Spanish Abstract Meaning Representation: Annotation of a General Corpus

Shira Wein, Georgetown University, Washington, DC, USA sw1158@georgetown.edu

Lucia Donatelli, Saarland University, Germany donatelli@coli.uni-saarland.de

Ethan Ricker, Georgetown University, Washington, DC, USA ear131@georgetown.edu

Calvin Engstrom, Georgetown University, Washington, DC, USA cle41@georgetown.edu

Alex Nelson, Georgetown University, Washington, DC, USA amn106@georgetown.edu

Leonie Harter, Saarland University, Germany leonie-harter@web.de

Nathan Schneider, Georgetown University, Washington, DC, USA nathan.schneider@georgetown.edu

Abstract Abstract Meaning Representation (AMR), originally designed for English, has been adapted to a number of languages to facilitate cross-lingual semantic representation and analysis. We build on previous work and present the first sizable, general annotation project for Spanish AMR. We release a detailed set of annotation guidelines and a corpus of 486 gold-annotated sentences spanning multiple genres from an existing, cross-lingual AMR corpus. Our work constitutes the second largest non-English gold AMR corpus to date. Fine-tuning an AMR-to-Spanish generation model with our annotations results in an absolute BERTScore improvement of 8.8%, demonstrating initial utility of our work.

1 Introduction

Abstract Meaning Representation (AMR) represents the core meaning of a sentence as a directed, rooted graph focused on predicate-argument structure (Banarescu et al., 2013) (figure 1). Nodes correspond to concepts and labels denote relations between concepts. Labels can be core roles functioning as predicates or arguments, or other attributes such as `:location` or `:manner`.

While there are large AMR-annotated corpora available for English, cross-lingual adaptations of AMR are necessary if AMR is to be useful as an interlingua or intermediate representation for cross-lingual tasks (Xue et al., 2014). Recent work has adapted AMR to a variety of languages (§2.1), evaluating cross-lingual efficacy of rolesets, word senses, and how effectively AMR relations capture "who is doing what to whom" in languages other than English.

As AMR aims to abstract away from morphosyntax, its graph structure is closer to logic than a syntactic parse. For English, AMR removes information such as number, definiteness, tense, word class, and word order. Yet, in many languages, morphosyntactic information in languages other than English carries rich, important semantic information beyond the "sugar" AMR intends to avoid. Therefore, it is important when developing non-English AMR annotation schema to both consider consistency with work in other languages (primarily English) as well as effectively reflecting the semantics of the language being annotated as much as possible.

Spanish is one of the most widely spoken languages in the world. There has been one previous proposal for adapting AMR to Spanish: Migueles-Abraira et al. (2018) presented a corpus of 50 representative annotations for a Spanish translation of (*The Little Prince*) (LPP) (§2.2). While Migueles-Abraira et al. (2018) noted that English AMR failed

```
(a) (s / say-01
        :ARG0 (p / prince
            :mod (l / little))
        :ARG1 (h / hurry-01)
            :ARG1 (t / they)
            :degree (g / great))
(b) (d / decir-01
        :ARG0 (p / príncipe
            :mod (p2 / pequeño))
        :ARG1 (a / apresurado
            :domain (t / th-pers-pl-sinnombre)
            :degree (m / muy)))
```

Figure 1: English (a) and Spanish (b) AMRs for the sentence *"They are in a great hurry," said the little prince.* (*"Tienen mucha prisa," dijo el principito.*) in PENMAN/text-based notation. The Spanish annotation from Migueles-Abraira et al. (2018) is adapted to our schema; `th-pers-pl-sinnombre` is an abbreviation of `third-person-plural-sinnombre` (§3.5) in this example AMR.

to adequately capture semantic phenomena in Spanish, they indicated that accurate representation could be accomplished by adding specific roles and constructions. For example, the English and Spanish AMRs in figure 1, which annotate parallel sentences, have two syntactic divergences due to inherent differences between the languages (Wein and Schneider, 2021).

We extend this prior work on Spanish AMR and present the first substantial Spanish AMR corpus of 486 gold-annotated Spanish AMRs (§4). Specifically, we annotate the Spanish sentences from the "Abstract Meaning Representation 2.0 - Four Translations" dataset (Damonte and Cohen, 2020), a corpus from the news domain that has become a popular resource for evaluation of cross-lingual AMR parsers (Blloshmi et al., 2020; Procopio et al., 2021; Cai et al., 2021) and that spans more genres than LPP.

