https://dx.doi.org/10.1590/1678-460X202339355159

Articles

The AMR-PT corpus and the semantic annotation of challenging sentences from journalistic and opinion texts

O corpus AMR-PT e a anotação semântica de sentenças desafiadoras de textos jornalísticos e opinativos

Marcio Lima Inácio[1,2]
Marco Antonio Sobrevilla Cabezudo[3]
Renata Ramisch[4]
Ariani Di Felippo[5]
Thiago Alexandre Salgueiro Pardo[6]

[1] Universidade de Coimbra. Coimbra – Portugal. https://orcid.org/0000-0002-0875-4574. E-mail: mlinacio@dei.uc.pt

[2] Universidade de São Paulo. São Carlos – Brasil.

[3] Universidade de São Paulo. São Carlos – Brasil. https://orcid.org/0000-0001-7625-9914. E-mail: msobrevillac@usp.br

[4] Redação Nota 1000. São Paulo – Brasil. https://orcid.org/0000-0003-3372-6150. E-mail: renata.ramisch@redacaonota1000.com.br

[5] Universidade Federal de São Carlos. São Carlos – Brasil. https://orcid.org/0000-0002-4566-9352. E-mail: arianidf@gmail.com

[6] Universidade de São Paulo. São Carlos – Brasil. https://orcid.org/0000-0003-2111-1319. E-mail: taspardo@icmc.usp.br

ABSTRACT

One of the most popular semantic representation languages in Natural Language Processing (NLP) is Abstract Meaning Representation (AMR). This formalism encodes the meaning of single sentences in directed rooted graphs. For English, there is a large annotated corpus that provides qualitative and reusable data for building or improving existing NLP methods and applications. For building AMR corpora for non-English languages, including Brazilian Portuguese, automatic and manual strategies have been conducted. The automatic annotation methods are essentially based on the cross-linguistic alignment of parallel corpora and the inheritance of the AMR annotation. The manual strategies focus on adapting the AMR English guidelines to a target language. Both annotation strategies have to deal with some phenomena that are challenging. This paper explores in detail some characteristics of Portuguese for which the AMR model had to be adapted and introduces two annotated corpora: AMRNews, a corpus of 870 annotated sentences from journalistic texts, and OpiSums-PT-AMR, comprising 404 opinionated sentences in AMR.

Keywords: *corpus annotation; knowledge representation; semantics.*

RESUMO

Abstract Meaning Representation (AMR) *é uma linguagem de representação semântica bastante popular em processamento de línguas naturais (PLN). Ela codifica o significado das sentenças em grafos orientados (enraizados). Para o inglês, há um grande* corpus *com anotação AMR que subsidia métodos e aplicações de PLN. Para a anotação de* corpora *em línguas que não sejam o inglês, incluindo o português brasileiro, têm-se aplicado estratégias automáticas ou manuais. As automáticas se baseiam essencialmente no alinhamento entre* corpora *paralelos e na herança da anotação AMR, enquanto as estratégias manuais focalizam na adaptação das diretrizes originais de anotação AMR (para o inglês) em função da língua-alvo. Ambas as estratégias, automática ou manual, precisam lidar com certos fenômenos linguísticos desafiadores. Neste trabalho, exploram-se características do português para as quais o modelo AMR foi adaptado e apresentam-se dois* corpora *anotados: AMRNews,* corpus *composto por 870 sentenças anotadas, provenientes de textos jornalísticos, e o* corpus *OpiSums-PT-AMR, contendo 404 sentenças opinativas em AMR.*

Palavras-chave: *anotação de corpus; representação de conhecimento; semântica.*

1. Introduction

Natural Language Processing (NLP) is a research field that aims at developing computational systems that are able to perform tasks involving interpretation and/or generation of natural languages such as automatic translation and summarization, sentiment analysis, text simplification, and speech recognition and synthesis, among several other tasks (Jurafsky & Martin, 2008).

NLP has significantly advanced in the last decade due to the good results obtained with artificial neural networks, in particular, deep learning (Goodfellow et al., 2016) and distributional word embedding models as *word2vec* (Mikolov et al., 2013) and BERT (Devlin et al., 2019). Despite such recent advances, Natural Language Understanding (NLU) or Natural Language Interpretation (NLI) has remained as a trending challenging topic in the NLP community. Defined as the subtopic of NLP that deals with machine reading comprehension, NLU is considered an AI-hard or AI-complete problem (Yampolskiy, 2013).

Given the considerable commercial interest in NLU because of its application in large-scale content analysis, recent works focused on different semantic representation languages have emerged. Some examples are the semantic representation used in the *Groningen Meaning Bank* (Basile et al., 2012), *Universal Conceptual Cognitive Annotation* (UCCA) (Abend & Rappoport, 2013), *Universal Decompositional Semantics* (White et al., 2016), and the model used in the *Parallel Meaning Bank* (Abzianidze et al., 2017).

