections, assessing its effectiveness through perplexity measures.

Table 9 outlines the hyperparameters fine-tuned in GPT alongside the final losses, indicating these are the only parameters available for fine-tuning. Notably, Experiments 1 and 2 exhibit a higher validation loss than the training loss, suggesting overfitting. This outcome may stem from training for 3 epochs with only 2,000 samples (400 corrections per language), which, despite GPT-3.5's efficiency with fewer examples, proves insufficient. Experiments 5, 6, and 7 achieved lower validation losses,

| Modality | Necessity | Obligation | Recommendation | Possibility |
|---|---|---|---|---|
| Auxiliary | need | have/must | should | can/might/could/may |
| Example | I need to read | I have to read | I should read | I might read |
| AMR | need-01 | obligate-01 | recomend-01 | possible-01 |
| MSL | :mode necessity | :mode obligation | :mode recommendation | :mode possibility |

Table 8: Table showing how we encode the Modality. Rows: Modality, Auxiliary verb, example sentence, AMR representation, MSL representation. Column: row type, and modalities.

| Experiment | Epochs | Batch | Training Loss | Validation Loss |
|---|---|---|---|---|
| Experiment 1 | 3 | 2 | 0.0019 | 0.5876 |
| Experiment 2 | 3 | 2 | 0.0271 | 0.4033 |
| Experiment 3 | 2 | 2 | 0.0961 | 0.3069 |
| Experiment 4 | 2 | 2 | 0.2138 | 0.1875 |
| Experiment 5 | 2 | 4 | 0.1567 | 0.1538 |
| Experiment 6 | 2 | 4 | 0.1405 | 0.1425 |
| Experiment 7 | 2 | 4 | 0.6379 | 0.3071 |

Table 9: Hyper-parameters for fine-tuning GPT-3.5.

which can be attributed to increased batch sizes, which reduced the number of steps and enhanced model generalization.

Delving into specifics with Table 10, Experiment 1 attempted to include two types of message: one with all possible graph relations and another without. The model was tasked with generating the correct graph from the original sentences, leading to high computational costs and poorer performance compared to Experiments 2 and 3. These latter experiments opted for a simplified message format, excluding extraneous relations, which improved outcomes.

Experiments 4 and 6 introduced the silver sentences as input to simplify the task, prompting the model to replicate modifications in the graphs. This approach improved performance over previous experiments. Experiment 5, building on Experiment 4's prompts by adding more task context, unfortunately, increased computational costs and worsened losses. Experiment 7 aimed to minimize redundant spans in graph linearization to reduce decoding tokens, but this led to poorer performance, which is hypothesized to be due to increased difficulty in understanding the graphs. This suggests GPT-3.5's familiarity with AMR linearization might be leveraged, as it likely forms part of its training corpus.

## F Projecting Annotations to Other Languages

Here we give more details for step 7 of Section 3.2, where the corrected graphs created in the previous step are projected to other languages by leveraging the parallel corpus. For any given pair of sentence and graph in English, German, Spanish, French or Italian in our parallel data, we have a sentence in Arabic, Korean, Catalan, Chinese, Galician and Portuguese. To finetune GPT3.5 for the task we follow a very similar approach as 5, where the LLM receives as input the sentence and graph in the same language, but here it also receives the sentence in another target language and instead of generating a corrected version of the graph, we prompt the model to output the graph in the target language. Table 11 shows an example. At train time we use 400 corrected graphs per language in English, German, Spanish, French and Italian to create pair-wise training samples per pairs of languages, either as source or target, creating 3,960 training samples. At inference time we set English as the source, and Arabic, Korean, Catalan, Chinese, Galician and Portuguese as target ones to obtain their annotations for $\text{MSL}_{HQE}$.

## G Architectures used

Throughout our experiments two kinds of architectures are used. All our LLM-based annotations, used in steps 6 and 7 of Section 3.2 are based on GPT3.5, finetuned using OpenAI API in which only the number of epochs and batch size can be defined and are reported in Table 9. For the rest of the experiments described in this paper, we employ CLAP (Martinez Lorenzo and Navigli, 2024), an efficient implementation of SPRING (Bevilacqua et al., 2021), where the parsing task is framed as a seq2seq task for an Encoder-Decoder system. In our case, we use mT5-large (Xue et al., 2021) as the underlying LM with an expanded vocabulary with the necessary extra tokens specific to our MSL

**Experiment 1**
**Message Type 1**
**role**: "system", **content**: "You are a Semantic Parser."
**role**: "user", **content**: "Sentence: We have said: It's raining cats and dogs."
**role**: "assistant", **content**: "Graph: (z1 / say :aspect perfect :agent (z2 / we) :content (z3 / rain_cats_and_dog :aspect continuous))"