To support the annotation, we develop annotation guidelines that update and complete those previously established for Spanish (§3). As with prior work, we find that AMR's principle of abstracting away from morphosyntax creates challenges for representing meaning in agreement-rich languages such as Spanish; we present solutions that may be extendable to other languages that exhibit similar linguistic phenomena (§3.14). Our work adds to the development of non-English AMR schema and discusses how to balance consistency and compatibility with standard English AMR while capturing pertinent semantic information not explicitly encoded in English. Three annotators were involved (§4); their work is verified with detailed analysis of inter-annotator agreement and disagreement (§5). Our annotations are publicly available on GitHub.[1]

Finally, to underscore the utility of our gold annotations, we conduct an initial evaluation for a cross-lingual generation task (§6). We show that by fine-tuning an AMR-to-Spanish generation model we are able to achieve an 8.8% increase in BERTScore (Zhang et al., 2019) performance.

2 Related Work

2.1 Cross-lingual Adaptations of AMR

Though AMR was originally designed for English (Banarescu et al., 2013), AMR's abstraction away from morphosyntactic variation lends itself to cross-lingual adaptation by capturing shared semantic structure (Li et al., 2016). Cross-lingual adaptations of AMR have been developed and evaluated for Czech (Hajič et al., 2014), Chinese (Xue et al., 2014; Li et al., 2016), Spanish (Migueles-Abraira et al., 2018), Vietnamese (Linh and Nguyen, 2019), Korean (Choe et al., 2020), Portuguese (Sobrevilla Cabezudo and Pardo, 2019; Anchiêta and Pardo, 2018; Inácio et al., 2022), Turkish (Azin and Eryiğit, 2019; Oral et al., 2022), Persian (Takhshid et al., 2022), and Celtic languages (Heinecke and Shimorina, 2022).

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. For example, 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).

[1] The annotations are available at `https://github.com/shirawein/Spanish-Abstract-Meaning-Representation.git`. The associated sentences are available through the Linguistic Data Consortium.

2.2 Prior Work Adapting AMR to Spanish

Prior work has proposed an initial adaptation of AMR to Spanish (Migueles-Abraira et al., 2018) using English AMR guidelines (Banarescu et al., 2019) as a baseline to pilot annotation for Spanish sentences. Seven key linguistic phenomena were identified as necessary to add to English AMR to capture essential semantic information in Spanish: (1) NP ellipsis, (2) third person possessive pronouns, (3) third person clitic pronouns, (4) varied *se* usage, (5) gender, (6) verbal periphrases/verbal structure and locutions, and (7) double negatives. Guidelines were developed for the first four of these phenomena, and 50 representative sentences of the Spanish translation of *The Little Prince* were annotated. Spanish translations were made to be more literal so that they would be more semantically equivalent to the original translation of the work.

One limitation of the previous approach was the use of English PropBank (Kingsbury and Palmer, 2002; Palmer et al., 2005) for sense annotation instead of AnCora (Taulé et al., 2008) (§4.4), a similar resource developed for Spanish. English PropBank senses do not correspond one-to-one with their Spanish verbs and bias word meanings towards English-based semantics. Migueles-Abraira et al. (2018) chose rolesets from English PropBank instead of AnCora as it provided more coverage of words in the corpus. Spanish words were translated to English, and the sense from the English word was attached to the Spanish word (Migueles-Abraira, 2017).

A second limitation of the previous Spanish AMR annotation was the limited amount of change to the English AMR guidelines to incorporate Spanish linguistic phenomena. Recent work has assessed various differences between Spanish and English annotations of the existing Spanish AMR adaptation, classifying the type and cause of the identified differences (Wein and Schneider, 2021).

3 Aims and Guidelines

Our primary aims with the development of this corpus included the release of a (1) sizable, (2) general-purpose Spanish AMR corpus, which can be useful in the evaluation of cross-lingual AMR parsers, (3) which effectively represents Spanish semantics. We set out to meet these goals by (1) manually annotating 586 AMRs, (2) annotating the Four Translations dataset, often used for evaluation of cross-lingual AMR parsers, and (3) developing guidelines which consider a range of linguistic phenomena. In this section, we discuss the key considerations and linguistic phenomena we prioritize in our approach to Spanish AMR annotation.