In the NLU scenario, *Abstract Meaning Representation* (AMR) is a very popular and prominent semantic model, which arose to answer the need to build a semantic bank that includes different semantic phenomena. It aims at encoding the meaning of a sentence with a (relatively) simple representation in the form of a directed rooted graph (Banarescu et al., 2013). This representation includes semantic roles, named entities, spatial-temporal information and polarity, among other semantic information levels.

The AMR-annotated corpus for English is large, with approximately 39,000 sentences. The Chinese AMR corpus is also of respectable size, containing 10,149 sentences[7]. Differently from such situations, there are small annotated corpora for other languages, likely due to the high complexity that building this kind of corpora represents. It is unnecessary to highlight the relevance of building corpora for other languages. Annotated corpora provide qualitative and reusable data for building or improving existing methods and applications and serving as benchmarks to compare different approaches.

Some efforts have been conducted to build AMR corpora for non-English languages. Some tried to use AMR as an interlingua and automatically mapped the alignments between parallel corpora (Anchiêta & Pardo, 2018a; Damonte & Cohen, 2018; Xue et al., 2014). In general, these works exploit an AMR parser for English and parallel corpora to learn AMR parsers for other languages (such as Italian, Spanish, German, and Chinese). Other works tried to adapt the AMR guidelines to annotate corpora in other languages[8] (Sobrevilla Cabezudo & Pardo, 2019; Migueles-Abraira et al., 2018), leveraging its cross-linguistic potential.

It is a known fact that automatic alignments can accelerate the annotation process but can also result in some limitations dealing with syntactic phenomena that account for several cross-lingual differences. For example, Damonte and Cohen (2018) mention that the automatic alignments generate AMR corpora with several mistakes, mostly involving concept identification. In another work, Anchiêta and Pardo (2018a) show that hidden subject and complex predicates are some linguistic phenomena not taken into account in the creation of an AMR corpus for Brazilian Portuguese (BP) via automatic alignment.

Manual AMR annotation can be an interesting direction for corpora building despite increasing annotation time. Some works focused on this approach are proposed by Sobrevilla Cabezudo & Pardo (2019) and Migueles-Abraira et al. (2018). However, performing manually AMR annotation for other languages, such as BP, is not a trivial task since the semantic representation model proposed in AMR is biased towards English, as stated by its original developers (Banarescu et al., 2013).

[7] Available at https://catalog.ldc.upenn.edu/ LDC2019T07.

[8] Available at https://github.com/amrisi/amr-guidelines/blob/master/amr.md and detailed at https://amr.isi.edu/doc/amr-dict.html.

To build the first version of the AMRNews corpus in BP, Sobrevilla Cabezudo & Pardo (2019) manually annotated 299 sentences belonging to several news domains[9], adapting some of the current AMR guidelines. In order to increase the size of the corpus, we recently annotated 571 more sentences using the same strategy, resulting in the version 2.0 of the corpus with 870 annotated sentences.

We also focused on the annotation of opinions, creating the OpiSums-PT-AMR corpus. Concerning a different domain from AMRNews, it enables a more semantic-focused comparative analysis between texts from both domains. Furthermore, this initiative provides data to be used in future research within the Sentiment Analysis area, as semantic knowledge can be an important feature to be taken into account in this type of processing, as argued by Cambria et al. (2015). To this extent, we used as basis the OpiSums-PT corpus (López Condori et al., 2015), comprising 1,502 sentences from comments about 17 different products, among which 404 have been annotated in AMR within the scope of this paper.

In this work, we explore the BP challenging linguistic phenomena for which the AMR model had to be adapted and present and detail the two annotated corpora: the AMRNews and the OpiSums-PT-AMR corpora. We also take advantage of the different domains of each AMR corpus and present a comparative analysis between opinions and news, highlighting important differences on the occurrence of semantic phenomena between each type of text.

This paper is organized as follows. Section 2 introduces the AMR fundamentals. In Section 3, we describe some works related to AMR corpus building. Afterwards, in Section 4, we present both the AMRNews and Opisums-PT-AMR corpora and report their annotation methodology and evaluation. We also perform a statistical description of each corpus together with some comparative analysis between this data. In Section 5, we explore some phenomena of our corpora in Portuguese and the correspondent adaptations of the English guidelines. Finally, some final remarks are presented in Section 6.

[9] Differently from the previous work focused on the “The Little Prince” book (Anchiêta & Pardo, 2018a).

