# Assessing the Cross-linguistic Utility of Abstract Meaning Representation

Shira Wein
Georgetown University

Nathan Schneider
Georgetown University

*Semantic representations capture the meaning of a text. Abstract Meaning Representation (AMR), a type of semantic representation, focuses on predicate-argument structure and abstracts away from surface form. Though AMR was developed initially for English, it has now been adapted to a multitude of languages in the form of non-English annotation schemas, cross-lingual text-to-AMR parsing, and AMR-to-(non-English) text generation. We advance prior work on cross-lingual AMR by thoroughly investigating the amount, types, and causes of differences which appear in AMRs of different languages. Further, we compare how AMR captures meaning in cross-lingual pairs versus strings, and show that AMR graphs are able to draw out fine-grained differences between parallel sentences. We explore three primary research questions: (1) What are the types and causes of differences in parallel AMRs? (2) How can we measure the amount of difference between AMR pairs in different languages? (3) Given that AMR structure is affected by language and exhibits cross-lingual differences, how do cross-lingual AMR pairs compare to string-based representations of cross-lingual sentence pairs? We find that the source language itself does have a measurable impact on AMR structure, and that translation divergences and annotator choices also lead to differences in cross-lingual AMR pairs. We explore the implications of this finding throughout our study, concluding that, while AMR is useful to capture meaning across languages, evaluations need to take into account source language influences if they are to paint an accurate picture of system output, and meaning generally.*

## 1 Introduction

Semantic representations, which reflect the meaning of a text, are an important tool for downstream natural language processing tasks which rely on meaning inference. Semantic representations can be designed to capture specific aspects of meaning. Abstract Meaning Representation (AMR; Banarescu et al. 2013), for example, captures "who does what to whom," and focuses on predicate-argument structure, abstracting away from morphosyntactic details such as word order by encoding the core meaning of sentence or phrase as a directed, rooted graph. AMR has been widely studied in the NLP literature with regard to text-to-AMR parsing, AMR-to-text generation, and downstream applications (described in §2.1).

Banarescu et al. (2013) devised AMR only for English, disclaiming any intentions of using it as an interlingua (language-neutral semantics that could bridge across languages, such as in the paradigm of interlingual machine translation; Richens 1958; Dorr, Hovy, and Levin 2004). Nevertheless, versions of AMR have since been developed

Action editor: Nianwen Xue. Submission received: 1 May 2023; revised version received: 14 September 2023; accepted for publication: 27 October 2023.

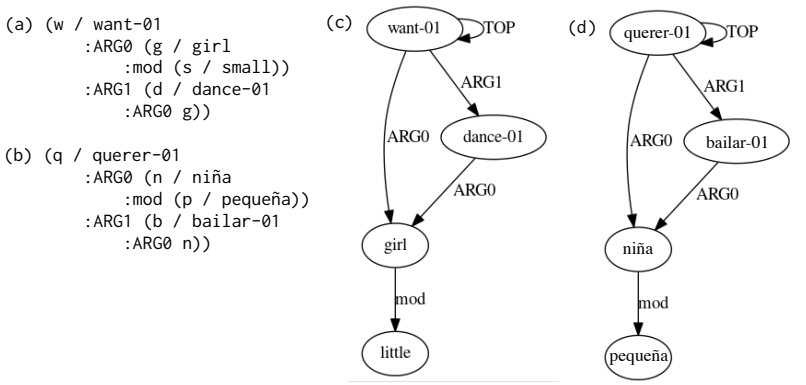



Figure 1: English (a) and Spanish (b) AMRs for the sentence "the little girl wants to dance" *la niña pequeña quiere bailar* in PENMAN (text-based) notation (Wein et al. 2022a), as well as the English (c) and Spanish (d) graph-based AMR illustrations.

to annotate text in many other languages (see §2 for a review). For example, Figure 1 shows Abstract Meaning Representations for parallel sentences in English and Spanish. Despite established literature on semantic differences in parallel sentences (which can arise due to syntactic differences in the languages or translation choices; Dorr 1990, 1994), attempts to account for the effect of the sentence's source language on AMR structure have been limited.

Previous work has characterized differences in AMR graphs ("AMRs") across languages (Urešová, Hajič, and Bojar 2014; Xue et al. 2014). In this article, we advance investigations into AMR as an interlingua by thoroughly assessing whether AMR can comprehensively reflect the meaning of languages other than English, and how this compares to more surface-level, non-hierarchical representations of sentence meaning.

To assess AMR as tool for capturing meaning cross-lingually, we first set out to quantify the amount of difference between AMRs of parallel sentences/cross-lingual AMR pairs (§3), by assessing the effect of source language (the language of the sentence parsed into an AMR) on AMR structure, and the degree of language effect by individual language. §3 investigates **RQ1: how can we measure the amount of difference between parallel AMRs, and using this measurement, what is the extent of the difference in parallel AMRs?** To measure this difference, we propose transferring the non-English tokens into English, leaving only underlying AMR graph structure to compare. Next, we determine the similarity between the underlying graph structures using Smatch, and uncover the critical finding that the language itself does have a sizable effect on AMR structure.

We then perform a finer-grained analysis (§4), introducing a novel taxonomy for annotating the types and causes of these language-based divergences. Applying this schema to a small set of divergent Spanish-English AMR pairs, we show that differences in parallel AMRs arise for three reasons: (1) translation choice, (2) annotation choice, and, importantly, (3) inherent differences between the languages. §4 answers **RQ2: why do AMRs for parallel sentences differ?**

Finally, in §5, we compare how AMR captures meaning of a sentence against string-level semantics (looking solely at the string/tokens, as judged by a human or machine, without using a symbolic meaning representation as an intermediary), for cross-lingual

sentences. Building on our findings from §3 and §4, which showed that AMRs capture differences encoded by source language, in §5 we explore **RQ3: how are language-based divergences captured at the AMR-level versus at the string-level?** We find that AMRs capture a finer-grained level of cross-lingual divergence than is able to be observed at the string-level.

This article incorporates content from prior publications: Wein et al. (2022b) in §3.1, a substantially expanded version of Wein and Schneider (2021) in §4, Wein and Schneider (2022) in §5.2, and Wein, Wang, and Schneider (2023) in §5.3. Here we synthesize this work and expand on it with new threads of investigation in the form of four supplementary experiments.[1]

Our findings provide critical insight into the applicability of AMR to cross-lingual settings and as a tool for capturing meaning across languages. What we uncover has implications for the broader study of cross-lingual meaning representations, and allows us to interpret claims about cross-linguistic phenomena such as universality. Further, our work elucidates the relevance of AMR (and other semantic representations) to cross-lingual applications, showing that it is necessary to consider the source language when extending AMR to non-English languages and domains/applications. Specifically, in order to effectively capture the meaning of non-English languages, we need to account for the effect of language itself on AMR structure.

## 2 Background

First we review the origins of the Abstract Meaning Representation schema (§2.1), existing AMR corpora (§2.2), the prior cross-lingual investigations of AMR (§2.3), cross-lingual work that has been done for other semantic representations (§2.4), and studies of semantic divergences in both sentence pairs and AMR pairs (§2.5).

### 2.1 Abstract Meaning Representation

The Abstract Meaning Representation (AMR) formalism is a graph-based representation of the meaning of a sentence or phrase. AMRs are rooted, labeled graphs where each node is an instance of a semantic unit. In AMR annotations, nodes reflect entities and events, and the edges are labeled with semantic roles. AMR aims to abstract away from surface details of morphology and syntax in favor of core elements of meaning, such as predicate-argument structure and coreference. With that in mind, sentences with the same meaning (and content word vocabulary) should be represented by the same AMR. English AMR annotations are unanchored—the nodes are not explicitly mapped to tokens in the sentence—but the concepts (semantic node labels) largely consist of lemmatized words from the sentence. The root of the AMR is the **focus**, an edge marked with `:argN` is a **core argument role** (where `N` is some number $\geq$0), and any other edge (e.g., `:opN`, `:domain`, `:manner`) is a **non-core role**.

Annotation of AMR is lightweight as it does not represent morphology, articles, or tense, but does require a fair amount of training. Inter-annotator agreement is measured using Smatch, which calculates semantic overlap between two AMRs (Cai and Knight 2013). Smatch works by inducing an alignment between nodes that maximizes the amount of overlap between the graphs. Specifically, Smatch quantifies the similarity of

1 Specifically, the experiments detailed in §3.2, §3.3, the qualitative analyses in §5.2.4 and §5.2.5, and the additional annotated corpus discussed in §4 are new to this article.

two AMRs by searching for an alignment of nodes between them that maximizes the $F_1$-score of matching (*node1*, *role*, *node2*) and (*node1*, `instance-of`, *concept*) triples common between the graphs.[2]

As a structured representation of meaning, AMR has proven useful in a variety of applications and domains. AMR has been: (1) applied to a range of downstream applications (such as information extraction (Zhang and Ji 2021; Zhang et al. 2021), summarization (Liu et al. 2015; Hardy and Vlachos 2018), and neural machine translation (Song et al. 2019; Li and Flanigan 2022), (2) extended/adapted to fit various domains (Vu, Le Nguyen, and Satoh 2022; Bonial et al. 2020; Abdelsalam et al. 2022; Mansouri, Oard, and Zanibbi 2022), and (3) leveraged in explainability efforts (Opitz et al. 2021; Xu et al. 2021b; Opitz and Frank 2022).

### 2.2 AMR Corpora

Though AMR was originally designed for English (Banarescu et al. 2013; Knight et al. 2014), AMR's abstraction away from morphosyntactic variation lends itself to cross-lingual adaptation by capturing semantic structure shared across languages (Li et al. 2016). Prior work extending AMR to other languages has developed language-specific annotation schemas, which (to varying extents) assess the relevant linguistic features which need be accommodated in the AMR annotations for that language. Additional prior work has also considered the compatibility of AMR for non-English languages.

| Language | Text | Reference |
|---|---|---|
| English | The Little Prince | Banarescu et al. (2013) |
| English | AMR 1.0 (news, discussion forums, etc.) | Knight et al. (2014) |
| English | AMR 2.0 (news, discussion forums, etc.) | Knight et al. (2017) |
| English | AMR 3.0 (news, discussion forums, etc.) | Knight et al. (2021) |
| English | Biomedical articles | Garg et al. (2016) |
| Chinese | The Little Prince | Li et al. (2016) |
| Chinese | Web collection | Extracted from Xue et al. (2013) |
| Spanish | The Little Prince | Migueles-Abraira, Agerri, and Diaz de Ilarraza (2018) |
| Spanish | AMR 2.0 data (news etc.) | Wein et al. (2022a) |
| Portuguese | The Little Prince | Anchiêta and Pardo (2018) |
| Portuguese | News, PropBank.Br | Sobrevilla Cabezudo and Pardo (2019) |
| Vietnamese | The Little Prince | Linh and Nguyen (2019) |
| Korean | ExoBrain | Choe et al. (2020) |
| Turkish | The Little Prince | Azin and Eryiğit (2019) |
| Turkish | The Little Prince | Oral, Acar, and Eryiğit (2022) |
| Persian | The Little Prince | Takhshid et al. (2022) |

Table 1: Gold-annotated AMR corpora by language.

2 Smatch is the original and default metric for comparing two AMR graphs for the same input sentence, but other metrics for AMR graph comparison have been developed—these are discussed in §5.1.

Cross-lingual adaptations of AMR have been developed in a variety of languages, shown in Table 1. AMR parsing experiments have also been performed for Celtic languages (Heinecke and Shimorina 2022) and Indonesian (Ilmy and Khodra 2020).

For English, annotation of *The Little Prince* English was released and described in Banarescu et al. (2013). English AMRs are also featured in AMR 1.0 (Knight et al. 2014), AMR 2.0 (Knight et al. 2017), AMR 3.0 (Knight et al. 2021), and BioAMR (Garg et al. 2016).[3]

AMR 2.0 - Four Translations contains natural language translations of the English AMR 2.0 sentences into four languages: Italian, Spanish, German, and Mandarin Chinese (Damonte and Cohen 2020). The sentences are only translated; this dataset does not provide parallel AMRs designed to closely mirror the target language sentences.

### 2.3 Cross-lingual Explorations of AMR

Abstraction can also create challenges, such that changes are required to the annotation schema to sufficiently account for language variation and pertinent linguistic phenomena in non-English AMR. AMR has been assessed as an interlingua, considering the types of differences which appear across AMR language pairs, for Czech (Urešová, Hajič, and Bojar 2014), Chinese (Xue et al. 2014), and Spanish (Wein and Schneider 2021), in comparison to English. Xue et al. (2014) explore the adaptability of English AMR to Czech and Chinese. They suggest that AMR may be cross-linguistically adaptable because it abstracts away from morphosyntactic differences. Cross-linguistic comparisons between English/Czech and English/Chinese AMR pairs indicate that many pairs align well. However, a comparison between English and Czech AMRs found that only 29 of 100 AMRs shared identical structure, and that key differences arose in event structure, multi-word expressions, and compound nouns (Xue et al. 2014). Also, the compatibility is higher for English and Chinese than for English and Czech.

Urešová, Hajič, and Bojar (2014) describe the types of differences between AMRs for parallel English and Czech sentences, and find that the differences may be either due to convention/surface-level nuances which could be changed in the annotation guidelines, or may be due to inherent facets of the AMR annotation schema. One notable cross-lingual AMR difference is from the appearance of language-specific idioms and phrases.

Additional work has explored whether structural differences across cross-lingual Chinese/English and English/Czech AMR pairs are due to syntactic idiosyncrasies (Xue et al. 2014); this information can be of use to machine translation because when AMR is incorporated into machine translation, these divergences could affect the quality of the system (Song et al. 2019; Nguyen, Pham, and Dinh 2021).

Cross-lingual studies of AMR primarily compare the structures between two AMRs representative of a parallel text. *AMRICA* visualizes and automatically aligns AMRs, including two AMRs of a sentence and its translation, to facilitate research into cross-lingual AMRs (Saphra and Lopez 2015). We take this as inspiration to use AMR pairs as a starting point for divergence classification between parallel texts.

Systems for cross-lingual AMR-to-text generation and text-to-AMR parsing (moving from an English AMR to non-English sentence and from a non-English sentence to an English AMR, respectively) have been designed as well (Ribeiro et al. 2021; Cai et al.

3 An automatically parsed corpus of approximately 2 million AMRs in the academic writing domain was also released by Zhao, Wang, and Lepage (2022).

2021; Lee et al. 2021; Xu et al. 2021a; Uhrig et al. 2021; Fan and Gardent 2020; Cai, Lin, and Wan 2021; Blloshmi, Tripodi, and Navigli 2020; Damonte and Cohen 2018; Wang, Li, and Xue 2018). Cross-lingual approaches to AMR parsing explore transfer learning techniques to generate parallel AMR annotations in multiple languages, and suggest that AMR can serve as a cross-lingual semantic representation capable of overcoming linguistic differences (Damonte and Cohen 2018; Zhu, Li, and Chiticariu 2019; Blloshmi, Tripodi, and Navigli 2020). Additional prior work has explored the role of structural divergences in cross-lingual AMR parsing (Blloshmi, Tripodi, and Navigli 2020; Damonte 2019). *XL-AMR* used transfer learning to automatically produce AMR annotations for Chinese, German, Italian, and French, and provides qualitative analysis suggesting that even with limited training data, the parser is able to manage many structural divergences across languages (Blloshmi, Tripodi, and Navigli 2020).

Though previous work has explored methods of characterizing the differences between pairs of cross-lingual AMRs, we aim to quantify the impact of the source language on AMR structure. In this article, we rigorously categorize both the types and causes of divergences in cross-lingual AMR pairs, identifying both quantitatively and qualitatively the actual impact of language on AMR structure. Prior work has motivated our study by pointing out that AMR is English-centric and may represent some languages more accurately than others; we extend this by assessing the effect of the source language itself on the structure of an AMR graph, and how that impacts the utility of AMR as a cross-lingual tool.

### 2.4 Related Work on Cross-lingual Meaning Representations

Cross-lingual investigations and adaptations have also been studied for other semantic representations. A multilingual extension of AMR, the Uniform Meaning Representation (Van Gysel et al. 2021), was developed to incorporate linguistic diversity into the AMR annotation process. The Uniform Meaning Representation framework adapts AMR with a focus on quantification and scope, as well as uniformity across languages.

In addition to AMR, several semantic representations have been used to capture meaning across languages (Van Gysel et al. 2019). Žabokrtský, Zeman, and Ševčíková (2020) summarized differences between formalisms (including AMR) for representing meaning across languages. We briefly highlight several such frameworks below.

The Universal Dependencies framework (Nivre et al. 2016; de Marneffe et al. 2021) has been used to analyze cross-lingual syntactic divergences (Nikolaev et al. 2020).

Universal Conceptual Cognitive Annotation (UCCA) annotates grammatical meaning while abstracting away from the syntax of the language (Abend and Rappoport 2013).

The Parallel Meaning Bank is a corpus of parallel texts with corresponding linguistic annotations and Discourse Representation Structure annotations projected from English (Abzianidze et al. 2017). Prior work towards the production of parallel meaning banks focused on the alignment of informative translations: translations where more details are included in the target translation than the source text (Bos 2014). Work on the Parallel Meaning Bank also includes a comparison of translations as being meaning-preserving or not, and these discrepancies as being largely due to: human annotation error, syntactic differences in definite articles, translation of proper names, or non-literal translations (van Noord et al. 2018).

Alignment for multilingual meaning representations has also been studied in relation to FrameNet, including recent work looking to produce a unified Multilingual FrameNet with alignments between all of the dozen FrameNet languages (Baker and Lorenzi 2020).

The Prague Czech-English Dependency Treebank (PCEDT) contains annotations of deep syntactic/shallow semantic dependencies in Wall Street Journal text, with gold annotations for English sentences as well as their Czech translations. PCEDT highlights some of the issues that arise out of automatically transferring an annotation schema to another language. Ultimately, the authors find that the annotation schema is not sufficiently fine-grained to provide a seamless conversion from annotation in one language to annotation in the other, and the difficulty of developing an annotation schema capable of this seamless transformation is unknown (Čmejrek, Cuřín, and Havelka 2004).