3.1 Use of English and Connection to English AMR Guidelines

Our guidelines are developed in reference to the English AMR Guidelines,[2] outlining the differences between our annotation schema of Spanish sentences and the annotation for English AMRs. As has been popularized in other non-English AMR corpora (Linh and Nguyen, 2019; Sobrevilla Cabezudo and Pardo, 2019), we maintain the role labels and canonical entity type list in English. For example, we use `:ARG0`, `:ARG1`, etc., as well as `:domain`, `:time`, etc., and `person`, `government-organization`, `location`, etc.

3.2 Verb Senses

We number verb senses according to the AnCora lexicon,[3] and supplement these with new senses for out-of-vocabulary lexemes and meanings encountered in our data (table 1). Usage examples for these senses are included in the guidelines.

3.3 Modality

The modal verbs *deber* ("must", "should") and *poder* ("might", "could") appear in table 1 in the list of words which appear in AnCora with other senses. Though meanings of *deber* and *poder* do appear in AnCora, we establish additional senses to mark modality. These modals take the same `:ARG1` structure as do their English modal equivalents—`recommend-01` and `possible-01`, respectively. These modals take the verb senses `deber-03` and `poder-04`.

[2] https://github.com/amrisi/amr-guidelines/blob/master/amr.md

[3] http://clic.ub.edu/corpus/en/ancoraverb_es

Verb	AnCora?	S#	English Translation
auditar	no	-01	to audit
disuadir	no	-01	to dissuade
vagar	no	-01	to wander
hervir	no	-01	to boil
desvanecer	yes	-02	to fade
sobrecargar	no	-01	to overload
congestionar	no	-01	to congest (traffic)
incriminar	no	-01	to incriminate
circunvalar	no	-01	to encircle
adular	no	-01	to flatter
salir	yes	-11	to go out (with someone)
entrelazar	no	-01	to interlace
zonificar	no	-01	to zone
embotellar	no	-01	to bottle up
deber	yes	-03	[modal] to recommend
poder	yes	-04	[modal] to be possible

Table 1: Table of verb senses specification for annotation of senses which are not covered by AnCora. The "Ancora?" column indicates whether the verb is included at all in AnCora, for any senses. If the verb appears for other sense, the S# (sense number) increases to the next available label.

3.4 Gender

In Spanish, all nouns have lexical gender (masculine or feminine), which affects agreement. Nouns relating to humans or animals will also be marked with natural/interpretable gender, such as *hermano* ("brother") versus *hermana* ("sister"). Either way, we remove only number information when lemmatizing the word for AMR, so *niños* (whether it means "boys", or "boys and girls") will always be represented with the concept *niño*, and *niñas* ("girls") with *niña*. If any agreeing adjectives appear, the gendered (singular) concept is annotated.

3.5 Pronoun Drop

Spanish belongs to a group of languages that allow pronoun drop (pro-drop), in which certain pronouns can be omitted if they are grammatically or pragmatically inferable from the surrounding linguistic context. Pro-drop in Spanish occurs only with subject pronouns and is permitted only in certain contexts (Española, 2010).[4] Migueles-Abraira et al. (2018) specify a special concept `sinnombre` ("nameless") for implicit references where no antecedent in context is represented in the AMR. We refine this approach to also encode person and number for these implicit entities following the standard format: `first-person-sing-sinnombre`, `first-person-plural-sinnombre`, etc.

For example, in *No sé que quiero* ("I do not know what I want"), there is an implicit subject *yo* ("I") that is reflected in the verbal agreement. We therefore specify `first-person-sing-sinnombre` as the agent. We choose to use `first-person-sing-sinnombre` instead of the reentrant `yo` ("I") as the conditions on the use of overt and dropped pronouns are typically subject to information structure, an important component of sentence meaning.

No sé que quiero. ("I do not know what I want.")

```
(s / saber-01
    :polarity -
    :ARG0 (f / first-person-sing-sinnombre)
    :ARG1 (h / querer-01
        :ARG0 f))
```

If the pronoun is present (e.g. *él, ella, usted,* etc.), the pronoun should be used in place of a `sinnombre` concept.

3.6 Polite Second Person Addressee

Usted ("you") can reflect either a polite usage of second person, or third person. When *usted* is used as a polite second person pronoun, the polite modifier should be added: `:mod-polite +`. This follows the same structure as `:polarity -`.