2. Abstract Meaning Representation

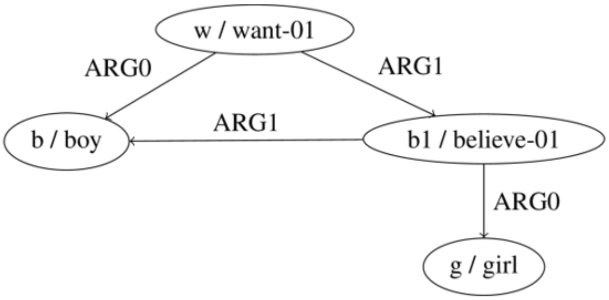
AMR is a semantic representation language designed to represent or encode the logical meaning of a sentence, abstracting away from elements of the surface syntactic structure, such as morphosyntactic information and word ordering (Banarescu et al., 2013). In a propositional-style logic, AMR is able to capture who is doing what to whom in a sentence. In such formalism, words that do not significantly contribute to the meaning of a sentence are left out of the annotation.

The AMR annotation is more frequently represented as a single-rooted directed acyclic graph with labeled nodes (concepts) and edges (relations) among them (see Figure 1). Nodes represent the main events and entities that occur in a sentence, and edges represent semantic relationships among nodes. AMR concepts are either (i) words in their lexicalized forms (e.g., boy), (ii) predicate-argument structure as defined by the PropBank resource (Palmer et al., 2005) (e.g., want-01), or (iii) special keywords such as "date-entity", "government-organization", and others. In the example of Figure 1, the concepts are `want-01`, `believe-01`, `boy` and `girl`, and the relations are `:ARG0` and `:ARG1`, represented by labeled directed edges in the graph. The symbols `w`, `b`, `b1` and `g` are variables and may be re-used in the annotation, corresponding to reentrances (multiple incoming edges) in the graph. Overall, AMR has become popular in the NLP research community due to its attempt to abstract away from syntactic idiosyncrasies and its wide use of other comprehensive linguistic resources, such as PropBank[10] (Palmer et al., 2005), supposedly being relatively simpler than other semantic languages.

Concerning its attempt to abstract away from syntactic idiosyncrasies, it may be seen that the AMR annotation examples in Figure 1 could be generated from the sentences "The boy wants the girl to believe him" and "The boy wants to be believed by the girl", which are semantically similar, but with different syntactic realizations.

[10] The PropBank project created a corpus of text annotated with information about basic semantic propositions. More information at https://propbank.github.io/.

Figure 1 — AMR graph for the sentence "The boy wants the girl to believe him"



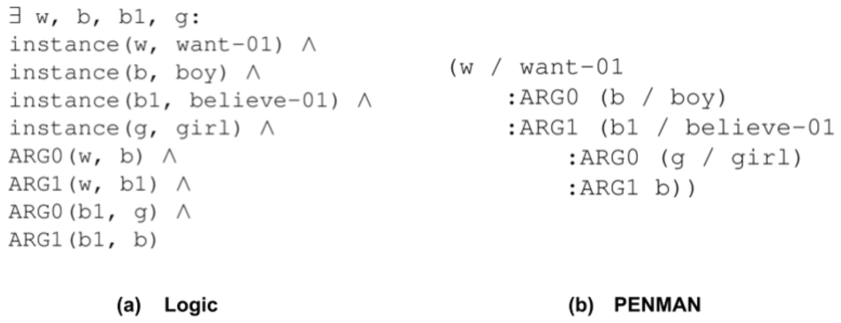
In relation to the use of linguistic resources, Figure 2 shows a predicate-argument structure (or *frameset*) provided by PropBank, which is essentially a verb linked to a list of possible arguments and their semantic roles. In this case, the frameset `want.01` represents the "desire to possess or do (something)" sense. It has two arguments, `Arg0` and `Arg1`, with the semantic roles `wanter` and `thing wanted`.

Figure 2 — Example of PropBank frameset

Frameset **want.01** "possession desiring"
Arg0: wanter
Arg1: thing wanted
Ex: [$_{Arg0}$ I] want [$_{Arg1}$ a flight from Ontario to Chicago].

Furthermore, AMR also offers approximately 100 additional relations, which are used to annotate different types of information, such as quantities (for example,`:quant`, `:unit`, `:scale`), dates (`:day`, `:month`, `:year`,`:weekday`), and others (`:mod`, `:manner`, `:location`, `:name`, `:polarity`). AMR may also be represented in first-order logic (see (a) in Figure 3) or in the PENMAN notation (Matthiessen & Bateman, 1991) (see (b) in Figure 3), for easier human reading and writing.