### 2.5 Semantic Divergences

Semantic divergences are differences between sentences which are purportedly parallel. Semantic divergences can appear in cross-lingual sentence pairs for a variety of reasons, and are an important phenomenon of language to consider when working with parallel semantic representations; cross-lingual AMR pairs may also encode these divergences seen at the string-level.

Translation divergences occur when translation from one language to another results in a different meaning or structure (Dorr 1994). These translation divergences can appear due to translation choices or to syntactic differences between the languages (Dorr 1990; Dorr and Voss 1993). Additional divergences can be introduced when automatically extracting and aligning parallel resources (Smith, Quirk, and Toutanova 2010; Zhai, Max, and Vilnat 2018; Fung and Cheung 2004).

The implications of these translation divergences include difficulties when using parallel texts for downstream tasks, because it can be difficult to identify why or how parallel sentences differ. For example, a parallel corpus, such as a work of fiction, likely contains some non-literal translations. When training a machine translation system on this parallel corpus, these divergences present a problem if looking to produce as literal a translation as possible.

Divergences have been explored with respect to synonymy (Gaillard et al. 2010) and diachronically (Montariol and Allauzen 2021).

Other studies have addressed whether and how given sentence pairs diverge (Carpuat, Vyas, and Niu 2017; Vyas, Niu, and Carpuat 2018; Briakou and Carpuat 2020, 2021; Zhai, Illouz, and Vilnat 2020). Carpuat, Vyas, and Niu (2017) classify divergences in parallel corpora using a cross-lingual textual entailment system to identify less equivalent sentence pairs. Related work has identified semantic divergences in parallel texts, classifying sentences as being divergent or non-divergent (Vyas, Niu, and Carpuat 2018).

The approach taken by Briakou and Carpuat (2020) to detecting string-level semantic divergences involves training multilingual BERT (Devlin et al. 2019) to rank sentences diverging to various degrees. They introduced a dataset called Rational English-French Semantic Divergences (REFreSD), which consists of sentence pairs from the French-English WikiMatrix (Schwenk et al. 2019). REFreSD sentence pairs are annotated with three types of divergences (subtree deletion, phrase replacement, and lexical substitution) based on a tree model (Briakou and Carpuat 2020). Other work to automatically classify divergences used a hierarchical alignment scheme of Chinese and English parse trees, enabling the identification and quantification of translation divergences (Deng and Xue 2017). Zhai, Illouz, and Vilnat (2020) detected non-literal translations in order to produce corpora of creative translations, to be used when pre-training translation models.

Semantic divergences are often seen in AMRs for parallel sentences. An example of this is shown in Figure 2. The English sentence “Which is your planet?” is aligned to

English: Which is your planet?

```
(p / planet
    :poss (y / you)
    :domain (a / amr-unknown))
```

Spanish: ¿ De qué planeta eres ?
Literal translation: What planet are you from?

```
(s / ser-de-91
   :ARG1 (t / tú)
   :ARG2 (p / planeta
       :campo (a / amr-desconocido)))
```

Figure 2: Example of a Spanish-English AMR pair, where semantic divergence due to translation choice results in differing AMR structure, notably a different root.

"¿De qué planeta eres?" which literally translates to "What planet are you from?" The Spanish sentence is seemingly less awkward than the original English sentence, and more explicitly asks about planet of origin, as opposed to ownership of a planet. This is a semantic divergence (due to translation choice), resulting in the Spanish AMR having a different focus (root).

## 3 Quantifying Differences between Parallel AMRs

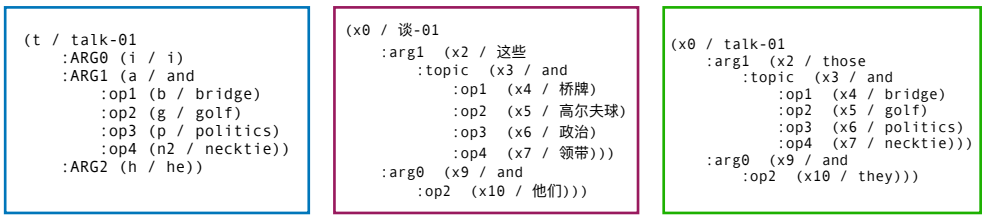
In this work, we assess the utility of AMR as a cross-lingual tool by studying the amount, causes, and types of differences in cross-lingual AMR pairs. Though AMR was originally designed for annotating English sentences, and not intended as an interlingua, it has since been adapted to a number of other languages, raising the question of how well it abstracts away from the particularities of individual languages.

In this section, we explore (1) how to measure the amount of structural difference between AMR pairs, and (2) what this measurable difference is.

In order to assess the applicability of AMR as a cross-lingual tool and explore what AMR captures between parallel sentences, we quantify the difference in AMR structure by language. First, we develop a novel method of comparing underlying graph structure for cross-lingual AMR pairs (§3.1). We compare AMRs which encode parallel sentences (in Chinese and English), and quantify the effect of source language on AMR structure after translating all of the Chinese tokens to English—leaving only differences in the underlying structure to measure.

Then, we move from differences in cross-lingual AMR pairs via manual annotation to differences in cross-lingual AMR pairs when automatically parsing (§3.2). Interestingly, we find that parsing directly from a non-English sentence causes a different AMR structure than if one translates that sentence to English first. We investigate how this reflects the aspects of meaning that AMRs capture.

Finally, having demonstrated the effect of one source language on AMR structure, we ascertain whether this appears regardless of language (§3.3). In order to do this, we measure language effect when parsing from multiple source languages (Spanish and Chinese).




(a) Original English annotation. (b) Original Chinese annotation. (c) Our annotation, the lexicalized version of the Chinese AMR.

Figure 3: Example of our annotation approach, showing the original English annotation (a), original Chinese annotation (b), and our lexicalized version of the Chinese annotation (c). The sentences annotated here are "I would talk to him about bridge, and golf, and politics, and neckties" and the parallel sentence: 和他们谈些桥牌呀，高尔夫球呀，政治呀，领带呀这些。

### 3.1 Effect of Source Language on AMR Structure

First, we set out to quantify the difference between parallel Chinese and English AMRs. To investigate AMR's ability to serve as an interlingua, previous work has explored methods of characterizing the types of differences between parallel AMR graphs (AMRs annotating parallel sentences in different languages). Additionally, cross-lingual text-to-AMR parsing and AMR-to-text generation assumes that non-English AMRs should be structured similarly to English AMRs, given that most cross-lingual AMR parsing has been compared against gold English AMRs (not gold in-language e.g. Spanish AMRs) (Wein and Schneider 2022). However, there has not yet been an effort to *systematically quantify* the effect on AMR structure of the language of the sentence being parsed (hereafter, the *source language*).

We hypothesize that regardless of any language-specific information in the AMR (i.e. even if the labels are made to be in the same language), the structure of AMRs across language pairs will likely differ because of the linguistic properties of the source sentence. To better understand the impact of language on AMR structure in the pursuit of effective evaluation of cross-lingual AMR pair similarity, we aim to quantify the amount of impact in parallel AMRs. Here we explore the effect of source language on AMR structure in the large annotated parallel corpus of Mandarin Chinese and English AMRs (Li et al. 2016).

To quantify the impact of source language on the AMR, we eliminate the measurable impact of lexical divergence and focus solely on structural divergences. To do this, we take a pair of parallel English and Chinese AMRs and manually translate every word in the Chinese graph into its English equivalent. This process of *lexicalizing* all Chinese tokens (replacing them with English tokens) ensures that the only quantifiable difference is AMR structure, so that we do not penalize lexical differences. Structural elements of the AMR are largely unchanged. We then evaluate via Smatch (Cai and Knight 2013), which is an algorithm to compare AMR graphs and calculate similarity. This produces a Smatch score quantifying the effect of source language on AMR structure. The choice of language pair was motivated by linguistic differences between Chinese and English, as well as the prominence of Chinese sentence–to–English AMR parsing (Damonte and Cohen 2020).

### 3.1.1 Dataset

For our annotation and analysis, we make use of parallel gold Chinese and English AMR annotations of the novel *The Little Prince*—the Chinese AMRs from the CAMR dataset (Li et al. 2016)[4] and their parallel English AMR annotations (Banarescu et al. 2013).[5] The 100 AMRs used are the first 100 annotations of both development sets, corresponding to the first 100 sentences of *The Little Prince*; 20 sentences were double-annotated. The average sentence length is 15.3 tokens for the 100 English sentences and 19.5 tokens for the 100 Chinese sentences. Since the Chinese AMRs do not include `:wiki` tags, we remove all `:wiki` tags from the gold English AMRs.

Note that *The Little Prince* was originally written in French, so both the English and Chinese versions are translations and may exhibit features of translationese and/or may be subject to differences due to French serving as a third pivot language (Koppel and Ordan 2011).

### 3.1.2 Approach

Our approach to manual relexification (visualized in Figure 3) consists of taking the CAMR annotation and replacing the Chinese concepts with English tokens. We want to replace the Chinese concepts with English tokens so that we do not penalize lexical differences (which are apparent as the words are originally in different languages), but rather, exclusively measure the structural differences between the AMRs. Specifically, this consists of a three-step process:

1. Manually translate the Chinese concepts to equivalent English tokens.
2. Check the parallel gold English AMR to identify synonyms of the manually generated translations of the Chinese concepts.
3. If a synonym (close enough in meaning such that faithfulness to the Chinese sentence is not lost) of the manually generated translation appears in the gold English AMR, the term from the English AMR is used to replace the manually generated translation. Otherwise, the manually generated translation is used.

Additionally, there are some terms that appear in the CAMR annotations which would not appear in English AMR annotations. For example, functional particles such as 就 (a central particle with a multitude of uses) appear in the CAMR annotation schema but role-marking prepositions and other morphosyntactic details do not appear in the English AMR annotation schema. We remove these functional particles from the Chinese annotations rather than attempt to translate them into English. No other structural changes are made to the Chinese AMR.

We trained two linguistics students bilingual in English and Chinese in our approach. Approximately 4 hours were spent per annotator to produce the annotations and a text editor was used.

### 3.1.3 Results and Analysis

We collect 60 relexified AMR annotations from each annotator, with 20 sentences overlapping so that we can calculate inter-annotator agreement (120 annotations total, on 100 unique sentences). We calculate the Smatch scores between the annotations (Chinese AMR with English concepts) and the corresponding gold English AMR.

We use the Smatch metric to calculate inter-annotator agreement (IAA), which ranges from 0 to 1, with 1 indicating the graphs identical. We find that the average IAA is

4 `https://www.cs.brandeis.edu/~clp/camr/res/blj_dev.txt`
5 `https://amr.isi.edu/download/amr-bank-struct-v1.6-dev.txt`

*English translation:* Nothing about him gave any suggestion of a child lost in the middle of the desert, a thousand miles from any human habitation.
*Annotation 1:*

```
(x0 / look-02
   :polarity  (x2 / -)
   :degree  (x3 / slightest)
   :arg0  (x4 / he)
   :arg1  (x5 / child
      :quant  (x6 / 1)
      :arg0-of  (x7 / lose-02
         :location  (x8 / desert
            :mod  (x9 / large)
            :mod  (x10 / uninhabited)))))
```

*Annotation 2:*

```
(x0 / seem-01
   :polarity  (x2 / -)
   :degree  (x3 / remote)
   :arg0  (x4 / he)
   :arg1  (x5 / child
      :quant  (x6 / 1)
      :arg0-of  (x7 / lose-02
         :location  (x8 / desert
            :mod  (x9 / huge)
            :mod  (x10 / uninhabited)))))
```

Figure 4: Both annotations (from Annotator 1 and Annotator 2) for one of the sentences in our dataset. Note that the annotators provided the English concepts and the structure of the annotation is derived from the parallel Chinese annotation.

0.8645, and the lowest IAA of all the sentence pairs in the data was 0.64. Inter-annotator agreement here measures lexical agreement between the translators. The reason overall IAA would not be 1 is because translation choices are being made when producing the annotations. For example, in Figure 4, one annotator felt that a more faithful translation of 像 is `seem`, while the other annotator decided that a more accurate translation would be `look`. The same is true for the difference between `slightest` and `remote`, as well as between `huge` and `large`. None of those terms (either item of any of the three pairs) are captured in the parallel gold English AMR, so these differences reflect translation choices and not errors in annotation.[6] This pair of annotations received an IAA score of 0.85.

The production of these annotations is motivated by the ability to then quantify the amount of difference between our annotations and the gold English AMRs. We use Smatch to quantify this difference as the standard similarity evaluation technique for AMR pairs.

The Smatch score for the gold English AMRs in comparison to the annotations is 41% for those produced by Annotator 1 and 44% for those produced by Annotator 2. These Smatch scores are over 60 sentence pairs each. This indicates that there is a sizable

6 In the case of Figure 4, the differences are lexical and this type of divergence could be captured with a metric such as S2match (Opitz, Parcalabescu, and Frank 2020). However, in this section we are focused on *structural* differences in the AMRs.

*English sentence:* "It has horns."
*Gold English annotation:*

```
(h / have-03
   :arg0 (i / it)
   :arg1 (h2 / horn))
```

*Chinese sentence:* " 还有犄角呢。"
*Annotation (Chinese AMR with English concept labels)*:

```
(x0 / say
   :arg1  (x2 / have-03
      :manner  (x3 / even)
      :arg1  (x4 / horn)))
```

Figure 5: Gold English AMR and our annotation for parallel sentences.

*English sentence:* "Boa constrictors swallow their prey whole, without chewing it.
*Gold English annotation*:

```
(s2 / say-01
   :arg0 (b2 / book)
   :arg1 (s / swallow-01
      :arg0 (b / boa)
      :arg1 (p / prey
         :mod (w / whole)
         :poss b)
      :manner (c2 / chew-01 :polarity -
         :arg0 b
         :arg1 p)))
```

*Chinese sentence:* 这本书中写道："这些蟒蛇把它们的猎获物不加咀嚼地囫囵吞下
*Annotation (Chinese AMR with English concept labels)*:

```
(x11 / writes-01
   :arg0  (x13 / book-01)
   :arg1  (x14 / swallow-01
      :arg0  (x15 / boa
         :mod  (x16 / these))
      :arg1  (x17 / prey
         :poss  (x25 / x15))
      :manner  (x19 / whole)
      :manner  (x21 / chew-01 :polarity -))))
```

Figure 6: Gold English AMR and our annotation for parallel sentences (some roles removed for brevity of presentation).

effect of source language on the structure of the AMR even with the Chinese labels being replaced, raising questions for how we evaluate cross-lingual AMR parsers.

We expect that some of the differences we capture in our approach are due to translation, and some differences are due to syntactic and semantic properties, as established by previous work comparing more similar languages (Spanish and English) (Wein and Schneider 2021). One example of a syntactic effect on AMR structure can be seen in Figure 5. This syntax-induced divergence arises out of the ability in Chinese to omit sentence subjects when they can be understood from context (pro-drop), which explains why the Chinese graph is missing an `:arg0` argument. It is likely that in addition

*English sentence:* And after some work with a colored pencil I succeeded in making my first drawing.
*Chinese sentence:* 于是，我也用彩色铅笔画出了我的第一副图画。
*Literal English translation of Chinese sentence:* So, I also drew my first drawing with colored pencils.

Figure 7: An English and Chinese sentence pair from the dataset, displaying slight variation in the translation.

to these syntactic differences (such as the divergence noted in the example in Figure 5), there are also differences in meaning in parallel sentences caused by the translation process, regardless of whether there is additionally syntax-induced divergence.

A more subtle effect of source language on AMR structure can be seen in Figure 6 relating to the `:arg1 prey`. In English, we have "swallow their prey whole," such that "whole" is a semantic modifier of "prey," denoted by `:mod`. In Chinese, the equivalent is 囫囵 (wholly, possibly barbarically) 吞下 (swallow). Wholly (囫囵) is annotated as `:manner` to the swallowing (吞下), instead of as the `:mod` of `prey`. We consider this a faithful and standard translation reflective of cross-linguistic differences between the "swallow whole" construction in English and the "wholly swallow" construction in Chinese. This difference is reflected in the AMR.

One example of sentences being slight variants of each other rather than literal translations is the sentence pair seen in Figure 7. The annotation (same for both annotators) received a Smatch score of 43% similarity with the gold English AMR. The majority of the sentences are closely parallel, so we expect that the difference we are quantifying is an effect of syntactic and semantic divergence between Chinese and English.[7]

#### 3.1.4 Accounting for Design Differences

A few relatively superficial differences in annotation guidelines between Chinese and English need to be accounted for, as they may impact the Smatch score without being a direct reflection of source language impact. We found four types of differences which have an impact on AMR structure:

- CAMR uses the concept `mean` for elaboration, often included to indicate parentheticals or colons (present in 3 AMR pairs)
- CAMR uses the concept `cause` instead of `cause-01` to refer to the cause of an event, which is considered a non-core role (in 4 AMR pairs)
- CAMR occasionally uses `:beneficiary` instead of `:arg2` to refer to indirect object (in 5 AMR pairs)
- While English AMR does not account for the sentence being a quotation, CAMR roots all quotations with `say` (in 13 AMR pairs). In English AMR, only the first sentence in the quotation, starting with open quotes, is rooted with `say`. In Chinese AMR, any sentence containing quotes is rooted with `say`.

As seen in Table 2, even when removing all AMR pairs noticeably affected by schema differences, the Smatch score similarity between our annotations and the gold English AMRs only increases incrementally, and a large effect of source language remains. This indicates that the dissimilarity we measure in AMR structure is not due to differences in annotation schema.

7 If Chinese and English gold AMRs are released in different domains in future work, it would be interesting to repeat this analysis on those texts and compare our findings.

| Removed Diff. | Anno.1/Gold | Anno.2/Gold |
|---|---|---|
| None | 41% | 44% |
| Mean | 43% | 44% |
| Cause | 41% | 44% |
| Beneficiary | 41% | 43% |
| Quotation | 41% | 43% |
| All | 42% | 45% |

Table 2: Smatch scores without each of the four design differences.

