D3.6

Multilingual semantic parsing - final report

Author(s): Rexhina Blloshmi, Federico Martelli, Roberto Navigli, Paola Velardi

Date: 28/07/2022

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

Deliverable Number: D3.6

Dissemination Level: Public

Delivery Date: 31/07/2022

Version: 1.0

Author(s): Rexhina Blloshmi, Federico Martelli, Roberto Navigli, Paola Velardi

2This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Deliverable/Document Information

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Document History

Version Date	Changes/Approval	Author(s)/Approved by
28/07/2022		Rexhina Blloshmi, Federico Martelli, Roberto Navigli, Paola Velardi

4This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Table of Contents

List of Tables

List of Figures

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

1 Introduction

The present document illustrates the research activities carried out in task 3.2 (work package 3) focused on semantic parsing, the computational task aimed at providing formal meaning representations of utterances (see Figure 1).

Allowing machines to interpret and understand natural language is one of the long-standing goals of Artificial Intelligence (AI) and Natural Language Understanding (NLU). Over the course of the last few years, several representations for semantic parsing have been proposed such as Elementary Dependency Structures (EDS, Oepen and Lønning, 2006), Prague Tectogrammatical Graphs (PTG, Hajič et al., 2012), Abstract Meaning Representation (AMR, Banarescu et al., 2013), Universal Conceptual Cognitive Annotation (UCCA, Abend and Rappoport, 2013), Universal Decompositional Semantics (UDS, White et al., 2016), Parallel Meaning Bank (Abzianidze et al., 2017, PMB). Although the vast majority of such representations focus on English, some attempts have been made to address multilinguality such as UCCA and PMB. Consequently, the task of semantic parsing is mainly focused on English and features language-specific constraints which hamper their scalability remarkably. Additionally, such approaches are not fully semantic and are not able to effectively abstract away from language-specific lexicons. Another limitation is that they rely heavily on supervision or on knowledge resources which seem to work only in specific tasks such as question answering.
With our work in task 3.2 we addressed both the issue of scaling representations multilingually and performing semantic parsing text more effectively in a number of respects.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Besides the work described in the intermediate report on semantic parsing (D3.4), in 2021 and 2022 we performed the following work:

1. **Symmetric PaRsIng aNd Generation** (*SPRING*, Bevilacqua et al. 2021), a novel Transformer-based symmetric approach which achieves state-of-the-art results in Text-to-AMR parsing and AMR-to-Text generation with a single seq2seq architecture;
2. **Speaking the Graph Languages** (*SGL*, Procopio et al. 2021, a), a novel framing of semantic parsing towards multiple formalisms as Multilingual Neural Machine Translation;

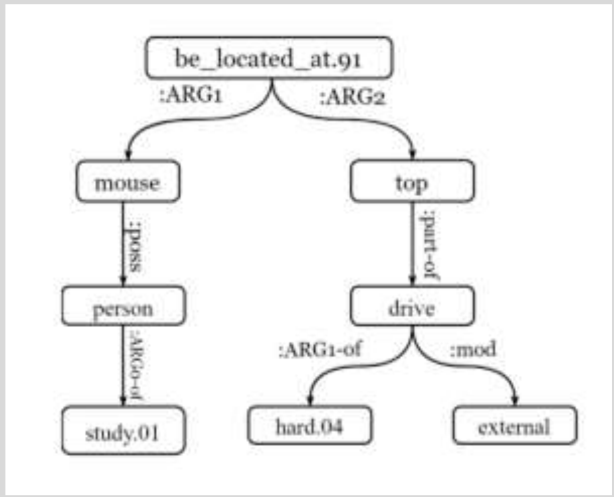

Figure 1 - AMR graph for the sentence "*The student's mouse is on top of the external hard drive*". Figure extracted by Navigli et al. 2021.

3. **BabelNet Meaning Representation** (*BMR*, Navigli et al. 2022; Martinez Lorenzo et al. 2022), a new language-independent formalism that abstracts away from language-specific constraints thanks to two multilingual semantic resources. We describe all of them in detail in the next sections.

Finally, we also highlight that, in this task and, more generally, within work package 3, great importance was attached to the manual creation of novel lexical-semantic resources which can be used for training and evaluation purposes in different NLP tasks and also for encouraging deeper connections between two related scientific fields such as lexicography and NLP (Martelli et al. 2021a). Such resources are created specifically for multilingual semantic parsing, but also for other closely-related mutually-beneficial tasks such as Word Sense Disambiguation (WSD, Maru et al. 2019), multilingual and cross-lingual Word in Context (Martelli et al. 2021b) and Idiomatic Expression Identification (Tedeschi et al. 2022).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Published Works in Task 3.2

- Di Fabio, A., Conia, S., & Navigli, R. (2019, November). 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)* (pp. 627-637). *Described in D3.4.*
- Blloshmi, R., Tripodi, R., & Navigli, R. (2020, November). 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)* (pp. 2487-2500). *Described in D3.4*.
- Procopio, L., Tripodi, R., & Navigli, R. (2021a, June). SGL: Speaking the Graph Languages of semantic parsing via multilingual translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 325-337).
- Bevilacqua, M., Blloshmi, R., & Navigli, R. (2021, May). One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 14, pp. 12564-12573).
- Martinez Lorenzo, A. C., Maru, M., & Navigli, R. (2022, May). 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)* (pp. 1727-1741).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

2. SPRING

We proposed for the first time an effective unified seq2seq approach based on a pretrained Transformer encoder-decoder architecture to generate either an accurate linearization of the AMR graph for a sentence or, vice versa, a sentence for a linearization of the AMR graph (Bevilacqua et al. 2021). Contrary to previous reports (Konstas et al. 2017), we find that the choice between competing graph-isomorphic linearizations does matter. We therefore studied three different approaches to graph linearization. To address the issue of scaling across domains, we also proposed a novel Out-of-Distribution (OOD) setting for estimating the ability of the Text-to-AMR and AMR-toText approaches to generalize on open-world data. The model, tested across settings and datasets, achieves the state of the art both on parsing and generation. SPRING is available at: github.com/SapienzaNLP/spring.

2.1 Method

SPRING performs Text-to-AMR and AMR-to-Text parsing by exploiting the transfer learning capabilities of BART (Lewis et al. 2020). In SPRING, AMR graphs are handled symmetrically: for Text-to-AMR parsing the encoder-decoder is trained to predict a graph given a sentence; for AMR-to-Text generation another specular encoder-decoder is trained to predict a sentence given a graph. In order to use the graphs within the seq2seq model, we transform them into a sequence of symbols using various different linearization techniques. Furthermore, we modify the BART vocabulary in order to make it suitable for AMR concepts, frames and relations. Finally, we define lightweight, non content-modifying heuristics to deal with the fact that, in parsing, seq2seq may output strings which cannot be decoded into a graph. For more implementation details we refer the reader to Bevilacqua et al. 2021.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

2.1.1 Graph Linearizations

Because seq2seq models require to input and output sequences of tokens, we need linearization techniques to transform graphs into sequences and vice versa. We use linearization techniques which are fully graph-isomorphic, i.e., it is possible to encode the graph into a sequence of symbols and then decode it back into a graph without losing adjacency information. We propose the following graph linearization techniques:

- **PENMAN** (Goodman 2020)**:** a standard format used to represent graphs in linearized form.
- **DFS-based:** Depth-First Search, on which PENMAN is based, is very attractive as it is quite closely related to the way natural language syntactic trees are linearized
- **BFS-based** The use of Breadth-First Search traversal is motivated by the fact that it enforces a locality principle by which things belonging together are close to each other in the flat representation

All the above linearizations are decoded into the same graph. However, in the PENMAN-linearized gold annotations, an edge ordering can be extracted from each AMR graph.