3.7 Third Person Possessives

We treat third person possessives similarly to the English annotation, using the `sinnombre` concepts discussed above. For example, we annotate *su coche* ("his car") the same way that "his car" is structured.

his car

```
(c / car
    :poss (h / he))
```

[4] Subject drop is viable in Spanish due to inflection of person and number in the verb. Other pro-drop languages permit the elision of pronouns in other positions. Future work can look at the impact of AMR's abstraction away from morphosyntactic information that allows phenomena such as pro-drop, especially in translation and generation tasks.

su coche (“his car”)

```
(c / coche
    :poss (e / third-person-sing-sinnombre))
```

The possessive pronoun *su* is ambiguous (“his”/“hers”/“its”), and could be annotated as `third-person-sing-sinnombre` (in the case of “his”), `second-person-sing-sinnombre` (as in “yours”), or `third-person-plural-sinnombre` (for “theirs”). These labels are only required when the use of *su* as a possessive pronoun is ambiguous. For example, in the case of *Sofía me mostró su auto* (“Sofía showed me her car”), *su* very likely refers to Sofia’s. However, in *Sofía copió su tarea* (“Sofía copied their homework”), this likely means that Sofia copied someone else’s homework; *su* would refer to some unnamed person, and would thus require the use of `third-person-sing-sinnombre`. Because *su* covers all third person possessives, this distinction requires some interpretation by the annotator based on context and meaning.

3.8 Third Person Clitic Pronouns

Clitics are treated as separate concepts, following (Migueles-Abraira et al., 2018). For example, *mandarlo* (“send it”) has a root of *mandar* (“send”) and an ARG1 of the item being sent: *lo* (“it”).

```
(m / mandar-01
    :ARG1 (l / lo))
```

3.9 Se Usage

Se has many uses in Spanish, including: (1) as a reflexive pronoun, (2) to denote the passive voice, (3) as a substitute for the indirect pronoun *le/les*, and (4) as an impersonal pronoun.

Se as a Reflexive Pronoun. Reflexives are represented via reentrancies as in English AMR. Two examples include the use of *se* in *ellos se perjudican* (“they are harmed”) and in *Pablo se ve* (“Pablo sees himself”).

Ellos se perjudican. (“They harm themselves.”)

```
(p / perjudicar-01
    :ARG0 (e / ellos)
    :ARG1 e)
```

Pablo se ve. (“Pablo sees himself.”)

```
(v / ver-01
    :ARG0 (p / person
        :name (n / name
            :op1 "Pablo"))
    :ARG1 p)
```

Se as a Passive Marker. When *se* reflects a passive voice for an omitted concept, we use the `:ARG0` role label with *se*.

Se venden casas rurales. (“Rural houses for sale.”)

```
(v / vender-01
    :ARG0 (s / se)
    :ARG1 (c / casa
        :mod (r / rural)))
```

Se as an Impersonal Pronoun. *Se* used to mean “one” is annotated with the concept `se-impersonal`.

No se debe beber. (“One should not drink.”)

```
(d / deber-03
    :polarity -
    :ARG0 (b / beber-01)
    :ARG1 (s / se-impersonal))
```

3.10 Double Negation

In Spanish, negation can be indicated by either single or double negatives, with double negatives sometimes providing emphasis. We annotate both single and double negation with the use of one polarity marker.

No hay ninguna persona. (“There is nobody.”)

```
(h / haber-01
    :polarity -
    :ARG0 (p / persona))
```

3.11 Suffixes

Derivational suffixes such as diminutives should be represented as modifier concepts. For example, *poquito* (“very little”) would be annotated with *poco* (“little”) being modified by *muy* (“very”).

```
(p / poco
    :mod (m / muy))
```

Another example would be *hombrecito* (“little man”), for which would *hombre* (“man”) receive the diminutive modifier of *pequeño* (“little”).

```
(h / hombre
    :mod (p / pequeño))
```

3.12 Words that Change Meaning When Singular Or Plural

In Spanish AMR as in English AMR, we annotate the concept as the singular of the entity even if it is plural. However, rarely in Spanish a word changes meaning if it is plural instead of singular. In this case we use the plural form of the word, such as *deber* (duty) versus *deberes* (homework), or *resto* (remainder) versus *restos* (human remains or rubbish). Additionally, we distinguish *algún* from *algunos*, for the case in which *algún* means “any” and *algunos* means “some.” Similarly, we distinguish *otros* (“others”) as a plural noun to mean a distinct group of “others,” and preserve the plural *otros* instead of making it singular as *otro* (“other”).