#### 3.1.5 Conclusion

Our case study between Chinese and English serves as an analysis of the impact of linguistic divergence between those two languages on AMR structure. For a small dataset of 120 Chinese AMRs with English concept labels annotated with our approach, we analyze Smatch score differences between our Chinese AMRs with English concept labels and the corresponding gold English AMRs. This represents a novel approach to quantifying effect of source language on AMR structure. Through our annotation process of translating Chinese concepts to English, we find that there is a dramatic impact on AMR structures, with Smatch scores between our annotations and the gold English AMRs falling below 50%. For comparison, inter-annotator Smatch scores within a single language (Chinese) in the same domain have been reported at 83% (Li et al. 2016).

From these Smatch scores, we are able to demonstrate that the source language has a dramatic effect on the structure of an AMR, even if the AMR is a gold annotation with no noise introduced by automatic parsing. This result has important implications for (1) identifying cross-linguistic inconsistencies in the AMR schema, and (2) interpreting scores in cross-lingual AMR parsing evaluations (Damonte 2019).[8]

As a meaning representation, it is critical that an AMR graph effectively reflect the meaning of the sentence being parsed, regardless of the language being parsed. These results suggest that source language has a noteworthy impact on AMR structure. This substantive impact on AMR structure motivates further consideration for source language when working with AMR cross-lingually—either in evaluating cross-lingual AMR parsers or when developing and comparing AMR schemas in new languages.

### 3.2 Translating before Parsing versus after Parsing: AMR versus String-level Semantics

While §3.1 discussed the effect of source language on manually annotated gold AMRs, here we move towards the effect of source language on *automatically parsed AMRs*. Prior work on cross-lingual AMR parsing has shown that first translating a non-English sentence into English, and then parsing the translated English sentence into an AMR, results in relatively high (67.6 for German, 72.3 for Spanish, 70.7 for Italian, and 59.1 for Chinese) Smatch score against the gold English AMRs (Uhrig et al. 2021). In §3.1, we found that lexicalizing the (gold) Chinese AMR into English *after* parsing leads to noticeably lower Smatch scores (Smatch similarity scores of 41% and 44%, for Annotator 1 and 2, respectively), even though there is no noise introduced by automatic parsing, given that the parsing is manual.

8 "Cross-lingual AMR parsing" typically refers to parsing a sentence from a language other than English into a standard English AMR.

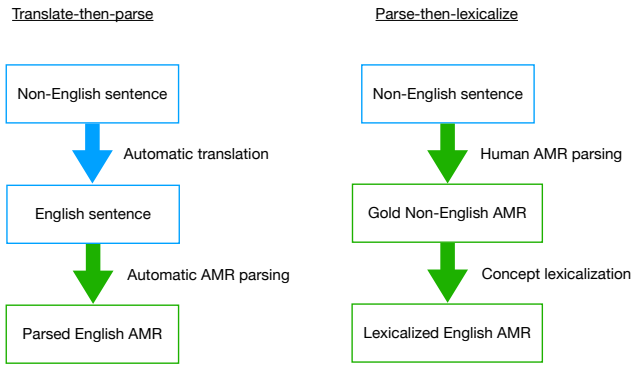



Figure 8: Flow diagram of the translate-then-parse (Uhrig et al. 2021) and parse-then-lexicalize §3.1 processes. Boxes corresponding to sentences are in blue, and boxes corresponding to AMR graphs are in green.

The question that then arises is why translating the sentence first and then parsing into an AMR ("translate-then-parse") results in more similar Smatch scores than lexicalizing the AMRs after parsing ("parse-then-lexicalize"). In this experiment, we consider the difference between (1) translating a non-English sentence into English, *and then* parsing that English sentence into an AMR, per Uhrig et al. (2021) versus (2) parsing a non-English sentence into non-English AMR, *and then* lexicalizing the concepts in the non-English AMR into English, per §3.1.

These two processes are depicted in Figure 8. Notably, Uhrig et al. (2021) demonstrate that translating a non-English sentence into English first, and then parsing that sentence using an English AMR parser, is an effective and accurate approach to parsing non-English sentences. §3.1, on the other hand, finds a large difference between parallel AMRs for which the non-English AMR was parsed then lexicalized.

These disparate results raises a question of the differences between parallel AMRs and parallel sentences, and how both relate to the act of translation itself. This work will look into what type of semantic difference is being captured across languages at the AMR-level, versus in text-form (by simply looking at the string), causing a difference in Smatch score in the previous set of results. To examine this question of the difference between translating then parsing versus parsing then lexicalizing (even when accounting for schema differences), we compute translate-then-parse scores for the AMRs presented in §3.1.

### 3.2.1 Results

We compare Smatch scores of AMRs for the same set of sentences produced via the parse-then-lexicalize method (detailed in §3.1, Chinese sentences parsed in Chinese AMR graphs, lexicalized into English) and the translate-then-parse method (Chinese sentences translated into English, parsed into English AMR graphs). We find that if first translating the Chinese sentence to English and then parsing, the similarity of those AMRs with the English AMRs is about 10% (via Smatch) higher on average than if comparing the gold AMRs without lexical differences. As shown in Table 3, the translate-then-parse scores were on average 9.61% Smatch higher than the average of the parse-then-lexicalize scores (comparing AMRs which were parsed from English translations of the original Chinese sentences, with AMRs parsed from the original English sentences). This is very

surprising because the automatic machine translation approach per Uhrig et al. (2021) leads to more parallelism between the AMRs than lexicalizing after parsing even when compared against gold human annotations.

This indicates that there are in fact some divergences that AMR is sensitive to that are removed when first translating the sentence, rather than faithfully representing the semantics via AMR of the original Chinese sentence, divergences included. This also suggests that some portion of the difference seen between the lexicalized Chinese AMRs and the gold English AMRs may be due to the literary genre because the translate-then-parse scores still indicate a sizable amount of semantic difference in the sentence pairs, with an average Smatch score of 52.44% when first translating the Chinese to English. Therefore, it is likely that some of these differences in the parse-then-lexicalize scores are due to translation in the literary work of *The Little Prince* making the sentences not exactly parallel, while some differences reflect semantic divergences captured by the AMR not easily visible in the string itself. Recall that the parsed-then-lexicalized AMRs are gold AMRs, so parser quality is not a potential cause of difference between translate-then-parse scores and parse-then-lexicalize scores.

#### 3.2.2 Qualitative Analysis

Qualitatively, we find that sentence pairs where Smatch similarity is higher via translate-then-parse than parse-then-lexicalize suffer from overemphasis of differences in lexicalization. Parallel AMRs where equivalent concepts are reflected as multiple words (a phrase or compound) in one AMR and one word in the other AMR result in a decrease in Smatch score (this issue of lexicalization in AMR is further discussed in §5.2). AMR/Smatch places more emphasis on these sorts of differences than a human would. The direct annotation from language (i.e. here from Chinese) maintains the lexicalization of the source language, while first translating the sentence to English conforms these n-to-one word pairs to what would typically appear in the English sentence.

Further, AMR is not robust enough to lexical paraphrases within the language or outside of the language, pointing to potential issues with AMR functioning cross-lingually. If expressed as one word in one language and multiple words in another language, the use of embeddings (such as in S2match (Opitz, Parcalabescu, and Frank 2020), XS2match (Wein and Schneider 2022), and Weisfeiler-Leman (Opitz, Daza, and Frank 2021)) won't actually mitigate the effect of structural differences induced by n-to-one word pairings across languages.

Also, we find that the quality of the automatic translation in the translate-then-parse method has a drastic effect on difference between translate-then-parse versus parse-then-lexicalize; when the automatic translation closely conforms to the target language (English), the similarity between the two AMRs is much higher than when the translation remains more faithful to the source language (Chinese).

Ultimately, this experimentation reveals that the effect of source language on AMR structure comes from directly parsing that language into an AMR, and that this effect is less pronounced when first translating to English. This finding suggests that the effect of source language on AMR is due to encoding from the language itself, and mitigating language effect through machine translation will lead to more English-like AMRs. AMR's approach to 'meaning' is sensitive to concept-level packaging, preserving certain nuances of construal (Trott et al. 2020) that a human might alter in paraphrasing/translation.

| Gold | | Auto | |
|---|---|---|---|
| **Anno. 1** | **Anno. 2** | **T+P** | **Diff** |
| 41% | 44% | 52% | 10% |

Table 3: Comparing the gold English and lexicalized Chinese AMRs for *The Little Prince* via Smatch. Anno. 1 and Anno. 2 are the Smatch scores between the lexicalized Chinese AMRs and the gold English AMRs, as reported in §3.1. T+P are our new translate-then-parse scores for the same dataset, first translating the Chinese sentences and then parsing them into English AMRs. The reported score is the Smatch score between those AMRs parsed from the Chinese sentences and the gold English AMRs. "Diff," indicating the average amount by which the translate-then-parse scores are higher than the parse-then-lexicalize, is calculated as follows: $AVG(T + P - AVG(Anno.1, Anno.2))$, averaging across every individual test point in the dataset.

| **En-Zh LPP** | **En-Es LPP** |
|---|---|
| 25% | 30% |

Table 4: Smatch scores for comparisons between: Chinese and English (En-Zh) gold AMRs annotating parallel sentences from *The Little Prince* (LPP), English and Spanish (En-Es) gold AMRs annotating parallel sentences from *The Little Prince*.

### 3.3 Spanish versus Chinese Language Effect

Thus far, we have identified that AMR, when parsing directly from the source language, captures some differences that are not captured when we first translate the Chinese sentence to English. It may be that the structural differences between Chinese and English are causing the AMRs to diverge, so we next perform a similar analysis on Spanish-English sentence pairs. To examine the effect of source language on AMR structure for two different languages, in the same genre, we perform a small Smatch analysis on gold Chinese and Spanish AMR annotations of *The Little Prince*.

Prior work comparing Chinese versus Czech AMRs to English AMRs has indicated that different languages may be more compatible with English AMR (Urešová, Hajič, and Bojar 2014). In order to address the potential amount of structural difference caused by the language itself (Chinese versus Spanish in comparison to English), we perform a basic comparison of the raw Smatch scores for Spanish and Chinese AMRs of *The Little Prince*. We use the Spanish *Little Prince* gold AMRs from Migueles-Abraira, Agerri, and Diaz de Ilarraza (2018) and the Chinese gold AMRs from Li et al. (2016). We find that the Spanish annotations of *The Little Prince* are not much more similar to the English annotations than the Chinese are, with the Chinese-English Little Prince Smatch score being 0.25 and the Spanish-English Little Prince Smatch score being 0.30. This suggests that in Experiment 1, the results are not exclusive to Chinese, because the amount of structural similarity for Spanish-English Little Prince AMRs is close to that for the Chinese-English Little Prince AMRs.

This experimentation confirms that the effect of language on AMR structure appears regardless of whether you are parsing from Chinese or Spanish, suggesting that this effect is not isolated to one language or another based on similarity to English.

Given the fact that these AMRs are not lexicalized, we also want to ensure that named entity overlap is not causing a sizable impact on Smatch score. In the Chinese

*The Little Prince* annotations, no English named entities are present. One book title ("The Real Story") is referenced in the English AMRs for the same sentences, and the Chinese tokens are ( 真实的故事) are also present in the parallel Chinese AMRs. This named entity does cause a divergent number of operators for that `name` concept, but as this is the only named entity in all of these AMRs, named entities do not play a notable role in this comparison. Likewise, in the Spanish *The Little Prince* annotations, there are no named entities appearing, and the same is true for the associated parallel English AMRs (this is a different subset of *The Little Prince* than the Chinese AMR subcorpus). Similarly, there is no token overlap for tokens which are not named entities (for either the Spanish and Chinese), ensuring that lexical overlap is not causing a sizable impact on Smatch score.

### 3.4 Summary of Findings: Quantifying Differences between Parallel AMRs

In this section, we have explored the impact of language on AMR structure, in order to quantify the amount of difference between parallel AMRs. We find that:
1. AMRs substantially structurally differ when annotating parallel sentences in different languages;
2. this divergence is less pronounced when first translating to English; and
3. this holds regardless of whether the source language is Chinese or Spanish.

This suggests that, in its current state, AMR is impacted by source language, which can pose a challenge to cross-lingual work. In the next two sections, we set out to understand these cross-lingual differences being captured by AMR by first analyzing the specific structural divergences, and then comparing AMR-level semantics to what is captured in text form.

## 4 Analyzing Cross-lingual Differences

Now that we have quantified differences in cross-lingual AMR pairs, we consider why these differences appear. Here, we will study more closely why exactly these differences appear.

A variety of factors come into play in translating from one language into another (Dorr 1990; Dorr and Voss 1993). The resulting parallel texts are not always completely equivalent in meaning. Differences/divergences between the source and target may reflect lexical or grammatical differences between the two languages, stylistic considerations, or simply the translator's preference for idiomatic phrasing.[9]

Identifying translation divergences may enable more nuanced use of parallel text in applications; for example, it has been shown that translation divergences have a measurable impact on machine translation (Vyas, Niu, and Carpuat 2018). This information on the types and causes of divergence enables these parallel texts to be more fully utilized in cross-lingual natural language processing tasks. Specifically, different types of semantic divergences impact the performance of neural machine translation systems differently (Briakou and Carpuat 2021), which motivates work to categorize and describe divergences in parallel texts.

Taking advantage of the fact that AMR resources and tools have been developed for a variety of languages (§2.2) and given the usefulness of AMR for downstream multilingual tasks such as machine translation (Song et al. 2019; Nguyen, Pham, and Dinh 2021), we seek to detect and explain divergences in cross-lingual AMR pairs. AMR

9 Semantic divergences are covered more fully in §2.5.

makes for a useful tool in the study of divergences by stripping away the grammar of the sentence to focus on the core aspects of meaning. Faced with ambiguity in language and multiple ways to express the same meaning, AMR condenses meaning in spite of variation to expose the key information of the sentence.

In order to explore why these cross-lingual AMR divergences appear, we analyze the causes and types of differences between parallel AMRs. We develop an annotation schema for these divergences, and annotate a small subset of existing Spanish-English AMR pairs with their AMR divergences (both type and cause).

Motivated by our coarse-grained, quantitative analysis of language-induced differences in AMR structure in the previous section, here we perform a fine-grained analysis of the types of structural differences found in cross-lingual AMR pairs. The structure of an AMR reflects the semantic relations of the sentence, so structural divergences (a difference in the structure of an AMR graph, whether it be the label or role) between multilingual AMR pairs serve as a reflection of semantic differences between a sentence and its translation.

We develop and present a categorization schema to identify both the type and the cause of the divergence (detailed in §4.1). The causes of divergence include semantic divergence, annotation divergence, and syntactic divergence. The types of divergence are rooting with different focus, adding/omitting non-core role difference, choosing a different non-core role, switching an argument and non-core role, adding/omitting an argument, and choosing a different argument. The reasons for annotating both the type and cause of a structural divergence are to (1) make the data more adaptable to cross-lingual NLP applications, (2) identify non-literal translations, (3) make AMR more cross-linguistically consistent (relevant for incorporating AMR into cross-lingual applications such as neural machine translation (Song et al. 2019)), and (4) investigate the ways in which annotation, semantics, and syntax play a role in cross-lingual AMR parsing (which are concerns for cross-lingual AMR parsing previously pointed out in (Damonte 2019)). Next, we show example annotations in §4.2.

We then annotate a set of 50 parallel English-Spanish AMRs annotations from *The Little Prince* (Migueles-Abraira 2017) and 200 English-Spanish AMR annotations from the Spanish annotations of AMR 2.0 (Wein et al. 2022a) using our divergence schema and make these annotations available online (presented in §4.3). Using this small set of gold annotated data, we are able to explore the comprehensiveness and meaningfulness of this annotation schema. We then examine the frequency and types of these divergences in each of the two datasets. Finally, we discuss the implications of our findings on AMR parsing and cross-lingual tasks §4.4.

### 4.1 Cross-lingual AMR Divergence Annotation Schema

We develop a categorization schema to be able to identify with granularity the type of structural divergence as well as the cause of the divergence between the two AMRs.

If an AMR pair has a structural divergence, meaning there is some difference in the way that the two AMRs are structured, there must be both a type of divergence (Figure 9) and a cause for the divergence (Figure 10).

*Cause of Divergence.*
The cause of the divergence can be:

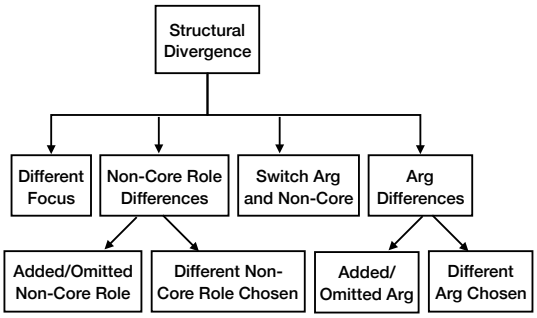



Figure 9: Types of structural divergence.

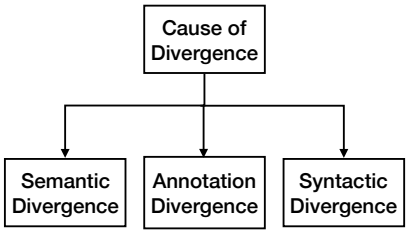



Figure 10: Causes of structural divergence.

- Semantic divergences ("sem"): due to translation choice, e.g. *an instant later / instantly*.[10]
- Annotation divergences ("anno"): due to annotation choice, e.g. rooting the AMRs with "erupt" versus "seem" in the case of *Volcanic eruptions are like fires in a chimney / Volcanic eruptions are like the fire of a chimney*. While the annotators could have converged on isomorphic AMRs, specific choices made by the annotators/ annotator disagreement led to divergence in the graphs.
- Syntactic divergences ("synt"): inherent differences between the languages, e.g. *belonged to a businessman / was of a businessman*. The annotators could not have agreed because the translated concepts are conveyed differently within the AMR structure.