2.1.2 Vocabulary

Since BART leverages a subword vocabulary and its tokenization is not well suited for AMR symbols, we expand the tokenization vocabulary of BART by including: i) all the relations and frames occurring at least 5 times in the training corpus; ii) constituents of the AMR tokens and iii) special tokens required for graph linearizations.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

2.1.3 Postprocessing

In SPRING we perform light postprocessing, in order to ensure the validity of the graph produced at parsing time. To do this, we restore parenthesis parity in PENMAN and DFS, and also remove any invalid token due to not being a possible continuation given the token that precedes it. For BFS, we recover a valid set of triples between each subsequent pair of tokens.

2.2 Experimental setup

To show the capability of SPRING both in the parsing (Text-to-AMR) and generation (AMR-to-Text) tasks we performed a variety of experiments whose setup we describe hereafter.

2.2.1 Datasets

In-Distribution. We evaluate SPRING on the standard evaluation benchmarks, which we refer to as the In-Distribution (ID) setting. The data used in this setting are the AMR 2.0 (LDC2017T10) and AMR 3.0 (LDC2020T02) corpora, which include, respectively 39,260 and 59,255 manually-created sentence-AMR pairs. AMR 3.0 is a superset of AMR 2.0. In both datasets the training, development and test sets are a random split of a single dataset, therefore they are drawn from the same distribution.

Out-of-Distribution. Since the ID setting does not allow estimates about performances on open-world data, which will likely come from a different distribution of that of the training set, we also propose a novel OOD setting. In this evaluation setting, we assess SPRING when trained on OOD data, contrasting it with the ID results. We employ the AMR 2.0 training set, while for testing we use three distinct Out-of-Distribution (OOD) benchmarks, covering a wide range of different genres:

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

1. New3, a set of 527 instances from AMR 3.0, whose original source was the LORELEI DARPA project – not included in the AMR 2.0 training set – consisting of excerpts from newswire and online forums;
2. TLP, the full AMR-tagged children's novel The Little Prince (ver. 3.0), consisting of 1,562 pairs;
3. Bio, i.e., the test set of the Bio-AMR corpus, consisting of 500 instances, featuring biomedical texts (May and Priyadarshi 2017).

Silver. In order to determine whether silver-data augmentation, another commonly used technique, is beneficial in both ID and OOD, we follow Konstas et al. (2017) and create pretraining data by running the SPRING parser using DFS (trained on AMR 2.0) on a random sample of the Gigaword (LDC2011T07) corpus consisting of 200,000 sentences.

2.2.2 Models

SPRING draws on BART with the augmented vocabulary. We use the same model hyperparameters as BART Large (or Base, when specified), as defined in the Huggingface[1] library. Models are trained for 30 epochs using cross-entropy with a batch size of 500 graph linearization tokens, with RAdam (Liu et al. 2020) optimizer and a learning rate of 1 × 10−5. The gradient is accumulated for 10 batches, while dropout is set to 0.25. We perform hyperparameter search to train and evaluate both the Text-to-AMR and AMR-to-Text models. At prediction time, we set the beam size to 5 following common practice in neural machine translation (Yang, Huang, and Ma 2018).

SPRING variants. We use models trained with the three linearizations, indicated as SPRING[lin], where [lin] is one of the linearizations: PENMAN (PM), DFS- (DFS) or BFS-based (BFS). In addition, we include variants of SPRING DFS using i) BART Base (base); ii) graph recategorization (+recat); iii) pretrained silver AMR data (+silver). BART baseline We also

[1] https://huggingface.co/

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

report results on a vanilla BART baseline which treats PENMAN as a string, uses no vocabulary expansion and tokenizes the graph accordingly.

2.2.3 Comparison Systems

In-distribution. In the ID setting, we use the AMR 2.0 benchmark to compare SPRING variants against the best models from the literature. To this end, we include the following Text-to-AMR parsers: i) Ge et al. (2019, Ge+), an encoder-decoder model which encodes the dependency tree and semantic role structure alongside the sentence; ii) Lindemann, Groschwitz, and Koller (2019, LindGK), a compositional parser based on the Apply-Modify algebra; iii) Naseem et al. (2019, Nas+), a transition-based parser trained with a reinforcement-learning objective rewarding the Smatch score; iv) Zhang et al. (2019b, Zhang+), a hybrid graph- and transition-based approach incrementally predicting an AMR graph; v) Zhou et al. (2020, Zhou+), an aligner-free parser (Zhang et al. 2019a) enhanced with latent syntactic structure; vi) Cai and Lam (2020a, CaiL), a graph-based parser iteratively refining an incrementally constructed graph. For AMR-to-Text, instead, we include the following: i) Zhu et al. (2019, Zhu+), a Transformer-based approach enhanced with structure-aware self-attention; ii) Cai and Lam (2020b, CaiL), a graph Transformer model which relies on multi-head attention (Vaswani et al. 2017) to encode an AMR graph in a set of node representations; iii) Wang, Wan, and Yao (2020, Wang+), a Transformer-based model generating sentences with an additional structure reconstruction objective; iv) Zhao et al. (2020, Zhao+), a graph attention network which explicitly exploits relations by constructing a line graph; v) Yao, Wang, and Wan (2020, Yao+), a graph Transformer-based model which encodes heterogeneous subgraph representations; vi) Mager et al. (2020, Mag+), a fine-tuned GPT-2 model (Radford et al. 2019) predicting the PENMAN linearization of an AMR graph. For AMR 3.0, which is a recent benchmark, there are no previous systems to compare against. Thus, we train the previous state-of-the-art parsing model of Cai and Lam (2020) on AMR 3.0 and perform the corresponding evaluation.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Out-of-distribution. In the OOD setting we compare the SPRINGDFS variants when trained on AMR 2.0 and test on OOD data against the best of the same variants trained on the corresponding ID training set when available (New3 and Bio).

2.3 Results

The results on the AMR 2.0 benchmark are reported in Table 1. Among the three different simple linearization models, i.e., SPRING DFS, SPRING BFS, and SPRING PM, the DFS-based one achieves the highest overall Smatch, obtaining slightly better results than the second best one, the PENMAN, and a wider margin over the BFS one.

We report in Table 2 the AMR 2.0 AMR-toText results. SPRING DFS achieves 45.3 BLEU points, improving the previous state of the art (Yao, Wang, and Wan 2020) by 11 points, and obtains very significant gains in chrF++ and METEOR as well.