**Message Type 2**
**role**: "system", **content**: "You are a Semantic Parser tasked with transforming natural language sentences into semantic graphs. Each graph comprises nodes that represent lemmatized concepts from the sentence, and edges depicting various relations between these concepts. The potential relations include: [:against, :age, :appearance, :composition, :context, :cost, :degree, :extent, :frequency, :identity, :participant, :location, :membership, :ord, :part, :path, :polarity, :possession, :purpose, :quality, :quantity, :related, :subevent, :target, :time, :scale, :compared, :content, :similar, :year, :month, :day, :hour, :contrast, :regardless, :agent, :co-agent, :patient, :co-patient, :theme, :co-theme, :accompanier, :experiencer, :stimulus, :manner, :medium, :source, :product, :instrument, :cause, :concession, :condition, :result, :and, :or, :multi-sentence, :op1, :op2, ..., :opX, :snt1, :snt2, ..., :sntX, :tense past, :tense future, :aspect continuous, :aspect imperfect, :aspect perfect, :aspect perfect_continuous, :mood imperative, :mood conditional, :mood subjunctive, :mode necessity, :mode obligation, :mode recommendation, :mode probability]. Each relation has its inverse form, which is denoted by adding _of at the end."
**role**: "user", **content**: "Sentence: We have said: It's raining cats and dogs."
**role**: "assistant", **content**: "Graph: (z1 / say :aspect perfect :agent (z2 / we) :content (z3 / rain_cats_and_dog :aspect continuous))"

---

**Experiment 2 & Experiment 3**
**role**: "system", **content**: "You are a Semantic Parser."
**role**: "user", **content**: "Sentence: We have said: It's raining cats and dogs."
**role**: "assistant", **content**: "Graph: (z1 / say :aspect perfect :agent (z2 / we) :content (z3 / rain_cats_and_dog :aspect continuous))"

---

**Experiment 4 & Experiment 6**
**role**: "system", **content**: "You are a Semantic Parser."
**role**: "user", **content**: "Sentence: We have said: It's raining cats and dogs. Silver Graph: (z1 / say :agent (z2 / we) :theme (z3 / rain :theme (z4 / and :op1 (z5 / cat) :op2 (z6 / dog))))"
**role**: "assistant", **content**: "Graph: (z1 / say :aspect perfect :agent (z2 / we) :content (z3 / rain_cats_and_dog :aspect continuous))"

---

**Experiment 5**
**role**: "system", **content**: "You are a Semantic Parser. Given a sentence and its potentially incorrect silver graph representation, correct the graph to accurately reflect the sentence's meaning."
**role**: "user", **content**: "Sentence: We have said: It's raining cats and dogs. Silver Graph: (z1 / say :agent (z2 / we) :theme (z3 / rain :theme (z4 / and :op1 (z5 / cat) :op2 (z6 / dog))))"
**role**: "assistant", **content**: "Graph: (z1 / say :aspect perfect :agent (z2 / we) :content (z3 / rain_cats_and_dog :aspect continuous))"

---

**Experiment 7**
**role**: "system", **content**: "You are a Semantic Parser."
**role**: "user", **content**: "Sentence: We have said: It's raining cats and dogs. Silver Graph: say :agent we) :theme rain :theme and :op1 cat) :op2 dog))))"
**role**: "assistant", **content**: "Graph: say :aspect perfect :agent we) :content rain_cats_and_dog :aspect continuous))"

---

Table 10: Example of prompts for training GPT. For predicting is the same format but with the role assistant message.

| |
|---|
| **role**: "system", **content**: "You are a Multilingual Semantic Parser." |
| **role**: "user", **content**: "English Sentence: We say: It's raining cats and dogs out. English Graph: (z1 / say :agent (z2 / we) :content (z3 / rain_cats_and_dog :aspect continuous)) Spanish Sentence: Decimos: Llueve a cántaros'. Spanish Graph:" |
| **role**: "assistant", **content**: "Spanish Graph: (z1 / decir :agent (z2 / nosotros) :content (z3 / llover_a_cantaros))" |

Table 11: Example of prompts for training GPT to project to other languages. For predicting is the same format but with the role assistant message.

layer, as SPRING did for AMR. For all experiments we use the Adam (Kingma and Ba, 2015) optimizer with a $5x10^{-5}$ learning rate, 2048 token batch size and a beam size of 5 at inference time. We train every model on a single NVIDIA® RTX 3090 graphic card with 24GB of VRAM. The average time to train each model was 48h.