In the case of semantic divergences, the Spanish translation is an inexact, non-literal, translation of the English sentence (or vice versa). For annotation divergences, the Spanish translation is a literal translation of the English sentence but the AMR was annotated differently nonetheless. Syntactic divergences arise because a feature of of the language (either English or Spanish), leading to a structural divergence in the AMR pair.

*Type of Divergence.*
As shown in Figure 9, the type of structural divergence can be a:
1. different focus ("focus")
2. a difference between the arguments/core roles ("diffarg" or "omitarg")

10 The examples given here for illustration are an English phrase and a literal translation of the Spanish phrase; they are presented in greater detail below.

3. the same label/feature being an argument in one AMR and a non-core role in the other AMR ("switch")
4. a difference between non-core roles ("diffnoncore" or "omitnoncore")

These four subcategories are listed by decreasing degree of granularity and effect on the structure of the rest of the AMR.

If the focus is different between the two AMRs, then the entire rest of the structure differs by definition, because the arguments and non-core roles that the focus can take on differ. Therefore, if an AMR pair is annotated with each having a different focus, then it is not possible to then annotate argument differences or non-core role differences.

In the paragraphs that follow, we include and explain the annotation of 7 illustrative examples.

### 4.2 Example Annotations

*Example of focus, sem.*
One example of a focus difference due to translation is shown in Figure 2. The focus, or root, of the AMR graph is the head/first node in the AMR. Here the focus differs because of the translation from the awkward "Which is your planet?" to the Spanish "¿De qué planeta eres?" which more explicitly asks about the geographic origin of the person being asked the question. The translation is a non-literal translation and the Spanish AMR reflects the Spanish sentence.

*Example of focus, anno.*
An example of a different focus due to annotation choice includes the AMR pair for the English sentence *Volcanic eruptions are like fires in a chimney.* and the parallel Spanish sentence *las erupciones volcánicas son como el fuego de una chimenea.* The focus difference here, where in the English sentence the focus is *erupt* and in the Spanish sentence the focus is *parecer* (seem), is due to annotation divergence in these gold annotations. The Spanish translation is a very literal translation of the English sentence. There are no other inherent language differences that place more emphasis on the "seeming" in the Spanish sentence either. Therefore, the difference is due to neither semantic divergence, being a literal translation, nor due to syntactic divergence; the difference is due to annotator discrepancy.

English: Volcanic eruptions are like fires in a chimney.

```
(e / erupt-0
    :ARG1 (v / volcano)
    :ARG1-of (r / resemble-01
        :ARG2 (f / fire
            :location (c / chimney))))
```

Spanish: Las erupciones volcánicas son como el fuego de una chimenea.
Literal translation: Volcanic eruptions are like the fire of a chimney.

```
(p / parecer
    :ARG0 (e / erupción
        :mod (v / volcánico))
    :ARG1 (s / ser-de-91
        :ARG1 (f / fuego)
        :ARG2 (c / chimenea)))
```

Similarly, if an argument is added/omitted or the label for an argument and non-core role are switched, then it is not possible to annotate a non-core role difference. It is still possible to have more than one structural divergence in an AMR pair because there are often multiple arguments and non-core roles in an AMR, so multiple differences may occur which do not explicitly encompass the others.

*Example of switch, anno.*
An AMR pair with English sentence: *To forget a friend is sad.* and Spanish sentence *Olvidar a un amigo es triste.* is an example of the infrequent "switch." This pair annotates *forget* in English as an `:ARG0`, while *olvidar* (forget) in Spanish is annotated as `:domain`. This is due to an annotation divergence. The English PropBank entry for sad-02 denotes `:ARG0` as the being the causer of the sadness which makes it an appropriate choice for the English, and the same guidelines were being referenced for the Spanish AMR annotation. Therefore this is due to annotation divergence, and likely is an annotation error in the Spanish AMR.

Additionally, an argument difference can be due to adding/omitting an argument, or because for the same argument label (e.g. `:arg0`) different arguments are chosen. If the same part of the sentence/feature of the AMR is featured as an `:arg0` in one AMR and an `:arg1` in the other AMR, this counts as two added/omitted arguments because the annotator judged that there was sufficient evidence for an `:arg0`, but this was omitted in the other AMR, and vice versa.

*Example of diffarg, synt.*
The `:ARG1` difference between *businessman* and *hombre* (man) is due to syntactic divergence. *Hombre de negocios* is the Spanish translation of *businessman*, and literally means "man of business," and thus is structured in the AMR as *hombre* (man), `:mod` *negocio* (business).

English: The fourth planet belonged to a businessman.
```
(b / belong-01
    :ARG0 (p / planet
        :ord (o / ordinal-entity :value 4))
    :ARG1 (b3 / businessman))
```

Spanish: El cuarto planeta era de un hombre de negocios.
Literal translation: The fourth planet was of a businessman.
```
(p / pertenecer
    :ARG0 (p2 / planeta
        :ord (e / entidad-ordinal :valor 4))
    :ARG1 (h / hombre
        :mod (n / negocio)))
```

Similarly to the `diffarg` label, a non-core role difference can be due to adding/omitting a non-core role, or because for the same argument label difference non-core roles are chosen.

*Example of diffnoncore, sem.*
This difference in the non-core attribute (`:time`/`:tiempo`) is due to semantic divergences, being a non-literal Spanish translation. The English sentence says "an instant later,"

while the Spanish translation literally says "instantly." This is thus reflect in the corresponding AMRs.

English: "I think it is time for breakfast," she added an instant later.

```
(a / add-01
    :ARG0 (s / she)
    :ARG1 (t / think-01
        :ARG0 s
        :ARG1 (t2 / time
            :purpose (b / breakfast-01)))
    :time (l / late
        :degree (m / more
            :quant (i / instant))))
```

Spanish: "Creo que es la hora de desayunar," añadió ella al instante.
Literal translation: "I think it's time for breakfast," she added instantly.

```
(a / añadir
    :ARG0 (e / ella)
    :ARG1 (c / creer
        :ARG0 e
        :ARG1 (h / hora
            :propósito (d / desayunar)))
    :tiempo (a2 / al-instante))
```

*Example with two omitnoncore, synt divergences.*
In English, `:grado` corresponds to `:degree`, so the degree is included in the Spanish sentence, but the domain is included in the English sentence. This is due to syntactic divergence because a natural Spanish translation of "that is funny" would literally translate in English to "how funny," making degree more appropriate than domain.

English: That is funny!

```
(f2 / funny
    :domain (t2 / that))
```

Spanish: ¡Qué gracioso!
Literal translation: How funny!

```
(g / gracioso
    :grado (t / tan))
```

*Example of no divergence.*
These AMRs are equivalent in every way, having equivalent sets of labels/relations, as well as the same arguments for each label.

English: Draw me a sheep...

```
(d / draw-01
    :ARG0 (y / you)
    :ARG1 (s / sheep)
    :ARG2 (i / i)
    :mode imperative)
```

Spanish: Dibújame una oveja...
Literal translation: Draw me a sheep...

```
(d / dibujar
    :ARG0 (t / tú)
    :ARG1 (o / oveja)
    :ARG2 (y / yo)
    :modo imperativo)
```

### 4.3 Annotated Data

In order to demonstrate the scope, effectiveness and utility of this annotation schema, we annotate two English-Spanish corpora: a small corpus of sentences from *The Little Prince* within the literary domain (Migueles-Abraira, Agerri, and Diaz de Ilarraza 2018), and a larger corpus of sentences from the news and web domain (Wein et al. 2022a).

First, we annotate a small corpus of English-Spanish sentences within the domain of literary works. We use a set of 50 English-Spanish AMR pairs, representing parallel sentences from *The Little Prince* (Migueles-Abraira, Agerri, and Diaz de Ilarraza 2018). Given the nature of *The Little Prince* being a literary work, Migueles-Abraira, Agerri, and Diaz de Ilarraza developed Spanish translations that would closely mirror the English versions so as to largely avoid the effect of literary translation. The English sentences selected from the larger dataset were chosen based on the fact that they collectively represent relevant linguistic issues hindering Spanish AMR annotation: NP ellipses, third person possessive pronouns, third person clitic pronouns, and the usage of the particle *se*.

In Wein and Schneider (2021), we annotated the 50 Spanish-English AMR pairs with a divergence classification. If there was no structural divergence, then none were listed. It is possible to have more than one structural divergence (type,cause pair) so any structural divergence is listed with its individual cause.

As it is difficult to draw firm conclusions about naturally occurring divergences from only these 50 pairs, for this article we expand our investigation by performing an additional 200 Spanish-English annotations (by the same annotator). While the 50 AMRs were manually edited to make the Spanish sentences more like the English, in this examination we make use of 200 Spanish-English AMR pairs which are both longer and also have not been edited to make the sentences more similar. This ensures that the semantic and annotation dissimilarity organic to cross-lingual AMR comparisons appears in addition to any syntactically-induced divergences. Therefore, we are able to examine the relationship between causes and types of divergences in this expanded, and less synthetically parallel, cross-lingual AMR setting.

Specifically, we annotate 200 sentences pairs from the Spanish AMR corpus (Wein et al. 2022a), aligned with their *AMR 2.0 - Four Translations* English counterparts (Knight et al. 2021): 100 sentences from the "Consensus" portion, which consists of DARPA GALE weblog and Wall Street Journal data; 40 sentences from the "DFA" portion, which consists of BOLT discussion forum source data; and 60 sentences from the "BOLT" portion, which consists of BOLT discussion forum machine translation data from Mandarin Chinese to English.

For the four subcorpora annotated (totaling 250 Spanish-English AMR pairs) in this study, their number of nodes in the dataset and average number of concepts per AMR are listed in Table 5.

| Source Data | Annotated Pairs | Eng Concepts | Concepts per Eng AMR |
|---|---|---|---|
| *The Little Prince* | 50 | 384 | 7.68 |
| Consensus (Wein et al. 2022a) | 100 | 1415 | 14.15 |
| DFA (Wein et al. 2022a) | 40 | 461 | 11.53 |
| BOLT (Wein et al. 2022a) | 60 | 1851 | 18.51 |

Table 5: AMR pairs annotated for divergences: existing 50 (Wein and Schneider 2021) and new 200 English-Spanish with their source data, number of annotated AMR pairs, datasets, document lengths for the English AMRs (via total number of concepts), and average number of concepts per English AMR. The data from *The Little Prince* originates from Migueles-Abraira, Agerri, and Diaz de Ilarraza (2018).

*The Annotation Process.*
All sentence pairs were annotated by the designer of the annotation scheme, who is fluent in both English and Spanish. There were few apparent difficulties in applying the scheme. The primary question was whether divergences that seemed like they were due to syntactic divergences were in fact inherent properties of the language or were imposed due to stylistic choices of the translator, as this relies on expert knowledge of both languages. Reliability of the annotations could be validated by extending the study to include additional annotators and calculating inter-annotator agreement.

*Distribution of Corpus.*
Figure 11 shows the breakdown of cause of divergence for the 50 pairs from *The Little Prince*, as well as the relationship between type of divergence and cause. Specifically, it contains the number of instances of every structural divergence caused by each of the three causes, with the types of structural divergences in rows and the causes of structural divergences in columns. While different focus and different `arg` chosen are regularly due to annotation choice, different non-core role chosen or omitted non-core role are much more often due to translation choice. Of the 50 multilingual AMR pairs, 13 pairs had no structural divergence, 11 pairs had multiple divergences (9 pairs with 2 divergences and 2 pairs with 3 divergences), and 26 pairs have one structural divergence. There were 49 divergences in total.

The results in Figure 11 demonstrate that there is a considerable discrepancy between type of divergence and cause. While non-core role divergences tend to be primarily due to semantic divergences, argument divergences tend to be due to annotation divergences or syntactic divergences. Change in focus tends to be due to an annotation divergence. The root serves as a representation of the central focus or topic of the sentence, so this relationship suggests that the central topic of the sentence is not changing between the parallel texts and instead the different focus annotation is due to annotator discrepancy.

The analysis of the data from *The Little Prince* revealed that, in a scenario where the cross-lingual AMR guidelines are the same and the sentences are altered to be more literally parallel to enable AMR parallelism, there are clearly differences which appear in Spanish-English AMR pairs due to the language itself, in addition to translation divergences and annotator choice. Now, we investigate the causes for and types of divergences corpus in a more organic cross-lingual scenario. This allows us to see how these diverges emerge in a situation where the AMRs are more dissimilar and there are differences in annotation approach/guidelines.

Figure 12 includes the types and causes of divergence for the 200 news and web annotations. Of the 200 AMR pairs, 21 had no divergence, and there were 245 divergences

| | sem | anno | synt | |
|---|---|---|---|---|
| focus | 2 | 6 | 2 | 10 |
| diffarg | 2 | 7 | 5 | 14 |
| omitarg | 0 | 2 | 2 | 4 |
| diffnoncore | 6 | 1 | 2 | 9 |
| omitnoncore | 4 | 1 | 5 | 10 |
| switch | 0 | 1 | 1 | 2 |
| | 14 | 18 | 17 | |

Figure 11: Number of instances of each structural divergence, and the number of times they were due to each cause of divergence, for the 50 Spanish-English annotations of *The Little Prince*. Greener cells indicate higher numbers and redder cells indicate lower numbers of instances. Sums across rows (types of divergence) and columns (causes of divergence) are found in clear cells to the right and bottom, respectively.

| | sem | anno | synt | |
|---|---|---|---|---|
| focus | 16 | 36 | 7 | 59 |
| diffarg | 18 | 37 | 12 | 67 |
| omitarg | 2 | 8 | 4 | 14 |
| diffnoncore | 19 | 24 | 13 | 56 |
| omitnoncore | 10 | 13 | 2 | 25 |
| switch | 3 | 17 | 4 | 24 |
| | 68 | 135 | 42 | |

Figure 12: Number of instances of each structural divergence, and the number of times they were due to each cause of divergence, for the new 200 annotations of the Spanish AMR corpus (Wein et al. 2022a).

total. The pairs had 0 to 3 divergences per AMR and there was an average of 1.23 divergences per AMR. Some similar patterns emerge in this set of results as the 50 annotation results, such as the fact that both different focus and different `arg` are two of the most comment types of divergences, both primarily appearing due to annotation choice.

Notably, there is a much higher proportion of annotation-induced divergences in this corpus, likely due to the fact that the sentences and AMRs were not designed to be more parallel, as they were for the 50 annotations of *The Little Prince*. This is also likely the same reason that we see a higher portion of "sem" divergences in this dataset. It is also possible that, given the fact that the Spanish AMR dataset had different guidelines that this could have led to the emergence of some the "anno" divergences.

We also see a higher rate of different non-core roles chosen. This is indicative of the fact that these AMRs have a larger depth and thus have more annotations where, for example, one portion of the AMR is affected by a different arg chosen whereas another portion of the AMR is affected by a different non-core role chosen (recalling that if an argument is changed, any "changed" dependents are not also marked as divergent).

### 4.4 Implications for AMR Parsing and Cross-lingual Tasks

The results of this study indicate that cross-lingual AMR pairs actually do sometimes inherently differ because of properties of the language, so these differences need to be accounted for when developing cross-lingual AMR systems.

As Xue et al. (2014) identifies divergences between Czech, Chinese, and English annotations in parallel texts which affect the degree to which AMR is able to serve as an interlingua, we similarly investigate the causes and ways in which AMR falls short of being an interlingua, tested on the cross-lingual case of Spanish and English. Notably, due to the granularity of our annotation schema, we are able to both describe and quantify the divergence. We identify pairs of AMRs as being more than solely divergent, by rigorously classifying each type of divergence, as well as attributing each divergence to one of three causes: semantics, annotation, or syntax.

Additionally, we show that the structure provided in AMR allows many divergences to be identified and categorized by an annotator, even cross-lingually. As such, AMR facilitates analysis that would be impossible from the strings or syntax trees alone. This motivates future work leveraging the semantic information captured by cross-lingual AMR pairs to identify sentences as being divergent or equivalent.

Finally, we find that, as one would expect, cross-lingual sentence pairs which are not made to be more parallel contain a higher rate of semantically and annotation choice–induced divergences than cross-lingual sentence pairs which are made to be parallel by enforcing similar English annotation practices and reducing semantic translation divergence.

### 4.5 Summary of Findings: Analyzing Cross-lingual Differences

We have presented our annotation schema for classification of structural divergences in cross-lingual meaning representations and produced 250 Spanish-English annotated examples. We demonstrate with our annotation schema and analysis of the annotated dataset that structural divergence in pairs of cross-lingual meaning representations can serve as a meaningful proxy of divergences between parallel texts. Therefore, tools which rely on highly literal translations, such as pre-trained machine translation systems, could benefit from applying this structural divergence annotation schema to cross-lingual Abstract Meaning Representations of the data.

Having found that there are measurable differences in AMR structure induced by the language of the sentence being parsed (§3), in this section, we set out to investigate the types and causes of these differences in structure. We find that AMR pairs do sometimes differ because of properties of the language being parsed—as well as because of annotation and translation differences. We have seen that there are specific differences in the language that are encoded into AMRs. In the final set of studies (§5), we explore how cross-lingual differences are encoded at the AMR-level versus in the string/text itself.

## 5 Comparing AMR vs. String-level Semantics Across Languages

In the previous two sections, we have showed that language-based differences are inherent to the structure of an AMR parsed from that language. In this section, we consider how what is captured by two AMR graphs differs from what is captured by their corresponding sentences, at the string-level (e.g., how does the meaning inferred

from simply scanning a sentence in its string form compare to the meaning inferred from reading an AMR?). The idea here is that AMR makes structural relations explicit, whereas shallower representations overlook them, which could give wrong impressions about meaning overlap. Specifically, we explore whether language-based divergences present themselves differently when captured by AMR versus by the string/text. We use "string-level" equivalence to describe the act of judging (either by machine or human) sentences as being equivalent based on the text itself, without using a symbolic meaning representation as an intermediary. AMR equivalence, on the other hand, identifies whether the AMR graphs of two sentences are equivalent or not (essentially, isomorphic).