Model	Recat.	Smatch	Unlab.	NoWSD	Conc.	Wiki.	NER	Reent.	Neg.	SRL
Ge+ (2019)	N	74.3	77.3	74.8	84.2	71.3	82.4	58.3	64.0	70.4
LindGK (2019)**	N	75.3	-	-	-	-	-	-	-	-
Nas+ (2019)**	N	75.5	80.0	76.0	86.0	80.0	83.0	56.0	67.0	72.0
Zhang+ (2019b)**	Y	77.0	80.0	78.0	86.0	86.0	79.0	61.0	77.0	71.0
Zhou+ (2020)*	Y	77.5	80.4	78.2	85.9	86.5	78.8	61.1	76.1	71.0
CaiL (2020a)*	N	78.7	81.5	79.2	88.1	81.3	87.1	63.8	66.1	74.5
CaiL (2020a)*	Y	80.2	82.8	80.0	88.1	86.3	81.1	64.6	78.9	74.2
SPRING^{DFS}	N	83.8	86.1	84.4	90.2	84.3	**90.6**	70.8	74.4	79.6
SPRING^{BFS}	N	83.2	85.7	83.7	90.3	83.5	90.2	70.9	70.9	78.2
SPRING^{PM}	N	83.6	86.1	84.1	90.1	83.1	90.2	71.4	72.7	79.4
BART baseline	N	82.7	85.1	83.3	89.7	82.2	90.0	70.8	72.0	79.1
SPRING^{DFS} (base)	N	82.8	85.3	83.3	89.6	83.5	89.9	70.2	71.5	79.0
SPRING^{DFS} +recat	Y	**84.5**	**86.7**	**84.9**	89.6	**87.3**	83.7	72.3	**79.9**	79.7
SPRING^{DFS} +silver	N	84.3	**86.7**	84.8	**90.8**	83.1	90.5	**72.4**	73.6	**80.5**

Table 1: Text-to-AMR parsing results (AMR 2.0). Row blocks: previous approaches; SPRING variants; baseline + other SPRING DFS. Columns: model; recategorization (Y/N); SMatch; Fine-grained scores. The best result per measure across the table is shown in bold. The best result per measure within each row block is underlined. Models marked with */* rely on BERT Base/Large.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

	BL	CH+	MET	BLRT
Zhu+ (2019)	31.8	64.1	36.4	-
CaiL (2020b)	29.8	59.4	35.1	-
Wang+ (2020)	32.1	64.0	36.1	-
Zhao+ (2020)	32.5	-	36.8	-
Mag+ (2020)	33.0	63.9	37.7	-
Yao+ (2020)	34.1	65.6	38.1	-
SPRING^{DFS}	45.3	73.5	41.0	56.5
SPRING^{BFS}	43.6	72.1	40.5	54.6
SPRING^{PM}	43.7	72.5	41.3	56.0
BART baseline	42.7	72.2	40.7	54.8
SPRING^{DFS} +silver	**45.9**	**74.2**	**41.8**	**58.1**

Table 2: AMR-to-Text generation results (AMR 2.0). Row blocks: previous approaches; SPRING variants; baseline +silver. Columns: measures. Bold/underline as in Table 2.

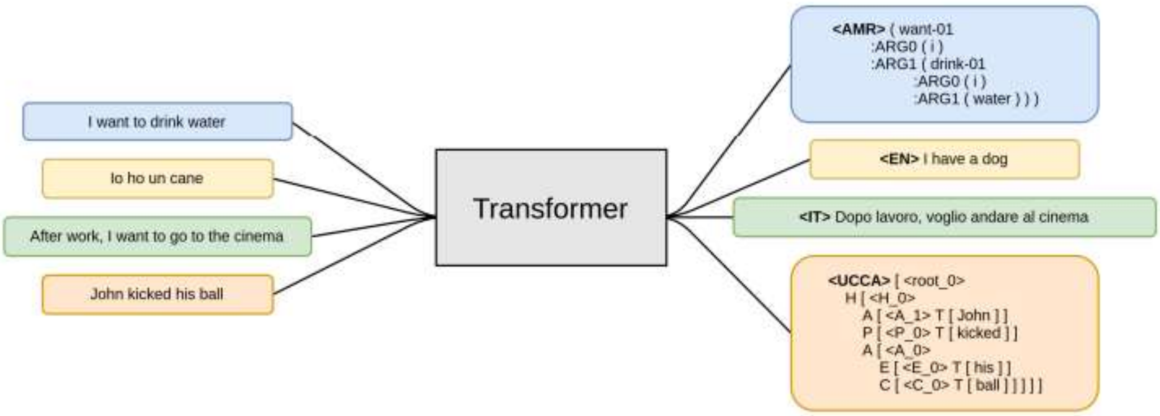

Figure 2: The SGL multilingual translation framework

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

3 SGL

As mentioned above, over the course of the last few years, directed graphs have gained a significant interest in graph-based semantic parsing, with the development of several formalisms. However, approaches capable of competitively scaling across formalisms represent a natural desideratum and recent works have started to explore this direction (Hershcovich et al., 2018; Oepen et al., 2019). Nevertheless, despite the promising results achieved, research in this direction has been hampered by the lack of training data affecting semantic parsing.

To address these drawbacks, we propose Speak the Graph Languages (SGL - Procopio et al. 2021a), a many-to-many seq2seq architecture capable of scaling across formalisms and languages. The key idea behind SGL is to train a seq2seq model with a Multilingual Neural Machine Translation (MNMT) objective, where, given an input text and an identifier denoting the desired output formalism, a single shared model has to learn to translate towards the corresponding linearized graph.

SGL provides the following contributions:

1. we reframe semantic parsing towards multiple formalisms and from multiple languages as multilingual machine translation;
2. As far as AMR parsing is concerned, SGL achieves competitive performances, surpassing most of its current competitors when exploiting a pre-trained Transformer;
3. We outperform all competitors in cross-lingual AMR parsing without ever seeing non-English to AMR examples at training time and push the current state of the art even further once we include these examples;
4. In the UCCA parsing task, we reach competitive results, outperforming a strong BERT-powered baseline (Hershcovich and Arviv, 2019).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

SGL is available at https://github.com/SapienzaNLP/sgl.

3.1 Method

We now describe SGL, a novel approach to graph-based semantic parsing. We first explain the graph linearizations employed for AMR and UCCA, along with their delinearizations. Subsequently, we describe the seq2seq modelling approach used and, finally, we present our multilingual framework.

Graph Linearizations. For *AMR parsing*, we first remove variables and wiki links by AMR graphs. Subsequently, we convert the AMR graphs into trees by duplicating coreferring nodes. In order to obtain the final linearized version of a given AMR, we concatenate all the lines of its PENMAN notation, replacing newlines and multiple spaces with single spaces. Instead, delinearization is carried out by assigning a variable to each predicted concept, performing Wikification, restoring co-referring nodes and, whenever possible, repairing any syntactically malformed subgraph. In both phases, we use the scripts released by van Noord and Bos (2017). As for *UCCA parsing*, we employ a Depth-First Search (DFS) approach: starting from the root, we navigate the graph, using square brackets to delimit subgraph boundaries and special variables to denote terminal and non-terminal nodes; remote edges are denoted by a special modifier appended to their labels, while reentrancies, that is, edges whose target is a node already seen, are handled by simply entering the respective variable. Similarly to AMR, delinearization is carried out by back-parsing this sequence into a UCCA graph, repairing malformed subgraphs whenever possible; moreover, as terminal nodes are anchored in UCCA, we remove those whose anchoring is impossible. The linearization and delinearization scripts for this schema are released along with the rest of our code.

Sequence-to-sequence Modelling. In order to perform sequence-to-sequence modelling, we employ neural seq2seq models based upon the Transformer architecture (Vaswani et al., 2017). Specifically, we use two different kinds of Transformer architecture, namely *Cross* and mBART (Liu et al., 2020). *Cross* is a randomly initialized Transformer closely following the

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

architecture depicted by Vaswani et al. (2017), except for a significant difference: we leverage a factorized embedding parameterization (Lan et al., 2020), that is, we decompose the large vocabulary embedding matrix into two smaller matrices. Instead, mBART is a multilingual Transformer pre-trained in many languages over large-scale monolingual corpora.