## H More Results Generation

Table 12 and Table 13 present additional results for the **Generation** and **Back-translation** experiments, using other evaluation metrics: Bert-score (Zhang et al., 2020), Chrf++ (Popović, 2017), and Rouge-L (Lin, 2004). Interestingly, the BLEU score for Arabic in MSL$_{HQE}$ is notably lower compared to other languages, reflecting BLEU's known limitations for Arabic (Bouamor et al., 2014). However, other metrics such as Bert-score in Table 13 reveal a much narrower performance gap, with MSL$_{HQE}$ achieving the highest performance across all languages.

## I More Results Parsing

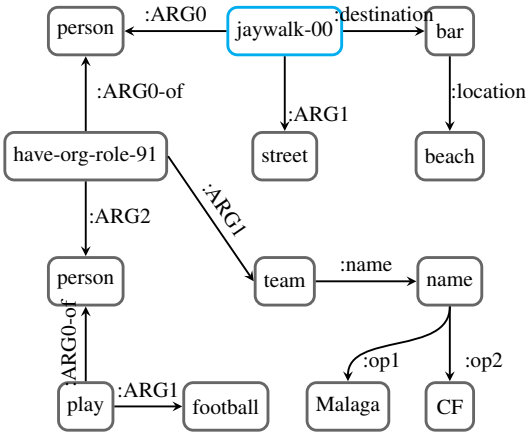
When we compare the SMATCH score of our system trained with MSL$_{HQE}$ (71.35) against the system trained in AMR score using the exact same architecture for English (83.0), we can observe how there are around 12 points of difference. However, as we mentioned in the paper, SMATCH values are not comparable across semantic representations. Our merging nodes and inclusion of multi-word expressions reduce the number of nodes in the graphs, reducing the number of matches between the gold and other reference graphs, and making the error have a bigger impact. For example, in Figure 11 we can see the impact of having an error in an AMR graph structure and having an error on the MSL representation. In the example, we observe how the predicted graph for MSL and AMR wrongly expresses the named entity of "Malaga CF" as "Malaga". However, since AMR tends to have more graph nodes to represent the same idea than an MSL, the error's impact in MSL is much higher (We can observe how 11(d) has 6 points lower than 11(c)). Analyzing the node counts in the training AMR 3.0 corpus graphs, AMR graphs contain *641,431* nodes, while MSL graphs for the same sentences have *423.485*.

| BLEU | Arabic | Catalan | German | English | Spanish | Korean | French | Galician | Italian | Portuguese | Chinese | AVG | SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $MSL_{AMR}$ | 0.00 | 20.21 | 33.48 | 40.60 | 44.47 | 2.80 | 37.36 | 15.00 | 38.28 | 25.50 | 1.27 | 26.06 | 16.74 |
| $MSL_{Silver}$ | 0.00 | 18.20 | 32.47 | 44.24 | 48.02 | 1.08 | 39.82 | 16.04 | 41.23 | 25.89 | 0.98 | 24.36 | 18.26 |
| $MSL_{HQ}$ | 0.00 | 16.37 | 39.00 | 55.19 | 57.00 | 1.15 | 51.24 | 16.08 | **51.80** | 25.08 | 1.07 | 28.54 | 23.07 |
| $MSL_{HQE}$ | **30.8**4 | **57.30** | **40.46** | **57.26** | **57.23** | **22.62** | **51.62** | **54.31** | 51.31 | **49.57** | **48.48** | **47.36** | 11.46 |
| Bert-score | Arabic | Catalan | German | English | Spanish | Korean | French | Galician | Italian | Portuguese | Chinese | AVG | SD |
| $MSL_{AMR}$ | 58.76 | 85.14 | 88.76 | 94.60 | 90.14 | 68.31 | 90.11 | 86.46 | 90.64 | 84.97 | 45.65 | 80.31 | 15.71 |
| $MSL_{Silver}$ | 56.27 | 85.91 | 88.71 | 94.44 | 90.74 | 68.31 | 91.38 | 86.98 | 90.59 | 86.72 | 45.29 | 80.48 | 16.35 |
| $MSL_{HQ}$ | 55.35 | 85.69 | **90.12** | 95.58 | 91.92 | 69.01 | 91.57 | 86.84 | 91.10 | 87.01 | 45.68 | 80.90 | 16.44 |
| $MSL_{HQE}$ | **89.88** | **92.11** | 90.00 | **96.22** | **92.01** | **84.96** | **91.90** | **91.62** | **91.19** | **91.10** | **89.30** | **90.94** | 2.69 |
| Chrf++ | Arabic | Catalan | German | English | Spanish | Korean | French | Galician | Italian | Portuguese | Chinese | AVG | SD |
| $MSL_{AMR}$ | 3.98 | 55.93 | 67.61 | 68.62 | 71.69 | 27.82 | 69.64 | 57.72 | 70.61 | 57.48 | 3.07 | 50.38 | 26.27 |
| $MSL_{Silver}$ | 7.35 | 54.27 | 66.93 | 71.36 | 73.49 | 29.63 | 73.57 | 57.81 | 74.61 | 59.13 | 4.51 | 52.06 | 26.22 |
| $MSL_{HQ}$ | 6.67 | 54.60 | **69.85** | 75.86 | 75.90 | 30.07 | **74.12** | 58.71 | 74.87 | 59.47 | 4.67 | 53.16 | 27.09 |
| $MSL_{HQE}$ | **64.83** | **76.01** | 69.76 | **76.05** | **76.70** | **58.68** | 73.97 | **73.76** | **75.27** | **72.45** | **41.93** | **69.04** | 10.55 |
| Rouge-L | Arabic | Catalan | German | English | Spanish | Korean | French | Galician | Italian | Portuguese | Chinese | AVG | SD |
| $MSL_{AMR}$ | 1.57 | 44.97 | 63.64 | 69.31 | 70.16 | 17.83 | 61.95 | 44.66 | 63 51 | 51.39 | 9.68 | 45.76 | 23.60 |
| $MSL_{Silver}$ | 1.57 | 45.08 | 63.33 | 73.65 | 72.67 | 16.00 | 65.56 | 44.73 | 69.26 | 54.03 | 14.71 | 47.32 | 25.69 |
| $MSL_{HQ}$ | 1.62 | 44.72 | 66.47 | 77.23 | **75.82** | 16.19 | 70.12 | 46.53 | 72.31 | 54.47 | 15.67 | 49.19 | 27.01 |
| $MSL_{HQE}$ | **55.33** | **75.25** | **67.17** | **78.94** | **75.82** | **47.68** | **71.76** | **72.84** | **73.07** | **71.01** | **77.44** | **69.67** | 9.69 |