We test this empirically as follows. First, we compare three AMR-based sentence similarity metrics and an embedding-based metric, BERTscore (Zhang et al. 2020), in a cross-lingual setting with respect to human judgments of semantic similarity.

Second, we compare how AMR-based metrics (again, versus string-based metrics) are able to capture *fine-grained divergences*. We hypothesize that AMR will be able to capture differences in meaning at a finer granularity than string-based metrics. We develop a novel approach to semantic divergence detection, which leverages the explicit semantics of AMR, and compare the amount of divergence detected by human string-level judgments versus AMR metrics. To do this for gold AMRs, we determine whether all of the string-level equivalent sentences correspond to isomorphic AMRs, showing that strings which are deemed equivalent may still have finer-grained differences captured in the AMR pair. Then, using automatically parsed AMRs, we are able to set a threshold which identifies the most semantically equivalent sentence pairs.

### 5.1 Background on Semantic Textual Similarity

One of the ways in which we compare how sentences and AMRs capture meaning in this section is through the lens of semantic textual similarity. Semantic textual similarity (STS) is the task of judging the graded semantic equivalence of two sentences (Agirre et al. 2016).

Recent work has incorporated AMR to measure semantic similarity for English. $S^3$BERT combined AMR metrics with Sentence-BERT by first partitioning Sentence-BERT embeddings into sub-embeddings, then training these sub-embeddings on separate components of AMR metrics (Opitz and Frank 2022). Also for English, Leung, Wein, and Schneider (2022) compared sentence similarity metrics which make use of vectors (e.g. BERTscore (Zhang et al. 2020)) and those which use AMR graphs (Smatch and S2match) to identify differences in those monolingual metrics. For generated English, Manning, Wein, and Schneider (2020) compared human judgments of output from AMR-to-text generation models against automatic metrics—BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), TER (Snover et al. 2006), CHRF++ (Popović 2017), and BERTscore—to see how well the automatic metrics correlate with human judgments of the generated text.

While prior research has either focused on combining explicit information from graph-based resources with vectors or comparing the two kinds of metrics monolingually, to the best of our knowledge, there has not been a direct, fine-grained comparison between these two kinds of representation across languages. Thus, we investigate whether such metrics facilitate *cross-linguistic* comparisons of sentence meanings, and how the explicit structure afforded by AMR corresponds with embedding representations

of sentences. We explore this below through cross-lingual comparisons of an embedding-based metric (BERTscore) and AMR-based metrics.[11]

### 5.2 AMR vs. Embedding-Based Metrics across Languages

In this subsection, we compare three AMR-based sentence similarity metrics and BERTscore in a cross-lingual setting through the lens of human judgments of semantic similarity, first quantitatively and then qualitatively.

#### 5.2.1 Cross-lingual Metrics

We develop three cross-lingual versions of AMR similarity metrics:

1. **XSmatch**, which translates tokens before applying Smatch (Cai and Knight 2013). We use the EasyNMT package,[12] specifically the Opus-MT model, to translate individual elements of the non-English AMR into English. Being a graph-comparison metric, Smatch compares AMR graphs rather than strings. Therefore, we are translating individual elements of the AMR and not the sentence itself. We also remove the word senses (numeric affixes to the concepts) for ease of translation and comparison.
2. **XSemBleu**, which is a cross-lingual version of SemBleu (Song and Gildea 2019). SemBleu is based on the machine translation metric BLEU (Papineni et al. 2002). Unlike Smatch, which searches for an alignment of variables between the two graphs, SemBleu instead converts each graph to a bag of *k*-grams for comparison. We again translate the non-English tokens in the non-English AMR; SemBleu does not break the AMR into triples (where we would translate specific elements of each triple), so we instead translate the entire non-English AMR to an English AMR. We do this by iterating token by token over the AMR and determining whether the current token needs to be translated. For example, parentheses, digits, and roles starting with a colon do not need to be translated. Similarly to XSmatch, XSemBleu translates concepts leading to a modified graph, but XSemBleu translates the text of the whole graph while XSmatch translates elements of individual triples. This approach to translation is more intensive than the translation required for individual tokens in XSmatch. Therefore, we aim to account for translation discrepancies and errors (e.g. part-of-speech discrepancies) by truncating the translations to the first `n` tokens. Given that instead of translating individual tokens in triples (as we did for XSmatch), we are translating the entire AMR string, in order to account for effects of more extensive translation e.g. part-of-speech discrepancies, we truncate/cut the translations to the first $n = 5$ tokens. Specifically, we truncate after translation for the non-English AMR, and also truncate for the English AMR. We use the default weights and smoothing function.
3. **XS2match**, which is a cross-lingual adaptation of S2match (Opitz, Parcalabescu, and Frank 2020). S2match incorporates word embeddings into Smatch to account for similarity of concept nodes without the same token being used. The current implementation of S2match relies on an external text file of embeddings, with the token being paired to an embedding in the file, and the embedding being retrieved from the text file for each token. To transport S2match to a multilingual format, we make use of the LaBSE (Feng et al. 2022) preprocesser and encoder. LaBSE embeddings are BERT-based cross-lingual sentence embeddings. We elicit a constant tensor of the word, preprocess it, and encode it to a LaBSE embedding. A benefit of XS2match is that, unlike XSmatch and XSemBleu, it

11 Full details of this study can be found in Wein and Schneider (2022).
12 `https://github.com/UKPLab/EasyNMT`

does not rely on neural machine translation practices that could unduly benefit a parser using the same translation tool (e.g. Uhrig et al. (2021)) through exact lexical matching.

#### 5.2.2 Human Evaluation

In order to compare AMR- and embedding-based metrics of similarity, we human judgments of sentence similarity. We collect such judgments for 100 Spanish-English sentence pairs and 150 Chinese-English sentence pairs which have associated gold AMRs on both sides. Both datasets are doubly annotated (meaning two people provided similarity judgments for each datapoint) by speakers fluent in both English and Chinese/ Spanish. The sentences for which we retrieved judgments come from the Chinese annotations of *The Little Prince* (Li et al. 2016) and the Spanish annotations (Wein et al. 2022a) of "AMR 2.0 - Four Translations" (Damonte and Cohen 2020). As a point of reference, we also measure correlation between human judgments and BERTscore, an embedding-based metric designed to assess the quality of generated sentences (Zhang et al. 2020).

We use both language pairs because Spanish and Chinese are quite different syntactically and vary noticeably in cross-lingual AMR performance. We also only use sentences which have associated gold AMRs, as opposed to existing sentence similarity data (Agirre et al. 2016), because we want to avoid introducing noise by relying on automatic parsers when comparing the AMR similarity with sentence similarity, or biasing our later assessment of cross-lingual parsers towards the parsers being used.

The parallel English sentences for both the Chinese and Spanish sentences are very related in meaning to the non-English sentences, so it was necessary to construct a dataset with varying degrees of sentence similarity (with all sentences still having associated gold AMRs).

In order to construct a Spanish-English dataset of varying similarities, 100 Spanish sentences from different genres in Damonte and Cohen (2020) were chosen. When pairing the Spanish and English sentences, so that not all pairs were parallel, we matched 25% of the Spanish sentences to English sentences with little to no relatedness, 50% of the Spanish sentences to English sentences with moderate relatedness, and 25% of the Spanish sentences retained their true translation pair, expected to have high similarity. These pairings were made manually.

A similar approach was used when constructing the Chinese-English dataset, with 66% of the dataset being mostly parallel and 33% of the dataset being mostly divergent (150 pairs total).

We asked human annotators to provide a score from 0 to 5 of how similar the content of a Spanish-English or Chinese-English sentence pair is. We use the task instructions from Agirre et al. (2016) as the basis for our instructions, "where 0 represents two sentences that are unrelated in meaning, and 5 indicates that the two sentences are perfect paraphrases of each other". We also added in degrees of similarity to the instructions to add clarity:

- (0) Completely unrelated
- (1) Not equivalent but share few subjects
- (2) Not equivalent but share some details
- (3) Roughly equivalent
- (4) Equivalent except for some details
- (5) Completely equivalent

We find that agreement for our sentence similarity protocol is high, with the correlation between annotator judgments being 0.93 for both the Spanish-English annotations

and the Chinese-English annotations.[13] The distribution of the sentence similarity scores is not uniform.[14]

### 5.2.3 Results

| | Smatch | XSmatch | SemBleu | XSemBleu | XS2match | BERTscore |
|---|---|---|---|---|---|---|
| Zh-En Anno. 1 | 0.43 | 0.40 | 0.20 | 0.42 | 0.51 | 0.76 |
| Zh-En Anno. 2 | 0.38 | 0.40 | 0.21 | 0.40 | 0.50 | 0.72 |
| Zh-En Anno. Sum | 0.41 | 0.41 | 0.21 | 0.42 | 0.51 | 0.75 |
| Zh-En BERTscore | 0.46 | 0.39 | 0.25 | 0.38 | 0.52 | 1.00 |
| Es-En Anno. 1 | 0.69 | 0.79 | 0.37 | 0.60 | 0.77 | 0.87 |
| Es-En Anno. 2 | 0.72 | 0.82 | 0.39 | 0.63 | 0.81 | 0.86 |
| Es-En Anno Sum | 0.72 | 0.82 | 0.38 | 0.63 | 0.80 | 0.88 |
| Es-En BERTscore | 0.74 | 0.82 | 0.41 | 0.62 | 0.79 | 1.00 |

Table 6: Pearson's correlation scores between the evaluation metrics (in the columns, along with BERTscore values of the sentences) and the human judgments of similarity (in the rows, with BERTscore, again). For each language pair, being Chinese-English and Spanish-English, we get the correlation with each of the two annotators as well as the sum of the similarity judgments.

Table 6 reports correlations between AMR-comparison metrics and BERTscore with human sentence similarity ratings. For BERTscore, we use the `bert-base-multilingual-cased` model with default settings.

First, note that the use of translation (in XSmatch and XSemBleu) is beneficial in SemBleu for both language pairs and in Smatch for Spanish-English. Applying translation to Chinese-English data has little effect on Smatch, for reasons discussed later in this subsection.

Comparing the three cross-lingual metrics, the two with the highest correlation to human judgments of sentence similarity are XSmatch and XS2match. While the correlation for Spanish-English is similar for those two multilingual metrics, though slightly higher via XSmatch, the correlation for Chinese-English is substantially higher using XS2match. As a result, we recommend XS2match as likely the best metric to use for cross-lingual AMR parser evaluation.

Notably, though perhaps unsurprisingly, correlation with the Chinese-English human annotations is lower for all metrics than correlation with the Spanish-English human annotations. This is likely not due to any issues with the human annotation itself, because the annotations still correlate well with BERTscore estimates of similarity, as seen in the final column of Table 6. Nonetheless, the Chinese-English human annotations are less correlated with BERTscore than the Spanish-English human annotations. Instead, the lower correlation of the Chinese-English annotations with BERTscore than the Spanish-English is likely due to lower performance on Chinese for the automatic machine translation systems and embeddings (for XSmatch, XSemBleu, and XS2match), as well as a greater degree of dissimilarity between the Chinese and English parallel AMRs than between the Spanish and English parallel AMRs (for all metrics). This greater degree of dissimilarity for certain AMR pairs has been studied previously (Xue et al. 2014) and

13 Annotator agreement for the SemEval task (Agirre et al. 2016) is not reported.

14 For the Chinese-English sentences, we initially collected judgments from a third annotator, but that annotator's interpretation of similarity was skewed towards saying most of the valid translations were completely equivalent, so the data was not informative for studying degrees of similarity. As a result we used the data from two other annotators.

is also evidenced here by the difference in the Smatch column in Table 6. The baseline Smatch similarity, with no multilingual component, is already much more correlated with human judgments for Spanish-English than for Chinese-English.

The monolingual Smatch score is already highly correlated with sentence similarity (for English-Spanish in particular, but for both language pairs broadly) because of structural similarity between the AMRs and matching between a subset of non-lexical nodes. For example, the Smatch scores aren't relying on lexical items as much as they are relying on the entities, e.g. shared named entities. This presence of names and named entities may also affect these correlation scores across languages because the Spanish-English text is from the news domain, which includes many country and person names, whereas the Chinese-English text is *The Little Prince*, which includes fewer of these named entities.

Even with the translation and truncation practices, XSemBleu correlation does not exceed XS2match correlation for either language pair. We hoped that SemBleu might be able to overcome structural differences between cross-lingual AMR pairs, but the undesirable presence of bias in the metric, which cannot be overcome without introducing a different bias (Opitz, Parcalabescu, and Frank 2020), likely led to the consequence of correlating less with the human annotations than the other metrics. Still, XSemBleu correlates fairly well with both language pairs.

We also measure correlation with scores from the BERTscore metric (Zhang et al. 2020), which uses the sentences directly and not the AMR graphs. BERTscore uses BERT-based models to compare embeddings of the words in the candidate and reference sentence via cosine similarity. We use BERTscore with the `bert-base-multilingual-cased` model as is the default for multilingual pairs. The last column of Table 6 shows that BERTscore achieves very strong correlations with human judgments. These results indicate that the embedding-based metric (BERTscore) is superior to the AMR-based metric overall (when not focusing on fine-grained divergences ). We also verify that sentence length is not a confounding variable in these judgments, with the correlation between average sentence length and human similarity score being only 0.07.

Reassuringly, the AMR metrics are not too far behind BERTscore, which further suggests that the AMR metrics are likely also capturing semantic similarity fairly accurately. This is unsurprisingly especially true for XS2match, which uses LaBSE embeddings (BERT-based cross-lingual sentence embeddings). Rows 4 and 8 of Table 6 compare the AMR metrics with BERTscore, showing that they are about as well correlated with each other as the metrics are with human judgments.

#### 5.2.4 Average XS2match and BERTscore Values

| XS2match score | BERTscore | Diff |
|---|---|---|
| 79% | 85% | −6% |

Table 7: On the Spanish-English parallel news texts, the average XS2match F1 score for the AMRs and the average BERTscore F1 score for the sentences. "Diff" indicates the average amount by which the BERTscore similarity is higher than the XS2match AMR similarity.

While §5.2.3 outlined how AMR-based XS2match and embedding-based BERTscore correlate with human judgments of similarity, in this experiment we consider overall average score produced by these two metrics on parallel sentences. This points to the general sensitivity of these metrics. In this experiment, we consider how we directly

compare XS2match (Wein and Schneider 2022) scores and BERTscore values for English-Spanish sentences, not using human judgments of similarity as an intermediary. Here, we ask again: "what is AMR capturing that string-level sentence comparison is not?" (and vice versa). We hypothesize that AMR and string-based metrics are somewhat complementary. To do this, we (1) collect XS2match scores between the aforementioned Spanish-English gold AMRs as well as (2) BERTscore values between those Spanish-English parallel sentences, and then (3) examine the pairs of sentences where the AMRs and sentences diverge notably to identify any patterns.

We work with 100 Spanish-English sentences and gold AMRs from the news domain. We use the 100 English AMRs/sentences from the AMR 2.0 dataset (Knight et al. 2017), and their parallel Spanish AMRs/sentences annotating the "AMR 2.0 - Four Translations" dataset (Wein et al. 2022a; Damonte and Cohen 2020). We use the news domain here to mitigate any potential genre effects or translation divergences, noting that the meaning similarity of the sentences in the AMR 2.0/"AMR 2.0 - Four Translations" dataset is very high, with the sentences being very faithfully translated.

Our results can be seen in Table 7. We find that the XS2match scores are generally lower than the BERTscore values, further confirming that AMR is more sensitive to certain differences which are not apparent in a string-based comparison. Pearson's correlation between the XS2match scores and BERTscore is somewhat low, with a correlation of 0.35. To test the potential impact of sentence length, we also collect Pearson's correlation scores for both metrics with sentence length. Sentence length is negatively correlated with BERTscore ($-0.47$), and less negatively correlated with XS2match ($-0.40$), indicating that XS2match is slightly less sensitive to sentence length than BERTscore.

#### 5.2.5 Qualitative Analysis: N-to-one Word Pairings

Qualitatively, we find that n-to-one word pairings across languages heavily affect AMR similarity. In particular within the news domain, syntactic divergences can manifest themselves within names, and AMR notation based on string order is not effectively able to capture that. For the AMRs in Figure 13, where the AMR structure is affected by the multi-word proper names, the XS2match score between the AMRs is 0.66, while the BERTscore between the sentences is 0.85. String alignment presents a problem for Smatch, in that parallel named entities with differing world alignment are not seen as equivalent in Smatch (e.g. in Figure 13, for the case of "International Atomic Energy Agency" and its parallel *Agencia Internacional de Energía Atómica*). Annotation choice and annotation schema differences also plays a role in divergences between the Spanish and English gold AMRs. For example, for "limited copies of the [...] report" (*copias limitadas del informe*) in Figure 13, the English AMR has an `:arg1` rooted by `publication` before denoting that it is a publication of a `copy`, while the Spanish AMR's `:arg1` is rooted directly by `copia` (copy).

These n-to-one word pairings pose one particular challenge to Smatch, contributing to the differences between AMR- and embedding-based metrics, and furthering the ways in which AMR is more sensitive to divergences than string-based comparisons.

### 5.3 Granularity of Meaning Captured by AMR and Embedding-Based Metrics

We have seen that AMR captures different (though overlapping) information from what a string or embedding encodes. We hypothesize strict semantic details (i.e., divergence from true semantic equivalence) that are often overlooked by simply reading the string are made explicit in the AMR, highlighting specific semantic information. In this subsection, we investigate the use and ability of AMR to capture meaning to a fine-grained degree.

**English sentence:** The International Atomic Energy Agency distributed limited copies of the IAEA report before a meeting on 11 September 2007 of the 35 members of the IAEA board.