3.13 Comparison with Previous Work

The most notable difference between our approach and that of Migueles-Abraira et al. (2018) is that theirs uses Spanish labels while ours uses English labels. Additional differences are largely due to our choice to break down the unnamed category of dropped entities into subcategories based on the type of noun phrase or pronoun. For NP ellipses (§3.5) and third person possessives (§3.7), we use the 6 tags outlined, which specify person and number. Migueles-Abraira et al. (2018) uses a standardized `ente` (“being”) concept with `sinnombre` (“nameless”) argument for NP ellipses and a `sinespecificar` (“unspecified”) argument for third person possessives. In comparison to our annotation in §3.7 for *su coche* (“his car”), the annotation in the corpus from Migueles-Abraira et al. (2018) separates entities (`ente`) and the possessive pronoun itself. Notably, this annotation focuses more on the morphosyntax than semantics:

```
(c / coche
    :posee (e / ente
        :sinespecificar (s / su)))
```

Our approach as well as that of Migueles-Abraira et al. (2018) represents clitics as if they were separated from the stem. We also both approach *se* as a reflexive pronoun in the same way via reentrancy. However, the approach of previous work omits *se* when it is used in the impersonal or passive voice, which we include via the `se-impersonal` concept and `ARG0` label, respectively (§3.9). We also address the issues of *se* as a substitute for *le* or *les* (§3.9), modality (§3.3), gender (§3.3), polite use of *usted* (“you”) (§3.6), double negation (§3.10), diminutive and augmentative suffixes (§3.11), meaning change in the singular versus plural (§3.12), and commas/decimals.

3.14 Limitations

Adapting standard English AMR to Spanish involves striking a balance between faithfully capturing the semantics of the Spanish sentence on the one hand, and mirroring the English annotation schema on the other. Here we discuss a few challenges.

Gender and Number Marking. The construction of Spanish interpretable/natural gender and its relationship to morphosyntax are open questions (Donatelli, 2019). In our annotation schema, we opted for simplicity, choosing not to explicitly annotate gender, but to leave any gender-bearing morphology as is in the concept. Migueles-Abraira (2017) encodes gender explicitly by converting all nouns to their masculine form, and adding a `:masc` or `:fem` role label.

Like in English AMR, number inflection is removed unless that would alter the meaning of the stem (§3.12). The possibility of encoding number and gender more explicitly is left to future work.

Idiomatic Expressions. Idiomatic expressions are difficult to annotate with AMR. As is the case for English, Spanish has numerous idiomatic expressions, phrases that have a meaning different to that of individual words in the phrase. Idiomatic expressions are annotated on a case-by-case basis. In the corpus, the majority of idiomatic expressions are either condensed into one concept (*por supuesto*, “of course,” becomes `por-supuesto`), or we must use a similar, pre-existing verb to convey the expression’s meaning, such as *tener prisa* (“to be in a rush”).

Limitations with AnCora. AnCora’s predicate lexicon only includes verbs, unlike English Prop-

Bank (Palmer et al., 2005), which has been extended beyond verbs to include noun, adjective, and complex predicates (Bonial et al., 2014). AnCora notably lacks adjective frames and numerous idiomatic/phrasal verbs. This posed a challenge when annotating many adjectives and (often more colloquial) verb phrases. When handling idiomatic verb usage, it is easy (but problematic) for annotators to default to using the structure of the equivalent English idiomatic structure, and substitute Spanish tokens into the English structure. Some AnCora rolesets were missing important core roles. Expanding AnCora or other Spanish propbank efforts would enhance any AMR annotations relying on it.

Mood. Spanish exhibits three grammatical moods: indicative, imperative, and subjunctive. English AMR assumes all sentences to be in indicative mood unless otherwise marked. There are two categories for additional moods: imperatives are marked with `:mode imperative` and expressive utterances with `:mode expressive`. As this is a very rudimentary treatment of the semantics of mood, we choose not to adapt it for Spanish AMR. Future work will look at how to integrate the subjunctive mood into Spanish AMR at both the verbal and sentential levels.