Multilingual Framework. In order to enable scalability across languages, we employ a data-driven approach: we replace the start token of the decoder with a special tag specifying the language the encoder representations should be unrolled towards. Since our focus is on semantic parsing, we perform oversampling on the AMR and UCCA datasets. Furthermore, when considering the parallel corpora from MT, we change the training direction with probability 0.5, hence allowing our model to see at training time both the X → EN and EN → X training directions.

3.2 Experimental setup

We assess the effectiveness of our proposed approach by evaluating its performance on all translation paths in which the target language is a graph formalism, the only exception being X → UCCA, with X being any language but English. This is due to the fact that, differently from AMR where cross-lingual AMR aims to produce English-based meaning representations (Damonte and Cohen, 2018), UCCA builds graphs on top of its tokens which are, consequently, inherently in the same language as the input text (Hershcovich et al., 2019); we leave exploring this direction to future work.

3.2.1 Datasets

AMR. For AMR parsing, we use AMR-2.0 (LDC2017T10) and its recently released expansion, AMR-3.0 (LDC2020T02), amounting, respectively, to 39 260 and 59 255 manually-created sentence-graph pairs.

Cross-Lingual AMR. We use Abstract Meaning Representation 2.0 - Four Translations

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

(Damonte and Cohen, 2020) to investigate the performance of SGL on cross-lingual AMR parsing. This corpus contains translations of the sentences in the test set of AMR-2.0 in Chinese (ZH), German (DE), Italian (IT) and Spanish (ES).

UCCA. We replicate the setting of the CoNLL 2019 Shared Task (Oepen et al., 2019). We train our models using the freely available UCCA portion of the training data; this corpus amounts to 6572 sentence-graph pairs, drawn from the English Web Treebank (2012T13) and English Wikipedia articles on celebrities.

3.2.2 Models

To carry out our experiments, we exploit both models namely Cross and mBART, a multilingual pretrained Transformer, to better grasp the effects of this joint multilingual framework in different regimes. Specifically, we investigate the following scenarios:

- models trained only on a single semantic parsing task (AMR or UCCA parsing) and without considering any parallel data, denoted by Crossst and mBARTst;
- models trained on both semantic parsing tasks and the MT data, denoted by Crossmt and mBARTmt.

3.2.3 Comparison systems

As comparison systems, we consider the works reported in Table 3, for a description of such approaches and metrics we refer the reader to section 2.2.3 and Procopio et al. 2021.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

	Model	Smatch	Unlabeled	No-WSD	Concepts	Wiki	NER	Reentrancies	Negations	SRL
AMR-2.0	Ge et al. (2019)	74.3	77.3	74.8	84.2	71.3	82.4	58.3	64.0	70.4
	Zhang et al. (2019b)	77.0	80.0	78.0	86.0	86.0	79.0	61.0	77.0	71.0
	Zhou et al. (2020)	77.5	80.4	78.2	85.9	86.5	78.8	61.1	76.1	71.0
	Cai and Lam (2020)	80.2	82.8	80.0	88.1	86.3	81.1	64.6	78.9	74.2
	Xu et al. (2020a)	80.2	83.7	80.8	87.4	75.1	85.4	66.5	71.5	78.9
	SPRING$_{bart}$	83.8	86.1	84.4	**90.2**	84.3	**90.6**	70.8	74.4	79.6
	SPRING	**84.5**	**86.7**	**84.9**	89.6	**87.3**	83.7	**72.3**	**79.9**	79.7
	$Cross_{st}$	70.7	75.1	71.3	80.3	75.7	78.9	58.9	58.6	68.5
	$Cross_{mt}$	78.1	82.1	78.7	85.1	80.6	85.0	66.6	71.5	75.2
	$Cross_{mt}^{ft}$	79.5	83.2	80.2	86.5	80.7	85.9	68.4	71.8	77.1
	$mBART_{st}$	81.7	85.1	82.1	88.4	81.9	90.3	70.7	73.4	79.7
	$mBART_{mt}$	81.9	85.3	82.3	88.5	81.0	89.4	71.0	75.3	80.0
	$mBART_{mt}^{ft}$	82.3	85.7	82.8	88.9	82.3	89.3	71.5	73.7	**80.4**
AMR-3.0	SPRING	**83.0**	**85.4**	**83.5**	**89.8**	**82.7**	**87.2**	70.4	**73.0**	78.9
	$Cross_{mt}^{ft}$	78.1	81.9	78.7	85.3	76.6	81.3	67.6	68.5	76.2
	$mBART_{st}$	80.0	83.2	80.5	86.6	77.2	86.3	70.0	68.5	78.4
	$mBART_{mt}^{ft}$	81.2	84.4	81.6	88.4	77.7	86.5	**71.1**	69.7	**79.7**

Table 3: Smatch and fine-grained results on AMR-2.0 (top) and AMR-3.0 (bottom).

	Model	DE	ES	IT	ZH
HT	AMREAGER	39.0	42.0	43.0	35.0
	AMREAGER$_{MT}$	57.0	60.0	58.0	50.0
	XL-AMR$_{\emptyset}$	32.7	39.1	37.1	25.9
	XL-AMR	53.0	58.0	58.1	41.5
	Sheth et al. (2021)	62.7	67.9	67.4	–
	$Cross_{mt}$	60.8	62.9	63.2	51.8
	$Cross_{mt}^{ft}$	61.8	63.7	64.1	52.6
	$mBART_{st}$	54.8	60.4	63.6	47.8
	$mBART_{mt}$	66.3	69.0	69.8	55.4
	$mBART_{mt}^{ft}$	65.8	69.2	69.6	54.8
	$mBART_{mt}^{ft}$ + AP	**69.8**	**72.4**	**72.3**	**58.0**
MT	Sheth et al. (2021)	66.9	69.6	71.0	–
	$mBART_{mt}^{ft}$ + AP	**73.3**	**73.9**	**73.4**	**64.9**

Table 4: Smatch scores on cross-lingual AMR parsing for both human (top, HT) and machine (bottom, MT) translations of the test set.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

3.3 Results

In this section, we illustrate the results obtained by SGL in two different settings, namely in AMR parsing and in cross-lingual AMR parsing.

3.3.1 AMR parsing

We report the Smatch and fine-grained scores that SGL and its current state-of-the-art alternatives attain on AMR-2.0 in Table 3 (top). Among the competing systems considered, for Bevilacqua et al. (2021) we report their BART-powered baseline (SPRINGbart) and their best performing model (SPRING).

We report the Smatch and fine-grained scores that SGL and its current state-of-the-art alternatives attain on AMR-2.0 in Table 3 (top). Among the competing systems considered, for Bevilacqua et al. (2021) we report their BART-powered baseline (SPRINGbart) and their best performing model (SPRING).

3.3.2 Cross-lingual AMR parsing

We now show the performances of SGL on cross-lingual AMR parsing in terms of Smatch score over Chinese (ZH), German (DE), Italian (IT) and Spanish (ES). For comparison, we report the results of the systems proposed by Damonte and Cohen (2018, AMREAGER), Blloshmi et al. (2020, XL-AMR) and Sheth et al. (2021); along with their best systems, we also show the strongest MT baseline reported in Damonte and Cohen (2018, AMREAGERMT) and the zero-shot configuration explored in Blloshmi et al. (2020, XL-AMRØ). We report the results of cross-lingual AMR parsing in Table 4.