**Spanish sentence:** La Agencia Internacional de Energía Atómica distribuyó copias limitadas del informe del OIEA antes de una reunión el 11 de septiembre de 2007 de los 35 miembros del consejo del OIEA.

```
(d / distribute-01
    :ARG0 (o / organization
        :name (n / name
            :op1 ``International''
            :op2 ``Atomic''
            :op3 ``Energy''
            :op4 ``Agency''))
    :ARG1 (p2 / publication
        :ARG2-of (c / copy-01
            :ARG1 (t / thing
                :ARG1-of (r / report-01
                :ARG0 (o2 / organization
                        :name (n2 / name
                    :op1 ``IAEA'')))))
        :ARG1-of (l / limit-01))
    :time (b / before
        :op1 (m / meet-03
            :ARG0 (p / person
                :quant 35
           :ARG0-of (h / have-org-role-91
                    :ARG1 (b2 / board
                        :poss o)
                  :ARG2 (m2 / member)))
            :time (d2 / date-entity
                :year 2007
                :month 9
                :day 11))))
```

(a) AMR for English sentence.

```
(c23 / distribuir-01
    :ARG0 (c37 / organization
        :name (c38 / name
            :op1 ``Agencia''
            :op2 ``Internacional''
            :op3 ``de''
            :op4 ``Energía''
            :op5 ``Atómica''))
    :ARG1 (c24 / copia
        :ARG1-of (c46 / limitar-01)
        :consist-of (c25 / informe
            :poss c38))
    :time (c51 / antes
        :op1 (c48 / reunir-03
            :ARG0 (c34 / miembro
                :quant 35
           :ARG0-of (c50 / have-org-role-91)
                :ARG1(c36 / consejo
                    :poss c38))
            :time (c52 / antes
               :op1 (c32 / date-entity
                    :day 11
                    :month 9
                    :year 2007)))))
```

(b) AMR for Spanish sentence.

Figure 13: Two parallel Spanish-English sentences and AMRs, which depict the effect of the structure of named entities on AMR divergence. (Note that the wikification of the English AMRs, omitted here in these examples, plays a small role in the artificial depression of the similarity between the two AMRs, as with other annotation schema differences.)

Translation between two languages is not always completely meaning-preserving, and information can be captured by one sentence which is not captured by the other. For example, consider the parallel French and English sentences from the REFreSD dataset (Briakou and Carpuat 2020) shown in Figure 14. The French sentence says "tous les autres édifices" (*all other buildings*) while the English specifies "all other *religious* buildings." Because the sentence goes on to list religious buildings, it could be inferred from context that the French is describing other *religious* buildings. The French author, for whatever reason, chose to exclude *religious*; the sentences thus convey the same overall meaning but are not *exactly* parallel. Under a strict or close analysis of the translation, these sentences could be considered divergent, because the meanings are not identical but at the string-level they are essentially equivalent (and are annotated as equivalent in the REFreSD corpus).

Semantic divergence (or conversely, semantic equivalence) detection aims to pick out parallel texts which have less than equivalent meaning. Semantic divergence detection plays an important role in many cross-lingual NLP tasks, such as translation studies (Dorr 1994) and machine translation (Carpuat, Vyas, and Niu 2017). Though semantic divergence across sentences in parallel corpora has been well-studied, current detection methods fail to capture the full scope of semantic divergence. State-of-the-art semantic divergence systems rely on perceived *string-level divergences*, which do not entirely encapsulate all semantic divergences.

All other *religious* buildings are mosques or Koranic schools founded after the abandonment of Old Ksar in 1957.

Tous les autres édifices sont des mosquées ou des écoles coraniques fondées à l'époque postérieure à l'abondance du vieux ksar en 1957.

Figure 14: Two parallel sentences from the REFreSD dataset marked as having no meaning divergence, for which the AMRs diverge.

Because implicit information can be critical to the understanding of the sentence (Roth and Anthonio 2021), we argue that a finer-grained measure of semantic equivalence is needed: a way to detect *strictly* semantically equivalent sentence pairs. In this work, we demonstrate that parsing sentences into Abstract Meaning Representation (AMR; Banarescu et al. 2013) graphs and comparing those graphs enables a finer-grained semantic comparison than simply comparing the sentences. We suspect that AMR may be useful in this case because it makes explicit every concept and relationship between those concepts present in the sentence, taxonomically categorizing each concept's role and argument.

Through analysis of data in two language pairs (English-French and English-Spanish), we demonstrate that string-level divergence annotations can be coarse-grained, neglecting slight differences in meaning. We find that comparing two AMR graphs is an effective way to characterize meaning in order to uncover finer-grained divergences, and this can be achieved even with automatic AMR parsers. Finally, we evaluate our AMR-based metric on a cross-linguistic semantic textual similarity dataset, and show that for detecting semantic equivalence, it is more precise than a popular existing model, multilingual BERTScore (Zhang et al. 2020).

#### 5.3.1 AMR for Identification of Semantic Equivalence

We leverage the semantic information captured by AMR to recognize semantic equivalence or divergence across parallel sentences. Figure 15, for example, illustrates a sentence pair with strictly equivalent meaning, along with the AMRs. Though the sentences differ with respect to syntax and lexicalization, the AMR graphs are structurally isomorphic. If the AMR structures were to differ, that would signal a difference in meaning.

Two particularly beneficial features of the AMR framework for this use case are the rooted structure of each graph, which elucidates the semantic focus of the sentence, as well as the concrete set of specific non-core roles, which are useful in classifying the specific relation between concepts/semantic units in the sentence. For example, in Figure 16 (also discussed in Figure 2) the emphasis on the English sentence is on possession—*your* planet—but the emphasis on the Spanish sentence is on place of origin, asking, which planet are you *from?* This difference in meaning is reflected in the diverging roots of the AMRs.

He later scouted in Europe for the Montreal Canadiens.

```
(s / scout-02
      :ARG0 (h / he)
      :ARG1 (c / continent
            :wiki "Europe"
            :name "Europe")
      :ARG2 (c2 / canadiens
            :mod "Montreal")
      :time (a / after))
```

Il a plus tard été dépisteur du Canadiens de Montréal en Europe. (*He later scouted for the Montreal Canadiens in Europe.*)

```
(d / dépister-02
      :ARG0 (i / il)
      :ARG1 (c / continent
            :wiki "Europe"
            :name "Europe")
      :ARG2 (c2 / canadiens
            :mod "Montreal")
      :time (p / plus-tard))
```

Figure 15: A pair of sentences and their human annotated AMRs, for which the sentences receive a "no meaning divergence" judgment in the REFreSD dataset, and are also equivalent per AMR divergence.

We find that non-core roles (such as `:manner`, `:degree`, and `:time`) are particularly helpful in identifying parallelism or lack of parallelism between the sentences during the annotation process. This is because AMR abstracts away from the syntax (so that word order and part of speech choices do not affect equivalence), but instead explicitly codes relationships between concepts via semantic roles. Furthermore, AMRs use special frames for certain relations, such as `have-rel-role-91` and `include-91`, which can be useful in enforcing parallelism when the meaning is the same but the specific token is not the same. For example, if the English and French both have a concession, but the English marks it with "although" and the French marks it with "*mais*" (but), the special frame role will indicate this concession in the same way, preserving parallelism.

*Granularity of the REFreSD Dataset.*
Another example, using sentences from the REFreSD dataset, is shown in Figure 17. These sentences are marked as having no meaning divergence in the REFreSD dataset but do have diverging AMR pairs. The difference highlighted by the AMR pairs is the `:time` role of reach/*atteindre*. The English sentence says that no. 1 is reached "within a few weeks" of the release, while the French sentence says that no. 1 is reached the first week of the release (*la première semaine*). In examples like this one it is made evident that string-level divergence (as appears in REFreSD) do not capture all meaning differences.

We explore the ability to discover semantic divergences in sentences either with gold parallel AMR annotations or with automatically parsed AMRs using a multilingual AMR parser, in order to enable the use of this approach on large corpora (considering that AMR annotation requires training).

We propose that an approach to detecting divergences using AMR will be a stricter, finer-grained measurement of semantic divergence than perceived string-level judgments.

Which is your planet?

```
(p / planet
    :poss (y / you)
    :domain (a / amr-unknown))
```

¿ De qué planeta eres ? (*Which planet are you from?*)

```
(s / ser-de-91
   :ARG1 (t / tú)
   :ARG2 (p / planeta
       :domain (a / amr-desconocido)))
```

Figure 16: Two parallel sentences and AMRs from the Migueles-Abraira, Agerri, and Diaz de Ilarraza English-Spanish AMR dataset, which diverge in meaning. The Spanish role labels are translated into English here for ease of comparison.

The use of a finer-grained metric would enable more effective filtering of parallel corpora to sentences which have minimal semantic divergence.

**5.3.2 Examining and Automatically Detecting Differences in Gold AMRs**
Here, we evaluate the ability of AMR to expose fine-grained differences in parallel sentences and how to automatically detect those differences. In order to do so, we produce and examine English-French AMR pairs, which is the first annotated dataset of French AMRs.

*Examination of Gold AMR Data.*
We focus on French for effective comparison with string-level semantic divergence models (because of the available resources), though it also makes for ideal candidates in a cross-lingual AMR comparison, as it is broadly syntactically similar to English. This suggests that the AMRs could be expected to look similar (though not exactly the same) as inflectional morphology and function words are not represented in AMR. Prior work has investigated the transferability of AMR to languages other than English, and has found that it is not exactly an interlingua, but in some cases cross-lingual AMRs align well. Additionally, some languages are more compatible (Chinese) with English AMR than other languages (Czech) (Xue et al. 2014).

*English-French AMR Parallel Corpus.*
In investigating the differences between the degree of divergence captured by AMR and string-level divergence, we aim to compare AMR similarity metrics with corresponding machine judgments of similarity at the string level. We compare human string-level judgments and AMR judgments for English-French parallel items. We produce gold AMR annotations for 100 sentences, which were randomly sampled, from the REFreSD dataset (Briakou and Carpuat 2020; Linh and Nguyen 2019).[15] For the French AMR annotation process, the role/argument labels were added in English as has been done in related non-English AMR corpora (Sobrevilla Cabezudo and Pardo 2019), and the concept (node) labels were in French.[16]

15 We also test our system on the full REFreSD dataset, using an automatic AMR parser.
16 Specific details of our production of these French AMRs can be found in Wein, Wang, and Schneider (2023).

Although the sales were slow (admittedly, according to the band), the second single from the album, "Sweetest Surprise" reached No. 1 in Thailand *within a few weeks* of release.

```
(r / reach-01
      :ARG0 (s / single :name (n / name :op1 "sweetest"
      :op2 "surprise")
            :ord (o / ordinal-entity :value 2)
            :source (a / album))
      :ARG1 "No.1"
      :location (c / country :name (n2 / name :op1 "Thailand"))
      :time (w / within
            :op1 (w2 / week
                  :quant (a2 / a-few)
                  :poss (r2 / release
                        :poss s)))
      :concession (s2 / slow-05
            :ARG1 (s3 / sales)
            :ARG1-of (s4 / say-01
                  :ARG0 (b / band)
                  :manner (a3 / admittedly))))
```

Même si les exemplaires ont du mal à partir (comme l'admet le groupe), le second single de l'album, Sweetest Surprise, atteint la première place en Thaïlande *la première semaine* de sa sortie.

```
(a / atteindre-01
      :ARG0 (s / single :name (n / name :op1 "sweetest"
      :op2 "surprise")
            :ord (o / ordinal-entity :value 2)
            :source (a2 / album))
      :ARG1 place :ord (o2 / ordinal-entity :value 1)
      :location (p / pays :name (n2 / name :op1 "Thailand"))
      :time (s2 / semaine
                  :ord (o3 / ordinal-entity :value 1)
                  :poss (s3 / sortie
                        :poss s))
      :concession (m / mal-05 :mod a-partir
            :ARG1 (s4 / exemplaires)
            :ARG1-of (s5 / admettre-01
                  :ARG0 (g / groupe))))
```

Figure 17: Two parallel sentences from the REFreSD dataset (Briakou and Carpuat 2020) marked as having no meaning divergence, but for which the AMRs diverge. The italicized spans in the text indicate one cause of AMR divergence.

*Findings from Corpus Annotation.*
In light of our research question considering whether AMR can serve as a proxy of fine-grained semantic divergence, we consider both qualitative and quantitative evidence. While producing this small corpus of French-English parallel AMRs, our suspicions that AMR would be able to more fully capture semantic divergence than perceived string-level divergence were confirmed. We uncovered a number of ways in which perceived string-level equivalence is challenged by the notion of AMR divergence.

Take the example in Figure 14. The difference between "religious" being applied in the French sentence and appearing in the English sentence is not captured by perceived string-level divergence, but is captured by AMR divergence.

| | AMR Divergent | AMR Equivalent |
|---|---|---|
| String-Level Divergent | 57 | 0 |
| String-Level Equivalent | 26 | 17 |

Table 8: Comparison between AMR Divergence annotations and String-Level Divergence REFreSD annotations for 100 French-English sentences.

Quantitative results appear in Table 8. AMRs are equivalent when the graphs are isomorphic (and contain the same concepts/roles, matching exactly), and the string-level equivalence is determined based on the gold labels from REFreSD. There are no instances where the string-level annotation claims that the sentences are divergent but the AMR annotations are equivalent. Conversely, there are 26 instances with AMR divergence but no perceived string-level semantic divergence. Therefore, AMR divergence is a finer-grained measure of divergence than perceived string-level divergence.

**5.3.3 Quantifying Divergence in Cross-Lingual AMR Pairs**
We have shown that not all pairs that humans considered equivalent at the string level receive isomorphic AMRs because they actually contain low-level semantic divergences. This suggests AMRs can be useful for more sensitive automatic detection of divergence. Now, we investigate whether we can automatically detect and quantify this divergence on gold AMRs via Smatch. In order to quantify this divergence in cross-lingual AMR pairs, we develop a simple pipeline algorithm which is a modified version of Smatch and incorporates token alignment. We test our modified Smatch algorithm on gold English-French AMR pairs and gold English-Spanish AMR pairs in comparison to the similarity scores output by Briakou and Carpuat (2020).

*Modified Cross-lingual Version of Smatch.*
Our simple pipeline algorithm extends Smatch, a measurement of similarity between two (English) AMRs (Cai and Knight 2013).[17] Smatch was designed to compare AMRs in the same language, with the same role and concept vocabularies. To compare AMR nodes across languages, the nodes first need to be cross-lingually aligned. This involves translating the concept and role labels. We take a simple approach of first word-aligning the sentence pair to ascertain corresponding concepts (most of which are lemmas of content words in the sentence). Our approach is similar to that of *AMRICA* (Saphra and Lopez 2015), but we use a different word aligner (fast_align rather than GIZA++[18]) and deterministic translation of role names if the labels are not in English. To align AMR graphs across languages, we word-align the sentence pairs, then map these alignments onto nodes in the graph. Role names are mapped deterministically based on a list from Migueles-Abraira (2017). We normalize the strings and remove sense labels from the English and French/Spanish concept labels. Finally, we run Smatch with the default number of 4 random restarts to produce an alignment. The Smatch score produced is an F1 score from 0 to 1 where 1 indicates that the AMRs are equivalent. This can be directly

17 This modified cross-lingual version of Smatch served as a preliminary/simplified version of XS2match, introduced in §5.2.
18 fast_align has been shown to produce more accurate word alignments, such as in the case for Latvian-English translation (Girgzdis et al. 2014).

used as a continuous value or converted to a binary judgment, where all non-1 pairs are divergent.

#### 5.3.4 Testing our Approach on Gold AMRs

One of the benefits of leveraging semantic representations in our approach to semantic divergence detection is that the identification of divergence boils down to determining whether the graphs are isomorphic or not (and accurate word alignment). This suggests that our pipeline algorithm should be highly effective at identifying whether AMR pairs are divergent or equivalent. In order to test our AMR-based approach to strict semantic equivalence identification, we first test on gold AMRs, which are created by humans and thus have no external noise from being automatically parsed.

We expect that our AMR divergence characterization would behave differently from a classifier of string-level divergence. This is because the string-level classification methods require specialized training data and as such learn to classify based on the perceived string-level judgments of semantic divergence. To test the strictness of our framing, we validate our quantification on gold English-French and gold English-Spanish cross-lingual AMR pairs.

| | **Equivalent** (17) | | | **Divergent** (83) | | | **All** |
|---|---|---|---|---|---|---|---|
| System | **P** | **R** | **F1** | **P** | **R** | **F1** | **F1** |
| Ours | 1.00 | 0.82 | 0.90 | 0.97 | 1.00 | 0.98 | 0.97 |
| BC'20 | 0.39 | 0.82 | 0.53 | 0.95 | 0.73 | 0.83 | 0.75 |

Table 9: FR-EN: Binary divergence classification on 100 gold French-English AMR/ sentence pairs, annotated for sentences from the REFreSD dataset. Precision (P), Recall (R), and F1 scores are reported for the Equivalent, Divergent, and All AMR pairs. We compare the performance of our model with the performance of the Briakou and Carpuat (2020) model, referenced as BC'20, on our finer-grained measure of divergence for the same English-French parallel sentences.

*Results on Gold English-French AMR Pairs.*

We test our pipeline algorithm on our 100 English-French annotated AMR pairs. As expected, the simple pipeline algorithm is very accurate at correctly predicting whether the cross-lingual pairs do or do not diverge according to the stricter criterion.

Table 9 showcases the ability of our AMR pipeline system and the Briakou and Carpuat (2020) system to identify these finer-grained semantic divergences. On these English-French AMR pairs, the F1 score for our system is 0.97 overall and 0.90 for equivalent (/isomorphic) AMR pairs. This high level of accuracy indicates we can reliably predict cross-lingual AMR divergence.

The Briakou and Carpuat (2020) system performs worse when using our finer-grained delineation of semantic divergence, achieving an F1 score of 0.75.[19] Unsurprisingly, the precision, recall, and F1 for their system is lower than the performance of our system, because theirs is not trained to pick up on these more subtle divergences. Note that on their own measure of divergence (perceived string-level divergence), the system achieves an F1 score of 0.85 on these same 100 sentences.