Figure 3 — Other AMR notations



```
∃ w, b, b1, g:
instance(w, want-01) ∧
instance(b, boy) ∧
instance(b1, believe-01) ∧
instance(g, girl) ∧
ARG0(w, b) ∧
ARG1(w, b1) ∧
ARG0(b1, g) ∧
ARG1(b1, b)
```

(a) Logic

```
(w / want-01
    :ARG0 (b / boy)
    :ARG1 (b1 / believe-01
        :ARG0 (g / girl)
        :ARG1 b))
```

(b) PENMAN

3. Related Work

Although AMR was not initially planned to be an interlingual semantic representation (Banarescu et al., 2013), some efforts in this line have been made to build non-English corpora. Nowadays, there are aligned and parallel AMR corpora available in Czech, Chinese, Spanish, and BP (Anchiêta & Pardo, 2018a; Damonte & Cohen, 2018; Xue et al., 2014), built mainly in a semiautomatic way.

Xue et al. (2014) is probably the first work that addressed the construction of AMR annotated corpora for non-English languages. Aiming to evaluate AMR's potential to work as an interlingua, the authors annotated 100 English sentences from the Penn TreeBank (Marcus et al., 1993) with AMR. Such sentences were translated to Czech and Chinese, and annotated with AMR as well. As a result, Xue et al. (2014) observed that the level of compatibility of AMR between English and Chinese is higher than between English and Czech.

Since annotating AMR manually is time consuming and demands a team of experts to perform reliable annotation, some efforts were made to develop AMR parsing and converting tools from other semantic representations. Vanderwende et al. (2015) proposed an AMR parser to convert logic form representations into AMR for English. As a result, AMR-annotated corpora for French, German, Spanish, and Japanese have been released. Damonte and Cohen (2018) also developed an AMR parser for English. In their work, they used parallel corpora to learn AMR parsers for Italian, Spanish, German, and Chinese, and discovered that the tools were able to overcome structural differences between the languages. Another result of this work is the method proposed by the authors to evaluate the parsers, which exempt the need of gold standard data for the target languages.

Despite their usefulness, automatic alignment and conversion strategies do not necessarily reflect the complexity of some linguistic phenomena in non-English languages, so manual annotation or revision is necessary. Thus, other annotating teams tried to adapt the AMR guidelines to their languages (Sobrevilla Cabezudo & Pardo, 2019; Linh & Nguyen, 2019; Migueles-Abraira et al., 2018). Migueles-Abraira et al. (2018) performed a manual AMR annotation of the book "The Little Prince", refining the original AMR guidelines and comparing Spanish and English in terms of similarity of the occurring phenomena. As a result of this work, the authors identified some relevant specific phenomena that proved to be challenging during the annotation process, such as ellipsis, third person possessives and clitic pronouns.

Concerning Portuguese, two AMR-annotated corpora were created. Both corpora used Verbo-Brasil[11] (Duran & Aluísio, 2015) as a lexical resource to annotate the framesets, that is based on the same representation scheme of the PropBank lexical repository. The first one was automatically built, leveraging the alignments between sentences of the "The Little Prince" book in English and Portuguese (Anchiêta & Pardo, 2018a). Specifically, such corpus is the result of an aligner based on pre-trained word embeddings and Word Mover's Distance function (Kusner et al., 2015) to match word tokens in the sentences and nodes in the corresponding AMR graphs. The Little Prince corpus has a rather unusual genre (tales), and is composed of sentences with restricted vocabulary, mainly related to the story. Furthermore, the number of sentences is small: only 1,527 annotated sentences. The other corpus is the AMRNews corpus, whose first version is described by Sobrevilla Cabezudo & Pardo (2019). The next section presents the current version of the corpus and reports its annotation methodology and evaluation, and also introduces a novel initiative on annotating opinions into the AMR formalism. Both corpora compound the AMR-PT corpus initiative.

[11] Available at http://143.107.183.175:21380/verbobrasil/.

4. The AMR-PT Initiative

General Description

The AMR-PT initiative comprises two corpora in different domains. One focuses on news texts, named AMRNews-PT, and the other one on opinionated texts, named OpiSums-PT-AMR. AMRNews is a news corpus with manually annotated sentences following the English AMR guidelines with some language-specific adaptations (Sobrevilla Cabezudo & Pardo, 2019). The journalistic texts were extracted from the Folha de São Paulo news agency[12]. The selected data came from different domains such as "Daily news", "World news", "Education", "Environment", "Sports", "Science", "Balance and Health", "Ilustrada", "Ilustríssima", "Power", "Tourism", "Food", and "Technology". To enrich the corpus, news sentences were also extracted from PropBank.Br[13] (Duran & Aluísio, 2012), since it already contains semantic role information, which makes the AMR annotation much easier. The document's download period was November 25th-28th, 2018. Currently, this corpus contains 870 manually annotated sentences and the size for each sentence is up to 23 tokens[14].

In its turn, the OpiSums-PT-AMR corpus comprises 404 manually annotated sentences from the OpiSums-PT (López Condori et al., 2015) corpus, which was created based on comments about 13 books — originally from the ReLi corpus (Freitas et al., 2014) — alongside opinions concerning four electronic products obtained from the Buscapé[15] e-commerce website. Each product has 10 comments with multiple sentences each. Every document also does not exceed 300 tokens.