3.3.3 UCCA parsing

We report in Table 5 the performance of SGL on UCCA parsing. We compare our approach with the original multi-task baseline (Oepen et al., 2019) and 3 transition-based parsers; in particular, we report the score of Che et al. (2019), the system that ranked first in both all-framework and UCCA parsing.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Model	Type	Score
Oepen et al. (2019)	multi-task	41.0
Hershcovich and Arviv (2019)	single-task	82.1
Hershcovich and Arviv (2019)	multi-task	73.1
Che et al. (2019)	multi-task	**82.6**
$Cross_{st}$	single-task	55.7
$Cross_{mt}$	multi-task	72.0
$Cross_{mt}^{ft}$	multi-task	75.1
$mBART_{st}$	single-task	77.0
$mBART_{mt}$	multi-task	79.9
$mBART_{mt}^{ft}$	multi-task	76.9

Table 5: UCCA results on The Little Prince

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Figure 3 - Equivalent sentences in different languages (left) and their BMR graph (right) with multiple possible lexicalizations. Figure extracted from Navigli et al. 2021, AAAI

4 BabelNet Meaning Representation

One of the biggest issues of semantic parsing formalisms is that they are not entirely semantic, meaning that many of their representation components are purely lexical and therefore bound to a specific language. Here we focus on AMR which, among the various approaches to semantic parsing, is the most popular one and is also widely applied in many downstream tasks, e.g. Machine Translation (MT)[2], in which such semantic representations can be used as interlingua (Richens 1958), dialogue, in which the human-computer interaction can be facilitated and supported by exploiting formal semantic representations of sentences, and many other tasks such as Information Extraction (IE) and paraphrase detection. Nonetheless, AMR also shows the aforementioned drawbacks. In fact, it inherently focuses on English. Furthermore, AMR is not anchored, i.e. there is no connection between the single tokens in

[2] We think that the BMR approach should be explored in terms of disambiguation biases since it might contribute significantly to reducing disambiguation biases affecting MT (Campolungo et al. 2022).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

the sentence under consideration and the nodes in the semantic graph obtained.

To address the aforementioned limitations, we propose a novel fully-semantic language independent approach called BabelNet Meaning Representation (BMR).

The BMR approach (BMR, Navigli et al., 2021, Martinez Lorenzo et al. 2022) aims at producing a fully semantic representation of a sentence by leveraging a directed acyclic graph and removing language-specific constraints. As illustrated in Figure 3, this is done by exploiting wide-coverage multilingual lexical-semantic resources detailed in the next subsection. Compared to the current state of the art, and specifically AMR, BMR provides various advantages: it effectively deals with multiword expressions and idioms, and it also encodes grammatical categories such as number, tense and aspect. Figure 4 illustrates how BMR produces better results than those achieved by AMR, since it is capable of encoding grammatical categories of the words *friends* and *tolerate* (see section 4.2 which describes the BMR approach in detail).

Thanks to its design, the BMR approach not only allows for an effective scalability to multiple languages, but it also paves the way for cross-modal representations including images, videos, speech and sound. Furthermore, BMR represents a novel semantic formalism which can be effectively used as an interlingua.

The BMR source code and data have been released and are currently available at: https://github.com/SapienzaNLP/bmr[3].

[3] The code used to convert the data from AMR to BMR makes use of sensitive data. Therefore, to use it, please provide us proof of the AMR licence issued by LDC at https://catalog. ldc.upenn.edu/LDC2020T02.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

```
—Original Sentence—
My friend did not tolerate his father's behaviour

—AMR—
(t / tolerate-01          :polarity -
         :ARG0 (p / person
                  :ARG0-of (h / have-rel-role-91
                           :ARG1 (i / i)
                           :ARG2 (f / friend)))
         :ARG1 (b / behave-01
                  :ARG0 (p2 / person
                           :ARG0-of (h2 / have-rel-role-91
                                    :ARG1 p
                                    :ARG2 (f2 / father)))))

—BMR*—
(t /  bn:00082138v      :polarity -      :timing -
         :agent (f /  bn:00036538n
                  :related (i / i))
         :theme (b / bn:00009656n
                  :related (f2 / bn:00009616n
                           :related f)))

—AMR Generated Sentence—
My friends do not tolerate the behavior of my father.

—BMR* Generated Sentence—
My friend did not put up with the behaviour of his father.
```

Figure 4 - AMR vs BMR

4.1 Lexical-semantic resources used in BMR

In order to obtain language-independent representations, the BMR approach relies on two wide-coverage lexical-semantic resources:

1. **BabelNet** (Navigli and Ponzetto 2010, 2012), a multilingual encyclopaedic dictionary and semantic network which covers approximately 500 languages. BabelNet is a

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

merger of computational resources such as WordNet (Miller 1995) and Wikipedia, available at https://babelnet.org.

2. **VerbAtlas** (Di Fabio et al., 2019), a manually-curated inventory of predicates organized into verbal frames, introduced in the ELEXIS project (see D3.4). VerbAtlas exploit several semantic roles which are illustrated in Table 6. The goal behind this resource is to gather all Wordnet's verbal synsets into semantically-coherent frames. VerbAtlas is available at http://verbatlas.org.

Agent	Material
Attribute	Patient
Beneficiary	Product
Cause	Purpose
Co-Agent	Recipient
Co-Patient	Result
Co-Theme	Source
Destination	Stimulus
Experiencer	Theme
Extent	Time
Goal	Topic
Instrument	Value
Location	

Table 6 - VerbAtlas semantic roles

4.2 The BMR approach

In this section, we describe the BMR approach, illustrating its advantages when compared to AMR.

4.2.1 Node Merging

Multiword expressions and idioms are rendered word by word in AMR, using node composition. Nevertheless, such an approach is not appropriate for an interlingual representation. In fact, often the meaning of multiword expressions and idioms cannot be derived compositionally, i.e. inferred from the meanings of its individual words. Therefore, in

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

BMR we exploit available BabelNet synsets to identify the meaning of a multiword expression or idiom, thus representing it with a single node.

To merge nodes, we first identify the words or multiword expressions represented by several nodes in the AMR graph. To do this, we lemmatize the original sentences in AMR 3.0 using the 3.1 version of the SpaCy software library (Honnibal and Johnson, 2015). Then, for each sentence, we search for the longest concatenations of lemmas matching a BabelNet synset lexicalization in BabelNet 5.0. Once the expressions have been identified, we use the automatic AMR aligner of Flanigan et al. (2014) to get the alignments between the tokens in the original sentence (and, consequently, the identified words and multiwords) and the graph nodes.

The automatic identification of multiwords can lead to poor node merging choices which, in turn, can result in wrong sense attributions. For instance, in the sentence “the rest of the world knows the same”, the multiword rest of the world is identified, even though its only meaning in BabelNet is that of “a team of players from many countries”, which is clearly not appropriate in the reported context. To address this issue, we asked our expert linguist to manually inspect all of the automatically detected multiword instances within the AMR 3.0 dataset in order to maintain, modify or delete them.