19 The Briakou and Carpuat (2020) system does not take AMRs as input, so we use the corresponding sentences as input for their system.

Of the 3 errors made by our algorithm (in all cases, classifying equivalent AMR pairs as divergent), 2 of the 3 are caused by word alignment errors. Named entities seem to pose an issue with fast_align for our use case.

| | **Equivalent** (13) | | | **Divergent** (37) | | | **All** |
|---|---|---|---|---|---|---|---|
| System | **P** | **R** | **F1** | **P** | **R** | **F1** | **F1** |
| Ours | 1.00 | 0.92 | 0.96 | 0.97 | 1.00 | 0.99 | 0.98 |
| BC'20 | 0.24 | 0.38 | 0.29 | 0.72 | 0.57 | 0.64 | 0.52 |

Table 10: EN-ES: Binary divergence classification with gold parallel AMRs. Included are Precision (P), Recall (R), and F1 for the Equivalent, Divergent, and All AMR pairs for our pipeline algorithm compared to the system by Briakou and Carpuat (2020), referenced as BC'20, on the same English-Spanish parallel sentences.

*Results on Gold English-Spanish AMR Pairs.*
In addition to testing our system on our English-French AMR annotations, we test our system on the 50 English-Spanish AMRs and sentences released by Migueles-Abraira, Agerri, and Diaz de Ilarraza (2018), who collected sentences from *The Little Prince* and altered them to be more literal translations. Recent work classified these AMRs via AMR structural divergence schema (Wein and Schneider 2021).

In Table 10, we measure the ability of our pipeline system and the Briakou and Carpuat (2020) system to detect semantic divergences at a stricter level, as picked up by the AMR divergence schema.

Our system performs similarly well on Spanish-English pairs as it did on the English-French pairs, described in Table 9. This demonstrates that our pipeline algorithm is not limited to success on only one language pair, and we further affirm that the simple pipeline algorithm is a reliable way to predict cross-lingual AMR divergence.

#### 5.3.5 Strictness Results using Automatic English-French AMR Parses

We have shown that we are able to use gold (human annotated) AMRs to capture a finer-grained level of semantic divergence, quantifiable via Smatch. We extend this further by determining whether fine-grained semantic divergences can be detected well even when using noisy automatically parsed AMRs. To do so, we compare the Smatch scores of automatically parsed AMR pairs with the human judgments output on the corresponding sentences by Briakou and Carpuat (2020).

To take the expensive human annotation piece out of the process, we show that automatic AMR parses can be used instead of gold annotations by establishing a threshold, instead of via binary classification. Therefore, we use the F1 score output by our pipeline algorithm as a *continuous score* and establish thresholds (described later in this section) to divide the data between divergent and equivalent.

We automatically parse cross-lingual AMRs for the entirety of the English-French parallel REFreSD dataset (1033 pairs). The REFreSD dataset is parsed using the mbart-st version of SGL, a state-of-the-art multilingual AMR parser (Procopio, Tripodi, and Navigli 2021). The (monolingual) Smatch score for the SGL parser, comparing our gold AMRs with the automatically parsed AMRs, is 0.41 for the 100 French sentences

using Smatch (0.43 using our pipeline algorithm)[20] and 0.52 for the 100 parallel English sentences using Smatch.

In doing error analysis, we find that the data points which are classified as having no meaning divergence but have extremely low F1 scores are largely suffering from parser error. We do find that there are pairs classified in REFreSD as having no meaning divergence at the string-level that do correctly receive low F1 scores. For example, the sentence pair in Figure 17, which has a REFreSD annotation of string-level equivalence and a gold AMR-level annotation of divergence, was assigned an F1 score of 0.3469.

Despite Smatch scores of 0.5 between the gold and automatic parses, both are usable for the task of detecting finer-grained semantic equivalence. To demonstrate the usefulness of our continuous metric of semantic divergence using automatically parsed AMR pairs, we develop potential thresholds at which one could separate data as being equivalent vs. divergent.

Because our metric is more sensitive, a practitioner could choose their own threshold by determining appropriate precision (how semantically equivalent they wanted a subset of filtered data to be) and recall (how much data they are willing to filter out) needs. This tradeoff is depicted in Figure 18. For example, if all pairs are marked as equivalent, precision would be approximately 40% on the REFreSD dataset if considering solely the "no meaning divergence" pairs equivalent.

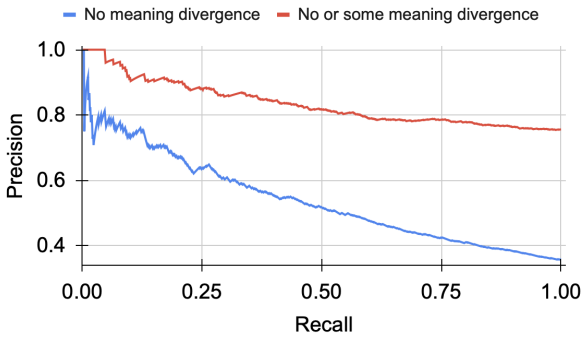



Figure 18: Precision/recall curve for equivalence detection in the 1033 sentence pairs in the full REFreSD dataset (English-French) using automatic AMR parses. Precision reflects the percent of sentences in which REFreSD human annotation was equivalent (as labeled as no meaning divergence in the blue/bottom curve, or as labeled as having either no or some meaning divergence in the red/top curve).

*Comparison against Model Probabilities.*
Though it is reasonable to assume that if the gold AMR annotations provide a distinctly finer-grained measure of divergence than string-level divergence then this would also be the case when using automatically parsed AMRs, we want to ensure the continued strictness of our methodology. To do this, we compare the values of our continuous metric and the probabilities produced by the Briakou and Carpuat (2020) system.

20 The SGL parser approaches cross-lingual parsing as the task of recovering the AMR graph for the English translation of the sentence, as defined in prior work (Damonte and Cohen 2018). The result is that the parses of French sentences are largely in English, and default to French concepts only for out-of-vocabulary French words. The alignments in our pipeline account for this to better reward the native French concepts.

Because the probabilities produced by the system described in Briakou and Carpuat (2020) are always very close to 1 (equivalent) or very close to 0 (divergent) and there are far more divergent instances than equivalent instances, median serves as a more effective form of comparison than mean between our F1 score and their probability score. Above the 0.7 threshold, the median F1 for our system is 0.7869 and mode is 0.8; the median probability for the Briakou and Carpuat (2020) system is 0.9990. For the 0.6 threshold, our median is 0.6667 and their median is 0.9871. Above the 0.5 threshold, our median is 0.5814 and their median is 0.8907, Because these numbers are lower for our system than their system, we confirm that our measure is a stricter measure of equivalence even when using the automatically parsed AMRs. If this type of semantic divergence detection system is being used in order to ascertain which items a human adjudicator should look at on a fixed budget, the absolute scores may matter less than rankings. We find that the rankings additionally differ drastically. Of the top 50 sentences ranked by AMR divergence (which range in AMR similarity score from 0.96 to 0.67), only 19 of the 50 appear in the 166 sentences scored 1.0 by the Briakou and Carpuat (2020) system.

#### 5.3.6 Sentence Similarity Evaluation with Automatically Parsed English-Spanish AMRs

Multilingual BERTscore (Zhang et al. 2020) is an embedding-based automatic evaluation metric of semantic textual similarity. Semantic textual similarity considers the question of semantic equivalence slightly differently because it rewards semantic overlap as opposed to equivalence.

As we have explored in previous sections, our AMR-focused approach in general is stricter than sentence-based measures of equivalence, in particular corpus filtering methods. Because our system is a stricter measure of semantic equivalence, it may be the case that our system can more precisely identify the most similar sentences than existing measures of sentence similarity. In this section, we look at the most semantically equivalent sentences in the dataset (as judged by our approach and as judged by multilingual BERTscore) in comparison to their human judgments of equivalence. Specifically, we aim to investigate: (1) whether the average human similarity score for the most similar n sentences is higher when ranked by our AMR-based metric versus when ranked by BERTscore, and (2) whether human judgments of sentence similarity for the most similar sentences are more correlated with our AMR-based metric than with BERTscore. We compare our AMR-based metric to multilingual BERTscore because it has been shown to work will in cross-lingual settings when comparing system output to a reference (Koto, Lau, and Baldwin 2021).

*Data.*

To perform this comparison, we use the 301 human annotated Spanish-English test sentences from the news down of the SemEval task on semantic textual similarity (Agirre et al. 2016).

*Smatch with Cross-Lingual AMR Parsing.*

For our analysis, we use the Translate-then-Parse system (T+P; Uhrig et al. 2021). Providing the Spanish sentences as input, T+P translates them into English, and then runs an AMR parser[21] on the English translation. Because the Spanish sentence was translated into English and *then* parsed, this automatic parse can be compared against

21 Via amrlib: `https://github.com/bjascob/amrlib`

the automatic parse of the original English sentence with plain Smatch (no cross-lingual alignment added).

As we have established, the noise introduced by automatic parsers can be overcome in our approach. We validate that the Smatch scores retrieved after using Uhrig et al.'s (2021) parser still bears some correlation with the Smatch scores on the aligned gold AMRs.[22]

*Sentence Similarity Results.*
The average human judgment score, on a scale of 0 to 5 with 5 being exactly equivalent, for all sentence pairs which have an AMR similarity score greater than 0.8 is 4.98. The average human judgment score for all sentence pairs which have a multilingual BERTscore similarity score greater than 0.8 is 4.89. Similarly, the average human judgment score for pairs with an AMR similarity score of greater than 0.7 is 4.86, while the average human judgment score for pairs with a multilingual BERTscore greater than 0.7 is 3.8. This is because multilingual BERTscore takes a much broader view of semantic equivalence. The true range of BERTscore values of semantic similarity tends to be confined within 0 to 1, and the metric is not particularly sensitive to smaller differences or errors (Hanna and Bojar 2021). This makes BERTscore a better choice generally for the question of semantic similarity, because it is more correlated with human judgments, but when assessing fine-grained semantic equivalence, our AMR-metric is more accurate than BERTscore.

While the human judgments occupy the full range of 0 to 5, the multilingual BERTscore values of these sentences range from 0.57 to 0.87, as shown in Figure 19. The AMR similarity score ranges from 0.11 to 0.98.

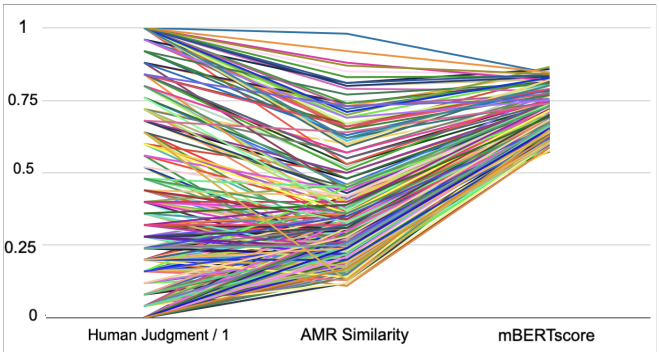



Figure 19: All data points normalized to a range of 0 to 1 for the Spanish-English sentence pairs from Agirre et al. (2016), including human judgment, AMR similarity score, and multilingual BERTscore. This displays the decreased range of multilingual BERTscore judgments in comparison to human judgments and AMR similarity.

This might suggest that a higher threshold should be used for multilingual BERTscore to achieve the same level of semantic granularity. However, our AMR similarity metric is also more correlated with human judgments for the most semantically equivalent

22 On the 50 Spanish-English sentences (from §4), the correlation between the Smatch scores (in comparison to the same gold AMRs) when using either the translation-then-parse method or the method of aligning concepts via fast_align is 0.31. This can be interpreted as a weak correlation. We find that both methods (translating the sentence first, or our pipeline algorithm aligning concepts in AMRs of different languages) work sufficiently well to capture the amount of divergence between cross-lingual AMR pairs.

sentences. For the top 20 items as ranked by AMR similarity, correlation with human judgments is 0.4068. But the top 20 items as ranked by multilingual BERTscore are not correlated with human judgments ($-0.0023$). When looking at all items above the multilingual BERTscore of 0.8, correlation with human judgment is 0.1645, whereas for all items above the AMR similarity score of 0.8, correlation with human judgment is 0.2675. Correlation is calculated with Pearson correlation. Overall, AMR similarity score correlates with human judgment at a coefficient of 0.8367, which is slightly lower than the 0.8605 correlation between multilingual BERTscore and human judgment.[23]

This evidence further supports that our metric is in fact a finer-grained measure of semantic equivalence, and is therefore better at identifying which sentences are exactly semantically equivalent.

#### 5.3.7 Conclusion

In this work, we have proposed a stricter measure of semantic divergence than existing systems which rely on perceived differences at the string level. We have demonstrated that parsing sentences into Abstract Meaning Representations and comparing those graphs facilitates a more detailed semantic comparison, when using either gold *or* automatically parsed AMR pairs.

Fine-grained semantic equivalence detection is not widely studied—yet it holds promise for a number of applications, including (for example) reducing the workload of human translators in post-editing of machine translation output. As it stands, MT systems which receive human post-editing, as well as translator aids which present MT output for human translators, present all sentences in the dataset to the human translator (Green, Heer, and Manning 2013). Being able to filter out exactly semantically equivalent sentence pairs would reduce this workload. Similarly, filtering out exactly semantically equivalent sentences can lessen the amount of annotation necessary for human evaluations of text (Saldías et al. 2022).

Other potential uses include cross-lingual text reuse detection (plagiarism detection), which asks whether one sentence is simply another sentence exactly translated (Potthast et al. 2011). Translation studies and semantic analyses could also benefit from the distinction between semantically equivalent sentence pairs and sentence pairs which have subtle or implicit differences (Bassnett 2013).

In additional to the potential engineering applications, our study provides important insight into how semantic differences are captured in AMR-based metrics versus at the string-level, indicating that a greater degree of semantic nuance is captured by the AMR encoding.

### 5.4 Summary of Findings: Comparing AMR vs. String-level Semantics

In §3 and §4, we showed that AMR structure is impacted by language and examined structural differences in cross-lingual AMR pairs. In this section, we assessed how such differences relate to other cross-lingual sentence-level measures of meaning overlap. We found that AMR-based metrics are generally less aligned with human judgments than embedding-based metrics and that AMR metrics are greatly affected by lexicalization. Still, we found that because of the explicit semantics captured by AMRs, the AMR graph-based metrics are especially well-suited to identify finer-grained divergences in

23 Opitz et al. (2023) similarly finds that BERT-based metrics are more correlated with human judgment than AMR-based metrics, and shows that BERT and AMR-based metrics can be complementary.

meaning than simply comparing sentence pairs. The findings in this section point to cross-lingual AMR-semantics and string-level semantics differing, and demonstrate that comparing AMRs highlights different (but overlapping) information versus a string-based comparison.

This affects our understanding of AMR as an interlingua or cross-lingual tool: it allows us to be conscious of the fact that the cross-lingual divergences seen in sentences are not entirely caused by the same effects that induce cross-lingual divergences in AMR pairs. Therefore, we need to adjust our expectations of what AMRs can tell us about cross-lingual meaning. AMR may not be sufficient to facilitate all aspects of semantic comparison, but does seem well-suited to certain components of meaning, such as semantic roles.

## 6 Conclusion

In this work, we have addressed the applicability of Abstract Meaning Representation (AMR) to cross-lingual settings. In order to investigate whether and how AMR is able to capture meaning across languages, we have considered AMR through multiple lenses. First, we measured the amount of difference between parallel AMRs in different languages, comparing the AMR graphs to determine the underlying effect of source language on AMR structure. Next, we developed an annotation schema to identify the types and causes of differences between AMR graphs. Finally, we compare sentence embeddings and AMR graphs as representations of cross-lingual meaning (dis)similarity.

Through a series of studies over several language pairs (English-Chinese, English-French, and English-Spanish), we conclude that, notably, source language has an effect on AMR structure, and in order to adapt AMR to individual languages, and the abstractions in AMR are not fully reflective of cross-lingual semantic correspondences understood by humans. Still, cross-lingual AMR is able to illuminate semantic divergences at a fine-grained level (more so than string/text-based semantics). Issues of lexicalization differences (linguistic phrase/token misalignment) across languages stand in the way of using AMR to its fullest extent across languages.

Crucially, we have found that AMR is impacted by source language and uncovered the causes (both linguistic and translation-induced) as well as the implications of that finding. This paves the way for future applications of AMR in cross-lingual contexts, potentially extending any of the many applications of English AMR—e.g. machine translation (Li and Flanigan 2022; Song et al. 2019), summarization (Kouris, Alexandridis, and Stafylopatis 2022; Inácio and Pardo 2021; Hardy and Vlachos 2018), event extraction (Garg et al. 2016; Rao et al. 2017; Li et al. 2015), toxic content detection (Elbasani and Kim 2022), etc.—to non-English settings.

Future work might pursue improved metrics for comparing AMRs across languages. Finally, investigating the differences between cross-lingual AMR and cross-lingual versions of other meaning representations would perhaps give further insight into what features of meaning representations are most useful for which cross-lingual tasks.

**Acknowledgments**

We thank anonymous reviewers for their feedback and thanks Zhuxin Wang, Wai Ching Leung, and Yifu Mu for their contributions to our prior publications. We also thank Rexhina Blloshmi, Eleftheria Briakou, Yitao Cai, Luigi Procopio, and Dongqin Xu for providing system results, and Roy Ilany, Yang Janet Liu, Siyao Logan Peng, and Yilun Zhu for providing associated human judgments. Components of this work were supported by a Clare Boothe Luce Scholarship.

## References


Abdelsalam, Mohamed Ashraf, Zhan Shi, Federico Fancellu, Kalliopi Basioti, Dhaivat J Bhatt, Afsaneh Fazly, et al. 2022. Visual semantic parsing: From images to Abstract Meaning Representation. *arXiv preprint arXiv:2210.14862*.

Abend, Omri and Ari Rappoport. 2013. UCCA: A semantics-based grammatical annotation scheme. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 1–12, Association for Computational Linguistics, Potsdam, Germany.