[12] Available at https://www.folha.uol.com.br/.

[13] PropBank.Br was the basis for the construction of the previously cited Verbo-Brasil repository.

[14] Due to the difficulty that the task of semantic annotation carries, the scope of this corpus was focused on annotating only short sentences (but guaranteeing that different domains are covered). In order to define what a short sentence is, the average number of tokens by sentence was calculated and this value was used as a threshold. Thus, sentences with a number of tokens below the average (in our case, it was 23 tokens) were selected.

[15] Available at https://www.buscape.com.br.

Annotation Procedure

Both annotations of journalistic and opinionated sentences followed the same process (Sobrevilla Cabezudo & Pardo, 2019). This means that it was guided by the original AMR guidelines[16] including the adaptations performed by the authors, and the lexical repository used was Verbo-Brasil (Duran & Aluísio, 2015).

The initial annotation focused on journalistic sentences (Sobrevilla Cabezudo & Pardo, 2019) and the team was originally composed of 14 annotators that belong to the areas of Computer Science and Linguistics, and with large experience in NLP. These annotators took part in two training sessions. In the first session, the task and the resources to be used were presented. The participants were trained by annotating sentences of PropBank.Br (Duran & Aluísio, 2012) for perceiving the difficulty of the task. The second session aimed at answering questions about the annotation, showing the inter-annotator agreement in the training stage, some common mistakes, and launching the annotation process. This process resulted in 299 annotated sentences.

In general, the annotation procedure consisted of two general steps. The first step aimed at analyzing the sentence structure, while the second step counted with the aid of the AMR Editor tool (Hermjakob, 2013), which produces an AMR PENMAN format to be exported into textual files of easy processing and consulting. Furthermore, the fundamentals of the manual annotation process in the first step were the following:

a) Identification of sentence type (i.e., default, comparative, superlative, coordinate, subordinate, and others), which determines whether it is necessary to build two or more sub-graphs (in case of coordinate or subordinate sentences) and then to join them using a conjunction (usually coordinate sentences) or a concept of the main sub-graph (in the case of subordinate sentences).

b) Concept identification, which was based on the AMR guidelines. Specifically, the annotators identify general concepts, either from the AMR guidelines or from Verbo-Brasil.

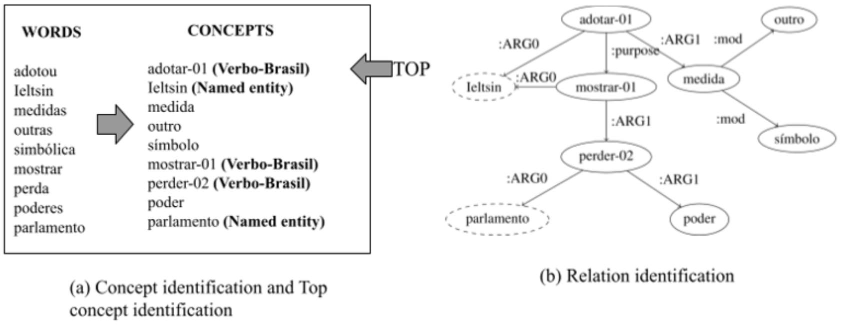
[16] Available at https://github.com/amrisi/amr-guidelines/blob/master/amr.md. The adopted version was the 1.2.6.

c) Identification of the main concept, which is done based on the two previous steps. To illustrate, the main verb could be the main concept in a default sentence.

d) Identification of relations among the identified concepts[17].

This sequence of actions (a-d) can be illustrated with Figure 4. To annotate the sentence "*Ieltsin adotou outras medidas simbólicas para mostrar a perda de poderes do Parlamento*" ("Yeltsin took other symbolic measures to show the loss of Parliament's power."), the annotators firstly identify that it includes a subordinate clause, which means that its correspondent AMR graph should have sub-graphs. Then, they identify the concepts. In Figure 4 (a), we see that some words became general concepts (i.e., `medida`, `outro`, `símbolo` and `poder`), named-entities (`Ieltsin` and `Parlamento`) or Verbo-Brasil framesets (`adotar-01`, `mostrar-01` and `perder-02`).

Figure 4 — Example of the manual annotation procedure (Sobrevilla Cabezudo & Pardo, 2019)



(a) Concept identification and Top concept identification

(b) Relation identification

Following the concepts identification, it was necessary to identify the graph root: in this case, the verb `adotar-01`, because it is the main verb of the main clause "*Ieltsin adotou outras medidas simbólicas*" ("Yeltsin took other symbolic measures"). Finally, the relations among all concepts were identified (e.g., `:ARG0` between `Ieltsin`

[17] The relations were extracted from Verbo-Brasil (for core relations) and AMR guidelines (for non-core relations).

and `adotar-01`), which encodes that `Ieltsin` has the `adopter` semantic role according to the corresponding frameset.