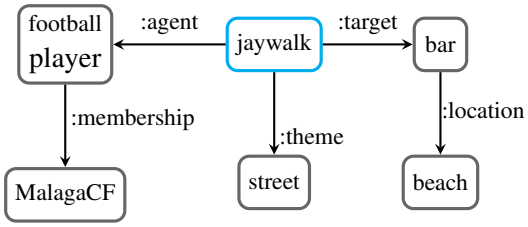
Table 12: Additional Evaluation Metrics for the **Generation** Experiment. Row Blocks: BLEU, Bert-score, Chrf++, and Rouge-L. Each block consists of four rows corresponding to each semantic representation: $MSL_{AMR}$, $MSL_{Silver}$, $MSL_{HQ}$, and $MSL_{HQE}$. There is one column per language, alongside columns for the average value across languages and the standard deviation. Best result per language and per Block in bold.

| | AMR | | | | | | Out Of Domain | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | DE | EN | ES | IT | AVG | SD | AR | CA | DE | EN | ES | KO | FR | GL | IT | PT | ZH | AVG | SD |
| AMR | 21.55 | 32.40 | 31.02 | 29.00 | 28.49 | 4.83 | 0.00 | 1.52 | 15.87 | 25.49 | 18.93 | 1.73 | 0.45 | 2.97 | 15.43 | 1.68 | 0.00 | 7.64 | 9.48 |
| BMR | 27.16 | 38.97 | 36.67 | 29.30 | 33.03 | 5.68 | 0.00 | 1.34 | 22.01 | 36.29 | 26.41 | 1.82 | 0.24 | 3.31 | 21.89 | 1.41 | 0.25 | 10.45 | 13.40 |
| MSL | **41.82** | **51.36** | **52.77** | **42.56** | **47.13** | 5.73 | **24.33** | **46.34** | **43.52** | **52.81** | **53.48** | **27.56** | **49.41** | **48.41** | **44.88** | **48.02** | **51.35** | **44.56** | 9.73 |
| Bert-score | DE | EN | ES | IT | AVG | SD | AR | CA | DE | EN | ES | KO | FR | GL | IT | PT | ZH | AVG | SD |
| AMR | 89.84 | 94.01 | 92.34 | 90.91 | 91.76 | 1.01 | 73.74 | 82.52 | 89.32 | 92.49 | 90.31 | 80.40 | 81.98 | 84.85 | 89.20 | 83.11 | 71.15 | 83.55 | 6.75 |
| BMR | 91.58 | 94.62 | 93.28 | 92.14 | 92.90 | 1.44 | 73.20 | 82.64 | 91.04 | 93.84 | 91.93 | 78.45 | 80.78 | 85.16 | 90.96 | 83.44 | 70.83 | 83.84 | 7.70 |
| MSL | **94.49** | **93.64** | **94.92** | **94.09** | **94.28** | 0.49 | **96.94** | **94.51** | **95.52** | **96.03** | **96.15** | **93.01** | **95.93** | **95.19** | **95.11** | **95.53** | **95.54** | **95.40** | 1.15 |
| Chrf++ | DE | EN | ES | IT | AVG | SD | AR | CA | DE | EN | ES | KO | FR | GL | IT | PT | ZH | AVG | SD |