4.2.2 Number, Tense and Aspect

Although AMR can encode textual information in its semantic structure, its formalism does not consider word components that are crucial to understand meaning, and that languages convey thanks to grammatical categories such as number, tense and aspect. Importantly, the need for the integration of such information into semantic parsing has been already mentioned in the literature (Donatelli et al., 2018; Bonial et al., 2019). To achieve this goal, we rely on SpaCy in order to retrieve the Penn Treebank part-of-speech tags (Marcus et al., 1993), which inherently provide information regarding number, tense, and aspect, for all the words and multiword expressions aligned with a node in the graphs. In practice, we account

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

for tense by enriching each verbal node with the semantic role :timing showing a value of + or − to indicate events that will take place in the future or that happened in the past, respectively. Similarly, we handle plurality of the nominal nodes by adding the :quantity relation followed by a + value. Lastly, we account for aspect by adding the relation :ongoing followed by a + mark to verbal nodes expressing the imperfective aspect (ongoing or usual actions).

4.2.3 Graph Conversion

Finally, using the multiwords and the alignments derived from the previous steps, we navigate the AMR graphs bottom-to-top and collapse together nodes referring to the same word or multiword expression (i.e., first reducing nodes closer to the graph leaves and then moving towards the graph root). As a result, we move from the original figure of 936,769 nodes of AMR 3.0 to 828, 483 in BMR 1.0, reducing the graph density by a notable 11.6%.

4.2.4 Graph Disambiguation

A crucial desideratum of interlingual representation of meaning is to be fully-linked to a (possibly multilingual) inventory of meanings. Therefore, in order to make nodes in BMR graphs language-independent, we enhance them with BabelNet synsets information.

Our last step is to add the disambiguation information to AMR 3.0 so as to finalize the conversion to BMR 1.0. To this end, we exploit different strategies:

a) we use the mapping from VerbAtlas frames to BabelNet synsets to assign word senses to nodes based on their lemmas;

b) we exploit the Wikipedia page information featured in AMR nodes which represents named entities to retrieve the corresponding synset which BabelNet identifies that page with;

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

c) we disambiguate the nodes without word senses using ESCHER (Barba et al., 2021), a state-of-the-art system for Word Sense Disambiguation, i.e., the task of automatically assigning a meaning to a word in context (Bevilacqua et al., 2021).

As a result of this process, we are able to assign a BabelNet synset to AMR content nodes in approximately 92% of cases (i.e., nodes aligned with content words), with 42,549 fully disambiguated graphs out of 59,255.

4.3 The BMR 1.0 dataset

In order to create the BMR 1.0 dataset, we use the mapping provided by Di Fabio et al. (2019) which provides links from VerbAtlas frames and arguments to PropBank. Specifically, we replace the original frames and semantic roles in the AMR 3.0 dataset with those of VerbAtlas. However, since this mapping is incomplete, e.g. predicates that OntoNotes labels as verbal, and non-verbal predicates, we asked an expert linguist to create a mapping between PropBank and VerbAtlas for the missing verbal predicates, and, with respect to the remaining links, to map them to BMR adapting previous semantic roles and creating new ones to better accommodate their argument structures. Table 7 shows all BMR semantic roles.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

BMR	AMR	VerbAtlas	examples
agent	-	AGENT	marry$_V$>man$_N$; suggest$_V$>mum$_N$
appearance	mod	-	complexion$_N$>pale$_A$; road$_N$>snowy$_A$
cause	cause	CAUSE/STIMULUS	issue$_N$>society$_N$; kill$_V$>evil$_N$
co-agent	accompanier	CO_AGENT	*(same as agent)*
co-patient	accompanier	CO_PATIENT	*(same as patient)*
co-theme	accompanier	CO_THEME	*(same as theme)*
composition	consist-of	MATERIAL	cup$_N$>metal$_N$; army$_N$>idiot$_N$
context	location, topic	TOPIC/ATTRIBUTE	boldness$_N$>mission$_N$; excellence$_N$>sports$_N$
cost	cost	ASSET	computer$_N$>euro$_N$; tuition$_N$>free$_A$
experiencer	-	EXPERIENCER	badness$_N$>customer$_N$; react$_V$>sector$_N$
extent	duration, extent	EXTENT	trip$_N$>mile$N$; work$_V$>day$_N$
identity	domain/meaning/role/example	-	boyfriend$_N$>lawyer$_N$; dog$_N$>animal$_N$
instrument	instrument	INSTRUMENT	book$_N$>eye$_N$; pasta$_N$>fork$_N$
interactor	-	-	book$_N$>librarian$_N$; treaty$_N$>actor$_N$
location	location	LOCATION	hotel$_N$>beach$_N$; water$_N$>jar$_N$
membership	employed-by/have-org-role-91	-	friend$_N$>company$_N$; nurse$_N$>hospital$_N$
part	part/subset/superset	-	book$_N$>beginning$_N$; car$_N$>wheel$_N$
patient	-	PATIENT	dry$_V$>skin$_N$; kick$_V$>ball$_N$
physical_prop	mod	-	hand$_N$>cold$_A$; train$_N$>fast$_A$
property	poss	-	patient$_N$>health$_N$; professor$_N$>book$_N$
purpose	purpose	PURPOSE	treaty$_N$>extradition$_N$; book$_N$>coloring$_N$
quality	mod	-	book$_N$>available$_A$; city$_N$>beautiful$_A$
quantity	quant	VALUE	cow$_N$>few$_A$; degree$_N$>42
related	have-rel-role-91	-	city$_N$>suburb$_N$; father$_N$>son$_N$
result	-	PRODUCT, RESULT	become$_V$>farmer$_N$; revolution$_N$>overthrow$_V$
source	source	SOURCE	book$_N$>author$_N$; trip$_N$>San Diego$_N$
target	beneficiary/destination/direction	BENEFICIARY/DESTINATION/GOAL/RECIPIENT	brutality$_N$>person$_N$; food$_N$>dog$_N$
theme	-	THEME	read$_V$>book$_N$; require$_V$>vitamin$_N$
timing	time	TIME	marry$_V$>then$_R$; struggle$_N$>current$_A$
url	hyperlink-91	-	website$_N$>https://verbatlas.org/

Table 7 - BMR semantic roles

4.4 Experimental setup

In order to show the impact of BMR, we evaluate our novel approach in different tasks, namely text generation, semantic parsing and translation through semantic parsing. To this end, we use different datasets which we describe in the next subsection.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Language	English (EN)				German (DE)				Italian (IT)				Spanish (ES)			
Model	AMR	AMR+	BMR	BMR*	AMR	AMR+	BMR	BMR*	AMR	AMR+	BMR	BMR*	AMR	AMR+	BMR	BMR*
BLEU	44.8	49.8	**50.7**	45.7	23.2	24.3	**24.8**	22.2	29.0	31.3	**31.4**	29.1	34.6	36.8	**37.3**	35.5
chrF++	73.4	76.0	**76.3**	72.1	55.8	57.0	**57.1**	54.7	60.7	62.1	**62.2**	60.0	64.0	65.2	**65.5**	63.7
METEOR	42.2	43.9	**44.3**	42.4	25.4	26.4	**26.4**	25.3	28.9	30.4	**30.5**	29.2	32.4	33.5	**33.7**	32.8
Rouge-L	68.2	71.7	**72.8**	69.7	49.3	50.7	**51.1**	49.7	51.9	54.2	**54.3**	52.4	57.4	60.9	**61.0**	59.8

Table 8 - Results for text generation

4.4.1 Datasets

In order to perform our experiments, we use *AMR 3.0* as well as the *BMR 1.0* dataset previously described. Additionally, we use *AMR+*, which includes a set of enhancements applied to AMR 3.0., and *BMR** which does not include lemmas in the nodes and limits to the BabelNet synset ID.