After the initial annotation and with some learned lessons, we focused on annotating both journalistic and opinionated sentences. This process was performed by three human experts with previous background on AMR and its guidelines and consists in (1) annotating a set of sentences and (2) discussing the hard cases and other interesting aspects of the annotation in an iterative way. So far, the annotation process resulted in 404 opinionated and 870 journalistic sentences (resulting from the previously annotated news corpus of Sobrevilla Cabezudo & Pardo (2019) and the newly 571 annotated sentences).

It is worth noting that, before the annotation of opinionated sentences, all texts were normalized using the Enelvo[18] tool (Bertaglia & Nunes, 2016), given that guidelines state that misspellings should be normalized during annotation. If the annotators think that there is a normalization error, they could check the original text and mention this in a note.

Finally, and in a similar way to Sobrevilla Cabezudo & Pardo (2019), each sentence with a verb that was not present in the Verbo-Brasil repository was not annotated and the given verb was added in a list to enable further development of the resource (in future work).

Evaluation

To compute the inter-annotator agreement, a random subset of all sentences to be annotated was shared among all the annotators and they could not discuss this subset. The agreement measure used was Smatch[19] (Cai & Knight, 2013). Unlike the work of Banarescu et al. (2013), which built a gold standard (using the total agreement

[18] Available at https://github.com/tfcbertaglia/enelvo.

[19] Available at https://github.com/snowblink14/smatch. It is interesting to notice that, differently from annotation efforts for other linguistic phenomena, the Smatch metric is the dominant metric for AMR annotation (instead of Kappa or other metrics (Banerjee et al., 1999)), following the original work on AMR (Banarescu et al., 2013). It evaluates the triples formed by the relations and the associated nodes in an AMR structure. Moreover, Smatch does an additional task of mapping the variables in the AMR representation in a way to maximize the results.

between the annotators), we calculated the inter-annotator agreement by comparing all annotations in an all-against-all configuration, obtaining the average of all inter-annotator agreements[20]. A set of 50 sentences was used for calculating the agreement within the journalistic corpus. Meanwhile, for the OpiSums-PT-AMR, 70 sentences of different lengths were initially selected to compose the agreement set, however, due to the complexity of the annotation process, only 17 were actually annotated by all three experts and, therefore, were considered to calculate the agreement for this specific corpus.

The overall agreement for the journalistic part of the corpora achieved a Smatch value of 0.73, which is a good value, considering that the inter-annotator agreement in the original AMR project ranged between 0.70 and 0.80. Besides, 34 (from the 50) sentences were annotated by 5-7 annotators in the initial procedure and the last 16 sentences were annotated by 3 annotators.

The annotation of sentences from the opinions domain resulted in an average agreement of 0.90, which can be considered high, when compared with other works on the matter. This can be due to the fact that the 17 sentences used are shorter and, therefore, easier to achieve some consensus.

From the obtained annotation, we can make a comparative analysis between the two domains within this AMR-BP initiative, pointing out how the texts differ in terms of semantic phenomena and how they are captured by the AMR representation.

News vs Opinionated texts

In total, the AMRNews corpus includes 4,192 concepts (excluding `name`) and 3,758 relations (excluding `:instance`), whilst OpiSums-PT-AMR comprises 3,064 concepts and 3,159 relations. As a first comparative analysis, we can observe the distribution of the different

[20] Finally, the annotated versions of the sentences belonging to the agreement sample that were included in the final corpus were chosen by an adjudicator (as more than one possible annotation exists).

types of phenomena captured by the concepts within the AMR graphs. These statistics can be seen in Table 1.

Table 1 — Statistics of concepts in both the AMRNews and OpiSums-PT-AMR corpora

Concepts	**Frequency**	
	AMRNews	**OpiSums-PT-AMR**
General concepts	1,977	1,770
Verbo-Brasil concepts	866	641
Named entities	311	125
Modal verbs	45	25
Amr-unknown*	80	7
Other entities and special frames	104	169
Constants**	660	215
Negative polarity	135	79

* AMR uses the concept "amr-unknown" to indicate wh-questions.
** Constants include numbers, strings and symbols that are not traditional concepts and, therefore, are not given variable names.

In a more detailed analysis, we present in Table 2 the 15 most frequent relations in each corpus. It is possible to see that the 3 most frequent relations are the same (and in the same order) in the two corpora. One point to remark in relation to the table is that, in the news texts, the sentences and expressions contained in them describe facts and usually use numbers to report quantities (through the `:quant` relation). More than this, some expressions collected until now describe imperatives like "*Arranje*!" ("Get it!"). Thus, the imperative mode (`:mode` relation) is frequent in the corpus. It is expected that, when the news corpus grows, these relations will change a bit.