4 Annotation Methodology

4.1 Dataset

We perform annotations on the "AMR 2.0 - Four Translations" dataset, which is released through the Linguistic Data Consortium (Damonte and Cohen, 2020) and has become a popular evaluation tool for cross-lingual AMR parsers (Blloshmi et al., 2020; Procopio et al., 2021; Cai et al., 2021). This dataset contains gold AMRs for English test split sentences from the AMR Annotation Release 2.0 (Knight et al., 2017) alongside translations of those sentences into Italian, Spanish, German, and Mandarin Chinese. The sentences originate mostly from news sources, including broadcast conversations, newswire and web text—genres broader than but complementary to the LPP corpus often used for AMR annotation. The corpus contains 1,371 Spanish sentences and 5,484 sentences total. Of the 1,371 Spanish sentences, we directly annotate 486, encompassing 9,540 words. There are five documents included in the Four Datasets dataset: Proxy reports from newswire data (Proxy), translated Xinhua newswire data (Xinhua), BOLT discussion forum source data (DFA), DARPA GALE weblog and Wall Street Journal data (Consensus), and BOLT discussion forum MT data (Bolt). For Consensus, Proxy, Bolt, and DFA, we annotate the first 100 sentences of the document. Xinhua is 86 sentences in total (averaging 22.37 words per sentence), so we annotate all 86 sentences. Consensus is originally 100 sentences (averaging 15.61 words per sentence), Proxy is originally 823 sentences (averaging 23.07 words per sentence), Bolt is 133 sentences (averaging 20.25 words per sentence), and DFA is 229 sentences long (averaging 17.83 words per sentence).

4.2 Annotator Training

Three undergraduate linguistics students, native English speakers with high levels of Spanish proficiency, were first trained in English AMR annotation. Annotators were then trained in our approach to Spanish AMR annotation, through discussions of our v1.0 Spanish AMR guidelines. *The Little Prince* corpus was used for practice annotation in both languages. Once trained, the annotators moved on to annotations of the Four Translations dataset. To verify annotator understanding, we completed adjudication on the test sets of English and Spanish annotations.

4.3 Collected Annotations

To validate our approach to annotation and the reliability of our annotations, we collect annotations from all three annotators for the first 50 sentences from the Proxy document. We are then able to perform inter-annotator agreement analysis on those overlapping annotations using Smatch, presented in §5. Other than those 50 Proxy annotations, all other annotations were distributed evenly between each of the three annotators. The three annotators produced 200, 190, and 196 annotations each. This results in a total of 586 annotations total, for 486 unique sentences, with Proxy 1–50 being annotated thrice (once by each annotator). After all annotations for the initial 50 sentences were produced, a final round of corrections were made for any errors in annotation (without changing any divergent judgment calls).

AMR annotation is expensive and time-consuming. Our 586 annotations took more than 200 hours to complete including some test annotations and correction of annotations. This is also a reflection of the sentences included in the AMR 2.0 - Four Translations dataset being especially difficult to annotate due to their complicated genre and length (approx. 20 words per sentence). To maximize the number of sentences with gold annotations, we refrained from double-annotating the remainder of the data beyond the aforementioned 50 sentences.

4.4 AnCora

We use the AnCora-Net Spanish lexicon of verbs (AnCoraVerb-ES) for verb sense annotation (Taulé et al., 2008). Similar to PropBank for English, the AnCora lexicon is comprised of predicates, accompanied by their argument structures and thematic roles. Each of the 2,647 predicate entries is also related to one or more semantic classes depending on its senses. AnCora also provides a lexicon of deverbal nominalizations, AnCoraNom-ES, which contains information regarding denotative type, WordNet Synset, argument structure, and the verb from which the noun is derived. As AnCoraNom-ES significantly overlaps with AnCoraVerb-ES, we choose not to use it in this work.

For all verbs or verb senses which did not appear in the AnCora corpus, we kept track of those instances in a table and supplemented the AnCora verb bank with 16 of our own. These added senses can be seen in table 1.

4.5 StreamSide Annotation Tool

Annotations were produced using the Streamside software (Choi and Williamson, 2021). The annotators annotate tokens in the sentence as concepts, and roles and arguments are then defined between these concepts as relations. While this software allows for annotation fitted to various languages, it is best accustomed to annotation using the English because the relevant PropBank roles (Kingsbury and Palmer, 2002; Palmer et al., 2005) are automatically populated. In our case, working on Spanish and using the AnCora rolesets (Taulé et al., 2008), the annotators needed to separately reference the arguments for each concept on the AnCora website.

4.6 Guidelines Development

We developed the guidelines by first outlining our approach to key Spanish linguistic phenomena, which we identified as potentially impacting Spanish AMR annotation. Our v1.0 guidelines discuss: (1) Use of English AMR Roles and Guidelines; (2) Pronoun Drop and NP Ellipsis; (3) Third Person Possessives; (4) *Se* Usage; (5) Gender; (6) Double Negation; (7) Diminutive and Augmentative Suffixes; (8) *Estar* (to be) as a Location.