| AMR | 56.16 | 70.60 | 63.52 | 59.77 | 62.51 | 6.17 | 0.26 | 28.17 | 49.74 | 62.63 | 52.44 | 26.45 | 22.84 | 33.71 | 50.64 | 28.77 | 0.67 | 31.39 | 20.36 |
| BMR | 62.50 | 74.34 | 68.17 | 65.58 | 73.83 | 5.03 | 0.20 | 27.18 | 57.03 | 71.06 | 60.76 | 19.10 | 19.71 | 34.94 | 58.32 | 28.94 | 0.48 | 34.33 | 24.44 |
| MSL | **76.25** | **72.04** | **78.06** | **76.84** | **75.79** | 2.62 | **67.57** | **78.49** | **79.06** | **80.91** | **82.40** | **75.61** | **80.30** | **80.38** | **78.44** | **80.13** | **51.71** | **75.90** | 8.95 |
| Rouge-L | DE | EN | ES | IT | AVG | SD | AR | CA | DE | EN | ES | KO | FR | GL | IT | PT | ZH | AVG | SD |
| AMR | 46.77 | 61.12 | 56.64 | 53.12 | 54.51 | 6.05 | 0.77 | 16.40 | 43.72 | 58.64 | 45.41 | 17.53 | 8.38 | 22.07 | 40.97 | 17.37 | 2.93 | 24.92 | 19.24 |
| BMR | 53.85 | 66.42 | 63.14 | 55.03 | 59.61 | 6.14 | 0.72 | 17.08 | 52.15 | 67.69 | 55.07 | 11.96 | 6.73 | 25.13 | 50.21 | 19.05 | 2.04 | 27.99 | 23.91 |
| MSL | **67.45** | **67.71** | **72.04** | **66.01** | **68.30** | 2.60 | **62.33** | **74.40** | **74.02** | **80.80** | **78.27** | **61.57** | **75.75** | **76.16** | **72.35** | **75.50** | **79.78** | **73.72** | 6.39 |

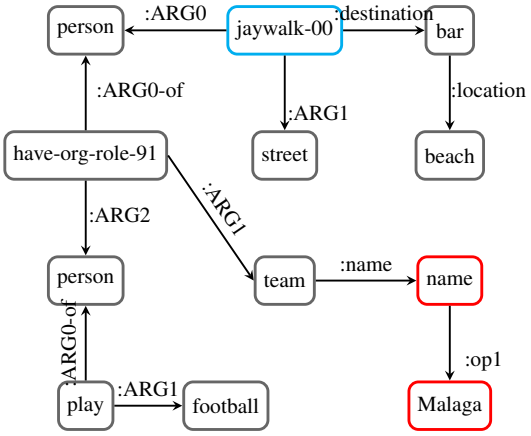
Table 13: Additional Evaluation Metrics for the **Back-translation** Experiment. Row Blocks: BLEU, Bert-score, Chrf++, and Rouge-L. Each block includes three rows, each representing a semantic formalism: AMR, BMR, and MSL. Column Blocks: AMR for four translation datasets and Out of Domain data (non-AMR). Each block presents a series of languages from the test set, followed by the average across these languages and the standard deviation. Best result per language and per Block in bold.



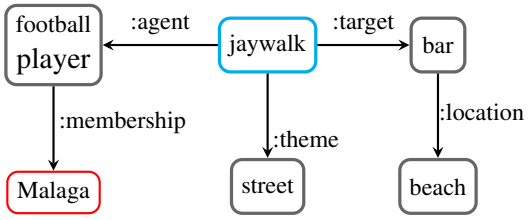

(a) Reference AMR Graph.




(b) Reference MSL Graph.




(c) Predicted AMR Graph. SMATCH = 98.0




(d) Predicted MSL Graph. SMATCH = 92.0

Figure 11: Example of differences SMATCH scores across formalisms. 11(a) Reference AMR graph. 11(b) Reference MSL Graph, 11(c) predicted example AMR graph. 11(d) predicted example of MSL graph.