Table 2 — Fifteen most frequent relations in both the AMRNews and OpiSums-PT-AMR corpora

OpiSums-PT-AMR			**AMRNews**		
Relation	**Frequency**	**Freq. (%)**	**Relation**	**Frequency**	**Freq. (%)**
ARG1	652	20.64%	ARG1	715	19.03%
op	624	19.75%	op	706	18.79%

ARG0	485	15.35%	ARG0	512	13.62%
mod	314	9.94%	name	311	8.28%
ARG2	208	6.58%	mod	268	7.13%
name	125	3.96%	ARG2	196	5.22%
domain	96	3.04%	polarity	169	4.50%
polarity	80	2.53%	domain	143	3.81%
time	67	2.12%	quant	105	2.79%
topic	56	1.77%	time	98	2.61%
poss	56	1.77%	location	75	2.00%
snt	55	1.74%	manner	49	1.30%
quant	44	1.39%	poss	48	1.28%
degree	39	1.23%	topic	45	1.20%
ARG3	38	1.20%	mode	36	0.96%

We can also note, from both Table 1 and Table 2 (through the `:name` relation), that news texts contain a higher proportion of named entities. However this phenomenon is still common in opinions, as `:name` is the 6th most common relation in OpiSums-PT-AMR. It is also worth pointing out that the `:degree` relation, used mainly with amplifiers and downtoners, are more common in opinionated texts, especially when taking into account its associated concept (`have-degree-91`), as can be seen in Table 3, which includes the ten most frequent framesets for both corpora.

Table 3 — Ten most frequent framesets in both the AMRNews and OpiSums-PT-AMR corpora

OpiSums-PT-AMR			**AMRNews**		
Frameset	**Frequency**	**Freq. (%)**	**Frameset**	**Frequency**	**Freq. (%)**
cause-01	44	5.27	ter-01	42	4.14
ler-01	42	5.03	contrast-01	29	2.86
ter-01	35	4.19	possible-01	27	2.66
gostar-01	33	3.95	dizer-01	24	2.36
contrast-01	27	3.23	fazer-01	23	2.27
escrever-01	25	2.99	haver-01	17	1.67
have-rel-role-91	21	2.51	querer-01	17	1.67
have-degree-91	21	2.51	acontecer-01	15	1.48
possible-01	17	2.04	saber-01	13	1.28
fazer-01	15	1.80	cause-01	13	1.28

Analyzing the results in Table 3[21], it is also important to mention that the higher frequencies of some concepts — such as `ler-01` (to read), `escrever-01` (to write) and `have-rel-role-91` (used to indicate personal relationship between people) — are due to the type of products about which the opinions are written, mainly books. A noteworthy observation to be made is that opinions have framesets used within contexts with some degree of sentiment associated, e.g., `gostar-01` (to like) and `have-degree-91`. Meanwhile, news texts have more descriptive concepts, such as `ter-01` (to have), `dizer-01` (to say), `fazer-01` (to do/to make) and `acontecer-01` (to happen), among others.

One of the limitations of annotating AMR in BP is related to Verbo-Brasil, as the lexical units that are not represented in this resource could not be annotated. We found 161 verbs that were not present in Verbo-Brasil, as *opinar*, *duvidar*, *ficar* (in the sense of "dating") and *devorar* (in the sense of "outperforming"). Overall, almost 14% of the analyzed sentences had verbs not found in Verbo-Brasil. These sentences were discarded and, therefore, not included in our corpora. Thus, this work also shows directions for developing and improving the linguistic resources used for building the annotated corpora, as well as adapting original methods and guidelines, which remain for future work.

5. Linguistic Phenomena and Adapted Guidelines

The experience of annotating news and opinionated corpora in BP with AMR allowed the identification of some challenging phenomena that could not (totally or partially) be represented with AMR. Thus, we conducted a linguistic analysis of them that could offer possible solutions to other language annotation teams that face similar issues. Although we are not able to propose definitive solutions to all the problems, we believe that they are possible satisfactory strategies.

[21] Some framesets in the table come directly from the English PropBank and not from Verbo-Brasil due to the original guidelines developed by Sobrevilla Cabezudo & Pardo (2019), in which modal verbs (`possible-01`) and some conjunctions (`contrast-01`, `cause-01`) are annotated in such way to keep consistency with the original AMR guidelines (Banarescu et al., 2013). Some other framesets are AMR-exclusive, for instance, `have-rel-role-01` and `have-degree-91`, and were kept in English.

The hard cases discussed here are diminutives, null subject, pronoun ambiguity, and multiword expressions. We do not aim, however, to present an extensive or exhaustive analysis for each example and issue in the corpus.