These v1.0 guidelines were developed *before* performing any annotation. Since starting annotation, there have been 9 further iterations of the guidelines, which both expand on the items included in v1.0 and incorporate additional items. We discuss the most notable elements of the guidelines in §3. After developing the first iteration of the guidelines (v1.0), any further changes required to the guidelines, as identified during the annotation process, were incorporated into the next iteration. All existing annotations were then uniformly altered by their annotators to match the most updated guidelines.

5 Evaluation

5.1 Inter-Annotator Agreement

Table 2 shows the inter-annotator agreement (IAA) scores for each pair of annotators on the 50 triple-annotated Proxy sentences. The IAA scores were calculated by averaging the Smatch scores across the 50 sentence pairs for the annotators. The Smatch (Cai and Knight, 2013) algorithm calculates the amount of overlap between the AMR graphs to determine similarity. Smatch using a hill-climbing method to determine the optimal alignment between the variables in the AMR graphs and outputs an F-score from 0 to 1, where 1 indicates that the AMRs are isomorphic.

The average IAA scores ranged from 0.83–0.89, a very promising range for AMR annotation agreement. Comparable work achieved Smatch inter-annotator agreement scores of 0.79 (Choe et al., 2020), 0.72 (Sobrevilla Cabezudo and Pardo, 2019), and 0.83 (Li et al., 2016). Other work on cross-lingual AMR adaptations which only had one annotator did not report IAA/Smatch scores.

Ann. 1 & Ann. 2	0.89
Ann. 1 & Ann. 3	0.86
Ann. 2 & Ann. 3	0.83

Table 2: Average inter-annotator agreement scores (via Smatch) for each pair of our three annotators on the first 50 sentences of the Proxy document.

5.2 Disagreement Analysis

Disagreements, which we define as any discrepancy that neither violates AMR guidelines nor deviates from the sentence's meaning, were common among all three annotators. The majority of disagreements are caused by differences in interpretation.

Entity versus Event Annotation. AMR takes a predicate-centric approach to annotation. While verbs are typically annotated as events and nouns are annotated as entities (concepts without a number), when nouns or phrases have verbal counterparts, this can cause differences among annotators. For example, *propuesta* ("proposal") could be annotated either as a noun or as a verb (`proponer-01`, "to propose"). We instruct annotators to annotate derived nouns as verbs and annotate related roles as arguments for increased expressivity.

Verb Sense Labels. Verb senses account for nuance in meaning depending on context. Sometimes annotators chose different rolesets when the meaning difference between senses was subtle. One notable example is the verb *reconocer* ("to recognize / acknowledge"). `Reconocer-01` refers to recognizing something as official or true, as in *reconocer el estado* ("to recognize the state"). Alternatively, `reconocer-02` maintains that meaning, but often precedes a subordinate clause, as in *reconocen que gané* ("they acknowledge that I won").

Non-Core Role Overlap. Finally, annotators had difficulty consistently choosing the same non-core role (`:poss`, `:mod`, etc.) when the roles could overlap in meaning. For example, *la carta del hombre* ("the man's letter") could be annotated differently depending on the interpretation of the man's relationship to the letter. An emphasis on the man's ownership of the letter elicits the `:poss` role, whereas emphasizing the letter's creation by the man elicits the `:source` role.

t5wtense	0.7389
Fine-tuned t5wtense	0.8265
XLPT-AMR	0.8534

Table 3: BERTscore results for: the output of the t5wtense generation model without any fine-tuning, t5wtense after fine-tuning with our data, and the state-of-the-art XLPT-AMR cross-lingual AMR generation model (Xu et al., 2021) on our test split.

6 Fine-tuning a Spanish Generation Model

AMR generation produces text from an AMR. To evaluate the utility of our dataset in practical NLP tasks, we fine-tune the t5wtense generation model of the AMR library `amrlib` to produce Spanish sentences.[5] The t5wtense generation model uses the pretrained HuggingFace T5 transformer to convert AMR graphs to text. We split our 486 annotations into 110 sentences (test) and 376 (training).[6]

We compare the fine-tuned system output and the un-tuned system output to the corresponding Spanish reference sentences from AMR 2.0 - Four Translations (Damonte and Cohen, 2020). We use BERTScore, an automatic evaluation metric for text generation (Zhang et al., 2019), to perform this comparison, as previous work has demonstrated that it is the automatic metric most correlated with human judgments for (English) AMR-to-text generation systems (Manning et al., 2020).

For evaluating Spanish text, the default BERTscore model is bert-base-multilingual-cased, which is the model we use here. Table 3 shows AMR-to-Spanish BERTscore results.