Importantly, all datasets share the same sentences, the only difference is their graph representation. We use the same training, development and test split as those in AMR 3.0. Additionally, for each set we obtain language-specific sets by translating the sentences using Machine Translation systems into German (DE), Italian (IT) and Spanish (ES) as done by Blloshmi et al. 2020. As far as testing is concerned, we use the 1,371 parallel sentences derived from the Abstract Meaning Representation 2.0 - Four Translations[4].

4.4.2 Tasks

In order to show the impact of BMR in semantic parsing, we evaluate our new approach in the following tasks: i) text generation, ii) semantic parsing, and iii) translation through semantic parsing.

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

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Model	BLE	CH+	MET	R-L
$\textbf{AMR}_{EN}$	44.8	73.4	42.2	68.2
$\textbf{AMR}_{REL}$	44.9	73.8	42.4	68.7
$\textbf{AMR}_{NOD}$	45.5	73.8	42.5	68.9
$\textbf{AMR}_{NUM}$	46.9	74.3	42.9	69.8
$\textbf{AMR}_{TEN}$	47.6	74.9	43.0	70.2
$\textbf{AMR}_{NT}$	49.0	75.4	43.5	71.1
$\textbf{AMR}+_{EN}$	49.8	76.0	43.9	71.7
$\textbf{BMR}_{EN}$	**50.7**	**76.3**	**44.3**	**72.8**

Table 9 - Ablation study - Impact of disambiguation in text generation

4.4.2.1 Text Generation

The text generation task consists in converting graph-based meaning representation into plain text. This task aims at investigating the effectiveness of the BMR approach when generating texts in many languages. In this case, the metrics which we used are: BLEU, chrF++, METEOR and Rough-L. Results obtained in this experiment are reported in Table 8. As illustrated, BMR allows us to achieve the best performance, even outperforming AMR+, which explicitly shows the benefits of exploiting disambiguated representations. Remarkably, BMR* attains competitive results, surpassing AMR without leveraging any lemma information in the nodes. Moreover, we conduct an ablation study to assess the impact of disambiguation and report our results in Table 9.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Model	SMT.	unlab.	noWSD	conc.	NER	neg.	reent.
AMR_{EN}	**82.1**	85.3	**82.6**	88.0	**89.0**	67.0	**73.0**
AMR+_{EN}	**82.1**	**85.8**	82.1	**90.0**	**89.0**	**75.0**	70.0
BMR_{EN}	78.6	82.2	78.6	82.0	83.0	63.0	65.0
BMR*_{EN}	78.7	82.2	78.6	82.0	83.0	63.0	65.0

Table 10 - Results for the semantic parsing task

4.4.4.2 Semantic Parsing

Semantic parsing is the computational task of producing a formal graph-based representation of raw text as described above. We evaluate this task, comparing the produced graph of the model against gold graphs using the SMATCH metric, which aims at creating a mapping 1-to-1 from the semantic unit of each graph and calculating the F1-score (see Table 10). This task can be used to assess the ability of generating each graph representation. In Table 10, we report the results of the semantic parsing experiment. As can be seen, BMR and BMR* obtain the worst performance. Furthermore, we observe that AMR+ obtains the same performance as AMR 3.0, which seems to confirm our intuition, based on which the drop in performance of BMR might be due to the disambiguation.

4.4.4.3 Translation through semantic parsing

Translation through semantic parsing allows us to assess the effectiveness of BMR as interlingua by merging the aforementioned tasks. BMR almost always obtains the best results in the various scenarios, even if the semantic parsing task is more challenging since, in that case, BMR has to disambiguate also the concepts (see Table 11). Therefore, even though the drop in performance in the semantic parsing task of BMR, the benefits of the disambiguation are more important in the final results.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Pairs	EN – EN				EN – DE				EN – IT				EN – ES			
Model	**AMR**	**AMR+**	**BMR**	**BMR***	**AMR**	**AMR+**	**BMR**	**BMR***	**AMR**	**AMR+**	**BMR**	**BMR***	**AMR**	**AMR+**	**BMR**	**BMR***
BLEU	45.3	49.3	**50.1**	45.1	23.0	**25.1**	24.4	22.8	29.0	30.7	**30.9**	29.1	34.0	36.5	**36.6**	35.4
chrF++	73.5	75.2	**75.4**	71.4	55.6	**56.8**	56.1	54.1	60.2	**61.4**	61.2	59.8	63.3	64.6	**64.8**	63.3
METEOR	42.3	43.5	**43.7**	41.9	25.4	**26.4**	26.0	25.1	28.7	29.8	**29.9**	29.1	32.0	**33.3**	33.2	32.6
Rouge-L	68.8	71.8	**73.0**	69.5	49.6	**50.8**	**50.8**	49.3	51.9	**53.5**	53.2	52.6	57.2	61.1	**62.2**	59.7

Table 11 - Results obtained in translation through semantic parsing (multilingual setting).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

4 Conclusion

In this deliverable, we described the research activities carried out in WP3, task 3.2 regarding multilingual semantic parsing, the task of creating formal graph-based representations starting from raw text.

The ambitious goal of task 3.2 was to develop innovative approaches to semantic parsing which address the high annotation costs required by current supervised approaches and, most importantly, to effectively scale to multiple languages. To this end, building upon our previous deliverable D3.4 in which we introduced VerbAtlas (Di Fabio et al. 2019) and XL-AMR (Blloshmi et al. 2020), an effective approach towards the achievement of the aforementioned goals, in this second deliverable we introduced three novel approaches to semantic parsing which enable multilinguality and fully semantic parsing with high performances.

As a suggestion for future research works, we strongly recommend to further explore this promising research area which might prove to be crucial in a wide range of NLP tasks. Specifically, we encourage the creation of multilingual models for semantic parsing covering an even higher number of languages, e.g. Asian languages, and to study the impact of semantic parsing on downstream tasks.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

References

Abend, O., & Rappoport, A. (2013, August). Universal conceptual cognitive annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 228-238).

Abzianidze, L., Bjerva, J., Evang, K., Haagsma, H., van Noord, R., Ludmann, P., ... & Bos, J. (2017). The Parallel Meaning Bank: Towards a Multilingual Corpus of Translations Annotated with Compositional Meaning Representations. *EACL 2017*, 242.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., ... & Schneider, N. (2013, August). Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse* (pp. 178-186).

Barba, E., Pasini, T., & Navigli, R. (2021, June). ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4661-4672).

Bevilacqua, M., Blloshmi, R., & Navigli, R. (2021, May). One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 14, pp. 12564-12573).

Blloshmi, R., Tripodi, R., & Navigli, R. (2020, November). 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)* (pp. 2487-2500).

Bonial, C. N., Donatelli, L., Ervin, J., & Voss, C. R. (2019). Abstract meaning representation for human-robot dialogue. *Proceedings of the Society for Computation in Linguistics*, *2*(1), 236-

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

246.

Campolungo, N., Martelli, F., Saina, F., & Navigli, R. (2022, May). DIBIMT: A Novel Benchmark for Measuring Word Sense Disambiguation Biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4331-4352).