Abzianidze, Lasha, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Association for Computational Linguistics, Valencia, Spain.

Agirre, Eneko, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, Association for Computational Linguistics, San Diego, California.

Anchiêta, Rafael and Thiago Pardo. 2018. Towards AMR-BR: A SemBank for Brazilian Portuguese language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan.

Azin, Zahra 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, Association for Computational Linguistics, Florence, Italy.

Baker, Collin F. and Arthur Lorenzi. 2020. Exploring crosslinguistic frame alignment. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 77–84, European Language Resources Association, Marseille, France.

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Association for Computational Linguistics, Sofia, Bulgaria.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Association for Computational Linguistics, Ann Arbor, Michigan.

Bassnett, Susan. 2013. *Translation studies*. Routledge.

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

Bonial, Claire, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. 2020. Dialogue-AMR: Abstract Meaning Representation for dialogue. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 684–695, European Language Resources Association, Marseille, France.

Bos, Johan. 2014. Semantic annotation issues in parallel meaning banking. In *ACL 2014*.

Briakou, Eleftheria and Marine Carpuat. 2020. Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1563–1580, Association for Computational Linguistics, Online.

Briakou, Eleftheria and Marine Carpuat. 2021. Beyond noise: Mitigating the impact of fine-grained semantic divergences on neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7236–7249, Association for Computational Linguistics, Online.

Cai, Deng, 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, Association for Computational Linguistics, Punta Cana, Dominican Republic.

Cai, Shu 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, Association for Computational Linguistics, Sofia, Bulgaria.

Cai, Yitao, Zhe Lin, and Xiaojun Wan. 2021. Making better use of bilingual information for cross-lingual AMR parsing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1537–1547, Association for Computational Linguistics, Online.

Carpuat, Marine, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79, Association for Computational Linguistics, Vancouver.

Choe, Hyonsu, Jiyoon Han, Hyejin Park, Tae Hwan Oh, and Hansaem Kim. 2020. Building Korean Abstract Meaning Representation corpus. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 21–29, Association for Computational Linguistics, Barcelona Spain (online).

Čmejrek, Martin, Jan Cuřín, and Jiří Havelka. 2004. Prague Czech-English Dependency Treebank: Any hopes for a common annotation scheme? In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 47–54, Association for Computational Linguistics, Boston, Massachusetts, USA.

Damonte, Marco. 2019. *Understanding and Generating Language with Abstract Meaning Representation*. Ph.D. thesis, University of Edinburgh.

Damonte, Marco and Shay Cohen. 2020. Abstract Meaning Representation 2.0 - Four Translations. Technical Report LDC2020T07, Linguistic Data Consortium, Philadelphia, PA.

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

Deng, Dun and Nianwen Xue. 2017. Translation divergences in Chinese–English machine translation: An empirical investigation. *Computational Linguistics*, 43(3):521–565.

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

Dorr, Bonnie. 1990. Solving thematic divergences in machine translation. In *Proceedings of the 28th Annual Meeting on Association for Computational Linguistics*, ACL '90, page 127–134, Association for Computational Linguistics, USA.

Dorr, Bonnie J. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

Dorr, Bonnie J, Eduard H Hovy, and Lori S Levin. 2004. Machine translation: Interlingual methods.

Dorr, Bonnie J. and Clare R. Voss. 1993. Constraints on the space of MT divergences.

Elbasani, Ermal and Jeong-Dong Kim. 2022. AMR-CNN: Abstract Meaning Representation with convolution neural network for toxic content detection. *Journal of Web Engineering*, 21(03):677–692.

Fan, Angela and Claire Gardent. 2020. Multilingual AMR-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2889–2901, Association for Computational Linguistics, Online.

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Association for Computational Linguistics, Dublin, Ireland.

Fung, Pascale and Percy Cheung. 2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable

corpus. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1051–1057, COLING, Geneva, Switzerland.

Gaillard, Benoît, Yannick Chudy, Pierre Magistry, Shu-Kai Hsieh, and Emmanuel Navarro. 2010. Graph representation of synonymy and translation resources for crosslinguistic modelisation of meaning. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 819–830, Institute of Digital Enhancement of Cognitive Processing, Waseda University, Tohoku University, Sendai, Japan.

Garg, Sahil, Aram Galstyan, Ulf Hermjakob, and Daniel Marcu. 2016. Extracting biomolecular interactions using semantic parsing of biomedical text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Girgzdis, Valdis, Maija Kale, Martins Vaicekauskis, Ieva Zarina, and Inguna Skadina. 2014. Tracing mistakes and finding gaps in automatic word alignments for Latvian-English translation.

Green, Spence, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 439–448.

Hanna, Michael and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Association for Computational Linguistics, Online.

Hardy, Hardy and Andreas Vlachos. 2018. Guided neural language generation for abstractive summarization using Abstract Meaning Representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 768–773, Association for Computational Linguistics, Brussels, Belgium.

Heinecke, Johannes and Anastasia Shimorina. 2022. Multilingual Abstract Meaning Representation for Celtic languages. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 1–6, European Language Resources Association, Marseille, France.

Ilmy, Adylan Roaffa and Masayu Leylia Khodra. 2020. Parsing Indonesian sentence into Abstract Meaning Representation using machine learning approach. In *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–6, IEEE.

Inácio, Marcio and Thiago Pardo. 2021. Semantic-based opinion summarization. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 619–628.

Knight, Kevin, Bianca Badarau, Laura Banarescu, Claire Bonial, Madalina Bardocz, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Martha Palmer, Tim O'Gorman, and Nathan Schneider. 2017. Abstract Meaning Representation (AMR) Annotation Release 2.0. Technical Report LDC2017T10, Linguistic Data Consortium, Philadelphia, PA.

Knight, Kevin, Bianca Badarau, Laura Baranescu, Claire Bonial, Madalina Bardocz, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Martha Palmer, Tim O'Gorman, et al. 2021. Abstract Meaning Representation (AMR) annotation release 3.0.

Knight, Kevin, Laura Banarescu, Claire Bonial, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Martha Palmer, and Nathan Schneider. 2014. Abstract Meaning Representation (AMR) Annotation Release 1.0. Technical Report LDC2014T12, Linguistic Data Consortium, Philadelphia, PA.

Koppel, Moshe and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Association for Computational Linguistics, Portland, Oregon, USA.

Koto, Fajri, Jey Han Lau, and Timothy Baldwin. 2021. Evaluating the efficacy of summarization evaluation across languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 801–812, Association for Computational Linguistics, Online.

Kouris, Panagiotis, Georgios Alexandridis, and Andreas Stafylopatis. 2022. Text summarization based on semantic graphs: An Abstract Meaning Representation graph-to-text deep learning approach.

Lee, Young-Suk, Ramon Fernandez Astudillo, Thanh Lam Hoang, Tahira Naseem, Radu Florian, and Salim Roukos. 2021. Maximum bayes smatch ensemble distillation for AMR parsing. *arXiv preprint arXiv:2112.07790*.

Leung, Wai Ching, Shira Wein, and Nathan Schneider. 2022. Semantic similarity as a window into vector- and graph-based metrics. In *Proceedings of the 2nd Workshop*

*on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 106–115, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid).

Li, Bin, Yuan Wen, Weiguang Qu, Lijun Bu, and Nianwen Xue. 2016. Annotating the little prince with Chinese AMRs. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 7–15, Association for Computational Linguistics, Berlin, Germany.

Li, Changmao and Jeffrey Flanigan. 2022. Improving neural machine translation with the Abstract Meaning Representation by combining graph and sequence transformers. In *Proceedings of the 2nd Workshop on Deep Learning on Graphs for Natural Language Processing (DLG4NLP 2022)*, pages 12–21, Association for Computational Linguistics, Seattle, Washington.

Li, Xiang, Thien Huu Nguyen, Kai Cao, and Ralph Grishman. 2015. Improving event detection with Abstract Meaning Representation. In *Proceedings of the First Workshop on Computing News Storylines*, pages 11–15, Association for Computational Linguistics, Beijing, China.

Linh, Ha 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, Association for Computational Linguistics, Florence, Italy.

Liu, Fei, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A. Smith. 2015. Toward abstractive summarization using semantic representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086, Association for Computational Linguistics, Denver, Colorado.

Manning, Emma, Shira Wein, and Nathan Schneider. 2020. A human evaluation of AMR-to-English generation systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4773–4786, International Committee on Computational Linguistics, Barcelona, Spain (Online).

Mansouri, Behrooz, Douglas W Oard, and Richard Zanibbi. 2022. Contextualized formula search using math Abstract Meaning Representation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4329–4333.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Migueles-Abraira, Noelia. 2017. A study towards Spanish Abstract Meaning Representation. Master's thesis, University of the Basque Country.

Migueles-Abraira, Noelia, 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)*, European Language Resources Association (ELRA), Miyazaki, Japan.

Montariol, Syrielle and Alexandre Allauzen. 2021. Measure and evaluation of semantic divergence across two languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1247–1258, Association for Computational Linguistics, Online.

Nguyen, Long HB, Viet H Pham, and Dien Dinh. 2021. Improving neural machine translation with AMR semantic graphs. *Mathematical Problems in Engineering*, 2021.

Nikolaev, Dmitry, Ofir Arviv, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Maria Saeboe, and Omri Abend. 2020. Fine-grained analysis of cross-linguistic syntactic divergences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1159–1176, Association for Computational Linguistics, Online.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, European Language Resources Association (ELRA), Portorož, Slovenia.

van Noord, Rik, Lasha Abzianidze, Hessel Haagsma, and Johan Bos. 2018. Evaluating scoped meaning representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European

Language Resources Association (ELRA), Miyazaki, Japan.
Opitz, Juri, Angel Daza, and Anette Frank. 2021. Weisfeiler-leman in the bamboo: Novel AMR graph metrics and a benchmark for AMR graph similarity. *Transactions of the Association for Computational Linguistics*, 9:1425–1441.
Opitz, Juri and Anette Frank. 2022. SBERT studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 625–638, Association for Computational Linguistics, Online only.
Opitz, Juri, Philipp Heinisch, Philipp Wiesenbach, Philipp Cimiano, and Anette Frank. 2021. Explainable unsupervised argument similarity rating with Abstract Meaning Representation and conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 24–35, Association for Computational Linguistics, Punta Cana, Dominican Republic.
Opitz, Juri, Letitia Parcalabescu, and Anette Frank. 2020. AMR similarity metrics from principles. *Transactions of the Association for Computational Linguistics*, 8:522–538.
Opitz, Juri, Shira Wein, Julius Steen, Anette Frank, and Nathan Schneider. 2023. AMR4NLI: Interpretable and robust NLI measures from semantic graphs. *arXiv preprint arXiv:2306.00936*.
Oral, Elif, Ali Acar, and Gülşen Eryiğit. 2022. Abstract meaning representation of Turkish. *Natural Language Engineering*, pages 1–30.
Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA.
Popović, Maja. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Association for Computational Linguistics, Copenhagen, Denmark.
Potthast, Martin, Alberto Barrón-Cedeno, Benno Stein, and Paolo Rosso. 2011. Cross-language plagiarism detection. *Language Resources and Evaluation*, 45(1):45–62.
Procopio, Luigi, Rocco Tripodi, and Roberto Navigli. 2021. 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*, pages 325–337, Association for Computational Linguistics, Online.
Rao, Sudha, Daniel Marcu, Kevin Knight, and Hal Daumé III. 2017. Biomedical event extraction using Abstract Meaning Representation. In *BioNLP 2017*, pages 126–135, Association for Computational Linguistics, Vancouver, Canada,.
Ribeiro, Leonardo F. R., Jonas Pfeiffer, Yue Zhang, and Iryna Gurevych. 2021. Smelting gold and silver for improved multilingual AMR-to-Text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 742–750, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.
Richens, Richard H. 1958. Interlingual machine translation. *The Computer Journal*, 1(3):144–147.
Roth, Michael and Talita Anthonio. 2021. UnImplicit shared task report: Detecting clarification requirements in instructional text. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 28–32, Association for Computational Linguistics, Online.
Saldías, Belén, George Foster, Markus Freitag, and Qijun Tan. 2022. Toward more effective human evaluation for machine translation. *arXiv preprint arXiv:2204.05307*.
Saphra, Naomi and Adam Lopez. 2015. AMRICA: an AMR inspector for cross-language alignments. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 36–40, Association for Computational Linguistics, Denver, Colorado.
Schwenk, Holger, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.
Smith, Jason R., Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of*

*the Association for Computational Linguistics*, pages 403–411, Association for Computational Linguistics, Los Angeles, California.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA.

Sobrevilla Cabezudo, Marco Antonio 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, Association for Computational Linguistics, Florence, Italy.

Song, Linfeng and Daniel Gildea. 2019. SemBleu: A robust metric for AMR parsing evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4547–4552, Association for Computational Linguistics, Florence, Italy.

Song, Linfeng, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using AMR. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Takhshid, Reza, Razieh Shojaei, Zahra Azin, and Mohammad Bahrani. 2022. Persian Abstract Meaning Representation. *arXiv preprint arXiv:2205.07712*.

Trott, Sean, Tiago Timponi Torrent, Nancy Chang, and Nathan Schneider. 2020. (Re)construing meaning in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5170–5184, Association for Computational Linguistics, Online.

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

Urešová, Zdeňka, Jan Hajič, and Ondřej Bojar. 2014. Comparing Czech and English AMRs. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 55–64, Association for Computational Linguistics and Dublin City University, Dublin, Ireland.

Van Gysel, Jens E., Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, 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*.

Van Gysel, Jens E. L., Meagan Vigus, Pavlina Kalm, Sook-kyung Lee, Michael Regan, and William Croft. 2019. Cross-linguistic semantic annotation: Reconciling the language-specific and the universal. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 1–14, Association for Computational Linguistics, Florence, Italy.

Vu, Sinh Trong, Minh Le Nguyen, and Ken Satoh. 2022. Abstract Meaning Representation for legal documents: an empirical research on a human-annotated dataset. *Artificial Intelligence and Law*, pages 1–23.

Vyas, Yogarshi, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. 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 1503–1515, Association for Computational Linguistics, New Orleans, Louisiana.

Wang, Chuan, Bin Li, and Nianwen Xue. 2018. Transition-based Chinese AMR parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 247–252, Association for Computational Linguistics, New Orleans, Louisiana.

Wein, Shira, Lucia Donatelli, Ethan Ricker, Calvin Engstrom, Alex Nelson, Leonie Harter, and Nathan Schneider. 2022a. Spanish Abstract Meaning Representation: Annotation of a general corpus. In *Northern European Journal of Language Technology, Volume 8*, Northern European Association of Language Technology, Copenhagen, Denmark.

Wein, Shira, Wai Ching Leung, Yifu Mu, and Nathan Schneider. 2022b. Effect of source language on AMR structure. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 97–102, European Language Resources Association,

Marseille, France.
Wein, Shira and Nathan Schneider. 2021. Classifying divergences in cross-lingual AMR pairs. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 56–65, Association for Computational Linguistics, Punta Cana, Dominican Republic.
Wein, Shira and Nathan Schneider. 2022. Accounting for language effect in the evaluation of cross-lingual AMR parsers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3824–3834, International Committee on Computational Linguistics, Gyeongju, Republic of Korea.
Wein, Shira, Zhuxin Wang, and Nathan Schneider. 2023. Measuring fine-grained semantic equivalence with Abstract Meaning Representation. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 144–154, Association for Computational Linguistics, Nancy, France.
Xu, Dongqin, Junhui Li, Muhua Zhu, Min Zhang, and Guodong Zhou. 2021a. XLPT-AMR: Cross-lingual pre-training via multi-task learning for zero-shot AMR parsing and text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 896–907, Association for Computational Linguistics, Online.
Xu, Weiwen, Huihui Zhang, Deng Cai, and Wai Lam. 2021b. Dynamic semantic graph construction and reasoning for explainable multi-hop science question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1044–1056, Association for Computational Linguistics, Online.
Xue, Nianwen, Ondřej Bojar, Jan Hajič, Martha Palmer, Zdeňka Urešová, and Xiuhong Zhang. 2014. Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1765–1772, European Language Resources Association (ELRA), Reykjavik, Iceland.
Xue, Nianwen, Xiuhong Zhang, Zixin Jiang, Martha Palmer, Fei Xia, Fu-Dong Chiou, and Meiyu Chang. 2013. Chinese Treebank 8.0. Technical Report LDC2013T21, Linguistic Data Consortium, Philadelphia, PA.
Žabokrtský, Zdeněk, Daniel Zeman, and Magda Ševčíková. 2020. Sentence meaning representations across languages: What can we learn from existing frameworks? *Computational Linguistics*, 46(3):605–665.
Zhai, Yuming, Gabriel Illouz, and Anne Vilnat. 2020. Detecting non-literal translations by fine-tuning cross-lingual pre-trained language models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5944–5956, International Committee on Computational Linguistics, Barcelona, Spain (Online).
Zhai, Yuming, Aurélien Max, and Anne Vilnat. 2018. Construction of a multilingual corpus annotated with translation relations. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 102–111, Association for Computational Linguistics, Santa Fe, New Mexico, USA.
Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proc. of ICLR*, Online.
Zhang, Zixuan and Heng Ji. 2021. Abstract Meaning Representation guided graph encoding and decoding for joint information extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–49, Association for Computational Linguistics, Online.
Zhang, Zixuan, Nikolaus Parulian, Heng Ji, Ahmed Elsayed, Skatje Myers, and Martha Palmer. 2021. Fine-grained information extraction from biomedical literature based on knowledge-enriched Abstract Meaning Representation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6261–6270, Association for Computational Linguistics, Online.
Zhao, Ming, Yaling Wang, and Yves Lepage. 2022. Large-scale AMR corpus with re-generated sentences: Domain adaptive pre-training on ACL Anthology corpus. In *2022 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 19–24.
Zhu, Huaiyu, Yunyao Li, and Laura Chiticariu. 2019. Towards universal semantic representation. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 177–181, Association for Computational Linguistics,

Florence, Italy.