After fine-tuning t5wtense, we see a marked improvement in performance, increasing in BERTscore by approximately 8.8% absolute (11.86% relative improvement). Current state-of-the-art cross-lingual generation (Xu et al., 2021) achieves a BERTscore of 0.8534 on the same test set,[7] which indicates that by fine-tuning on only 376 Spanish AMR annotations,

[5] `https://github.com/bjascob/amrlib`

[6] We split the data as follows: Training set: Bolt 1–100, Consensus 1–100, DFA 1–40, Proxy 51-100, Xinhua 1–86; Test set: DFA 41–100, Proxy 1–50.

[7] Xu et al. (2021) report SOTA scores using BLEU. We computed BERTscore on their system's output.

we are able to achieve results close to the current best performing model.[8] The marked improvement resulting from our fine-tuning demonstrates the utility of our corpus and suggests incorporating our data into more sophisticated generation or parsing models can lead to greater improvements.[9]

7 Conclusion

We have presented an updated approach to Spanish AMR annotation which considers a broader range of meaningful linguistic phenomena than previous work. Using updated guidelines, we constructed a corpus of 486 gold-annotated Spanish AMRs for the "AMR 2.0 - Four Translations" dataset, achieving high AMR inter-annotator agreement (0.83–0.89 IAA via Smatch). Gold Spanish AMRs will contribute to ongoing evaluation and training of cross-lingual AMR models; this is substantiated by our results in §6, which improved an off-the-shelf AMR-to-Spanish generation system by fine-tuning on our data. Little prior work on AMR has set out to develop large-scale gold corpora in languages other than English; our work suggests that this is a fruitful effort, both to foster a better understanding of the cross-lingual properties of AMR and to improve system performance on non-English NLP tasks.

Acknowledgements

We thank anonymous reviewers for helpful feedback. We also thank the coordinators of the Georgetown University RULE (Research-based Undergraduate Linguistics Experience) program. This work is supported by a Clare Boothe Luce Scholarship.

References

Rafael Anchiêta 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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

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

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2019. Abstract Meaning Representation (AMR) 1.2.6 specification. `https://github.com/amrisi/amr-guidelines/blob/master/amr.md`.

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

Claire Bonial, Julia Bonn, Kathryn Conger, Jena D. Hwang, and Martha Palmer. 2014. PropBank: semantics of new predicate types. In *Proc. of LREC*, pages 3013–3019, Reykjavík, Iceland.

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

[8] The Xu et al. (2021) system performance is hindered by the fact that two of the AMRs do not produce any output at all by this model. If we remove those two AMRs from consideration, the F1 score for the Xu et al. (2021) system is slightly higher, achieving a BERTScore F1 of 0.8695, while our fine-tuned results are 0.8266 on the same sentences.

[9] Xu et al. (2021) do not release their code, so the model cannot be fine-tuned.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Hyonsu Choe, 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, Barcelona Spain (online). Association for Computational Linguistics.

Jinho D. Choi and Gregor Williamson. 2021. Streamside: A fully-customizable open-source toolkit for efficient annotation of meaning representations.

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

Lucia Elizabeth Donatelli. 2019. *The Morphosemantics of Spanish Gender: Evidence from Small Nominals.* Georgetown University.

RAE Real Academia Española. 2010. *Nueva gramática de la lengua española manual.* Espasa.

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

Johannes Heinecke 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, Marseille, France. European Language Resources Association.

Marcio Lima Inácio, Marco Antonio Sobrevilla Cabezudo, Renata Ramisch, Ariani Di Felippo, and Thiago Alexandre Salgueiro Pardo. 2022. The AMR-PT corpus and the semantic annotation of challenging sentences from journalistic and opinion texts. *SciELO Preprints.*

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

Kevin Knight, 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.

Bin Li, 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, Berlin, Germany. Association for Computational Linguistics.

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

Emma Manning, 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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

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

Elif Oral, Ali Acar, and Gülşen Eryiğit. 2022. Abstract Meaning Representation of Turkish. *Natural Language Engineering*, page 1–30.

Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31:71–106.

Luigi Procopio, 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, Online. Association for Computational Linguistics.

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

Reza Takhshid, Razieh Shojaei, Zahra Azin, and Mohammad Bahrani. 2022. Persian Abstract Meaning Representation.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Shira Wein 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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dongqin Xu, Junhui Li, Muhua Zhu, Min Zhang, and Guodong Zhou. 2021. 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, Online. Association for Computational Linguistics.

Nianwen Xue, 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, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.