Cai, D., & Lam, W. (2019, November). Core Semantic First: A Top-down Approach for AMR Parsing. 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)* (pp. 3799-3809).

Cai, D., & Lam, W. (2020, July). AMR Parsing via Graph-Sequence Iterative Inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1290-1301).

Donatelli, L., Regan, M., Croft, W., & Schneider, N. (2018, August). Annotation of tense and aspect semantics for sentential AMR. In *Proceedings of the joint workshop on linguistic annotation, multiword expressions and constructions (LAW-MWE-CxG-2018)* (pp. 96-108).

Marco Damonte and Shay B. Cohen. 2018. Crosslingual Abstract Meaning Representation parsing. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1146–1155*, New Orleans, Louisiana. Association for Computational Linguistics.

Di Fabio, A., Conia, S., & Navigli, R. (2019, November). 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*

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 627-637).

Flanigan, J., Thomson, S., Carbonell, J. G., Dyer, C., & Smith, N. A. (2014, June). A discriminative graph-based parser for the abstract meaning representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1426-1436).

Ge, D., Li, J., Zhu, M., & Li, S. (2019, January). Modeling Source Syntax and Semantics for Neural AMR Parsing. In *IJCAI* (pp. 4975-4981).

Goodman, M. W. (2020, July). Penman: An open-source library and tool for AMR graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 312-319).

Hajic, J., Hajicová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., ... & Žabokrtský, Z. (2012, May). Announcing prague czech-english dependency treebank 2.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 3153-3160).

Hajic, J., Bojar, O., & Uresova, Z. (2014, August). Comparing Czech and English AMRs. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing* (pp. 55-64).

Hershcovich, D., Abend, O., & Rappoport, A. (2018, July). Multitask Parsing Across Semantic Representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 373-385).

Hershcovich, D., & Arviv, O. (2019, November). TUPA at MRP 2019: A multi-task baseline system. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning* (pp. 28-39).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Honnibal, M., & Johnson, M. (2015, September). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1373-1378).

Konstas, I., Iyer, S., Yatskar, M., Choi, Y., & Zettlemoyer, L. (2017, July). Neural AMR: Sequence-to-Sequence Models for Parsing and Generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 146-157).

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020, July). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871-7880).

Lindemann, M., Groschwitz, J., & Koller, A. (2019, July). Compositional Semantic Parsing across Graphbanks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4576-4585).

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Han, J. (2019, September). On the Variance of the Adaptive Learning Rate and Beyond. In *International Conference on Learning Representations*.

Mager, M., Astudillo, R. F., Naseem, T., Sultan, M. A., Lee, Y. S., Florian, R., & Roukos, S. (2020, July). GPT-too: A Language-Model-First Approach for AMR-to-Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1846-1852).

Marcus, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Using Large Corpora*, *273*.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Martinez Lorenzo, A. C., Maru, M., & Navigli, R. (2022, May). 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)* (pp. 1727-1741).

Martelli, F., Navigli, R., Krek, S., Kallas, J., Gantar, P., Koeva, S., Nimb, S., Sandford Pedersen, B., Olsen, S., Langemets, M., Koppel, K., Üksik, T., Dobrovoljc, K., Ureña-Ruiz, R., Sancho-Sánchez, J., Lipp, V., Váradi, T., Győrffy, A., László, S., Quochi, V., Monachini, M., Frontini, F., Tiberius, C., Tempelaars, R., Costa, R., Salgado, A., Čibej, J., & Munda, T. (2021a). Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages. In *Proceedings of eLex 2021.*

Martelli, F., Kalach, N., Tola, G., & Navigli, R. (2021b, August). SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 24-36).

Maru, M., Scozzafava, F., Martelli, F., & Navigli, R. (2019, November). SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations. 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)* (pp. 3534-3540).

May, J., & Priyadarshi, J. (2017, August). Semeval-2017 task 9: Abstract meaning representation parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 536-545).

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Naseem, T.; Shah, A.; Wan, H.; Florian, R.; Roukos, S.; and Ballesteros, M. 2019. Rewarding Smatch: Transition-Based AMR Parsing with Reinforcement Learning. In Proc. of ACL 2019, 4586–4592. Florence, Italy

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, *193*, 217-250.

Stephan Oepen, Omri Abend, Jan Hajic, Daniel Hershcovich, Marco Kuhlmann, Tim O'Gorman, Nianwen Xue, Jayeol Chun, Milan Straka, and Zdenka Uresova. 2019. MRP 2019: Cross-framework meaning representation parsing. In Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning, pages 1–27, Hong Kong. Association for Computational Linguistics.

Oepen, S., & Lønning, J. T. (2006, May). Discriminant-based MRS banking. In *LREC* (pp. 1250-1255).

Procopio, L., Tripodi, R., & Navigli, R. (2021a, June). SGL: Speaking the graph languages of semantic parsing via multilingual translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 325-337).

Procopio, L., Barba, E., Martelli, F., & Navigli, R. (2021b). MultiMirror: Neural Cross-lingual Word Alignment for Multilingual Word Sense Disambiguation. In *IJCAI* (pp. 3915-3921).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Richens, R. H. (1958). Interlingual machine translation. *The Computer Journal*, *1*(3), 144-147.

Tedeschi, S., Martelli, F., & Navigli, R. ID10M: Idiom Identification in 10 Languages.

Tiedemann, J., & Thottingal, S. (2020, November). OPUS-MT--Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

van Noord, R., & Bos, J. (2017). Neural Semantic Parsing by Character-based Translation: Experiments with Abstract Meaning Representations. *Computational Linguistics in the Netherlands Journal*, *7*, 93-108.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, T., Wan, X., & Yao, S. (2021, January). Better AMR-to-text generation with graph structure reconstruction. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence* (pp. 3919-3925).

White, A. S., Reisinger, D., Sakaguchi, K., Vieira, T., Zhang, S., Rudinger, R., ... & Van Durme, B. (2016, November). Universal decompositional semantics on universal dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1713-1723).

Yang, Y., Huang, L., & Ma, M. (2018). Breaking the Beam Search Curse: A Study of (Re-) Scoring Methods and Stopping Criteria for Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3054-3059).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Yao, S., Wang, T., & Wan, X. (2020, July). Heterogeneous graph transformer for graph-to-sequence learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7145-7154).

Yarowsky, D., & Ngai, G. (2001). Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Zhao, Y., Chen, L., Chen, Z., Cao, R., Zhu, S., & Yu, K. (2020, July). Line graph enhanced AMR-to-text generation with mix-order graph attention networks. In *Proceedings of the 58th Annual meeting of the association for computational linguistics* (pp. 732-741).

Zhang, S., Ma, X., Duh, K., & Van Durme, B. (2019a, July). AMR Parsing as Sequence-to-Graph Transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 80-94).

Zhang, S., Ma, X., Duh, K., & Van Durme, B. (2019, January). Broad-Coverage Semantic Parsing as Transduction. In *EMNLP/IJCNLP (1)*.

Zhu, J., Li, J., Zhu, M., Qian, L., Zhang, M., & Zhou, G. (2019, November). Modeling Graph Structure in Transformer for Better AMR-to-Text Generation. 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)* (pp. 5459-5468).

Zhou, Q.; Zhang, Y.; Ji, D.; and Tang, H. 2020. AMR Parsing with Latent Structural Information. In Proc. of ACL 2020, 4306–4319. Online.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.