Diminutives

From the 404 sentences of OpiSums-PT-AMR, there are five sentences with one diminutive case each (1-5). Such diminutives are basically formed by replacing the unstressed final vowel -o or -a of a word with the affix *-inho* or *-inha* according to its gender. There are other rules of diminutive formation[22] in BP, but there are no occurrences of them in the corpus.

1. Aquele filme meio [*chatinho*]$_{\text{adj}}$ e clichê que está passando na televisão. (“That rather boring and cliché film that is on television”)

2. Livro bem [*chatinho*]$_{\text{adj}}$ (“[A] pretty boring book”)

3. Muito [*engraçadinho*]$_{\text{adj}}$! ;) (“Very funny! ;)”)

4. Lindo, [*fininho*]$_{\text{adj}}$ e discreto. (“[It´s] Beautiful, very thin and discrete”)

5. Acaba se atrapalhando com a sua “[*anjinha*]$_{\text{noun}}$”. (“He/She ends up messing with his/her little angel’”)

In the examples (1) and (2), the diminutive form *chatinho* is used to temper an unpleasant quality. In (3) and (4), however, the meaning is quite different from (1) and (2); they have the meaning of “nice and…” or having a quality to exactly the desirable degree (i.e., *engraçadinho* > “good and funny”, and *fininho* > “good and thin”). As illustrated by sentence (5), diminutive forms very often connote cuteness, affection or pleasantness (more examples are: “*Que tal uma cervejinha gelada?*”

[22] Diminutive in BP can also be formed as follows: (i) with nouns and adjectives ending in *-s* or *-z*, the affix *-inho/-inha* is also added to the stem word (e.g., *japonês* (“Japanese man”) > *japonesinho* (“little Japanese guy”), and *voz* (“voice”) (fem.) > *vozinha* (“little voice”), and (ii) with all other nouns, the affix -zinho/-zinha is added to the word (e.g., *papel* (“paper”) (masc.) > *papelzinho* (“scrap of paper”), and *mão* (“hand”) (fem.) > *mãozinha* (“little hand”).

/ "What about a nice and cold beer?" or "*Adoro pezinho de bebê*" ("I love babies' little feet")[23].

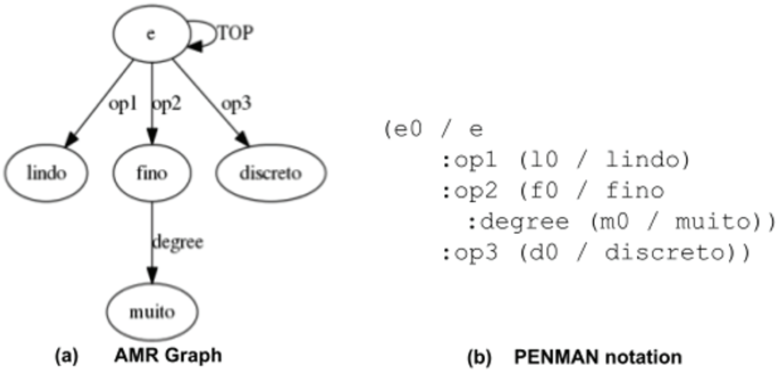
According to Alves (2006), diminutive forms can be classified in terms of their function in semantic diminutives and pragmatic diminutives. The first group expresses "reduced size /quantity / intensity" meanings, which are based on inherent properties or features of the objects. The second one expresses more subjective meanings and refers to how the speaker perceives objects and their properties, which are guided by social and cultural factors. Thus, we first classified the cases in semantic diminutives (4) and pragmatic diminutives (1, 2, 3 and 5) for understanding the different meanings of such words before the AMR annotation[24]. This task was strongly based on world knowledge, since the sentences are out of context, as it is established by the AMR guidelines.

We then turn to the AMR guidelines for diminutive annotation. While the semantic diminutive *fininho* in (4), for example, is easily represented in AMR with the `:degree` relation (as in Figure 5), which links two concepts, i.e., `fino` ("small") and `muito` ("very"), the pragmatic diminutive is much more difficult to represent, since it corresponds to non-literal meanings. In other words, the concepts represented by pragmatic diminutives do not literally mean a `:degree` relation, so using the same annotation in both constructions seems inappropriate. Consequently, we used two different annotation schemes for diminutives: while the semantic ones are represented as usual with the `:degree` relation (Figure 5), the pragmatic diminutives are not lemmatized, and the concept remains as a diminutive, as in Figure 6.

[23] It's worth noting that the same can happen with the augmentative. In our corpora, however, there was no occurrence of augmentative forms.

[24] The different meanings of diminutive are not an idiosyncrasy of Portuguese; however, we are not aware of specific AMR guidelines in the literature to annotate these cases.

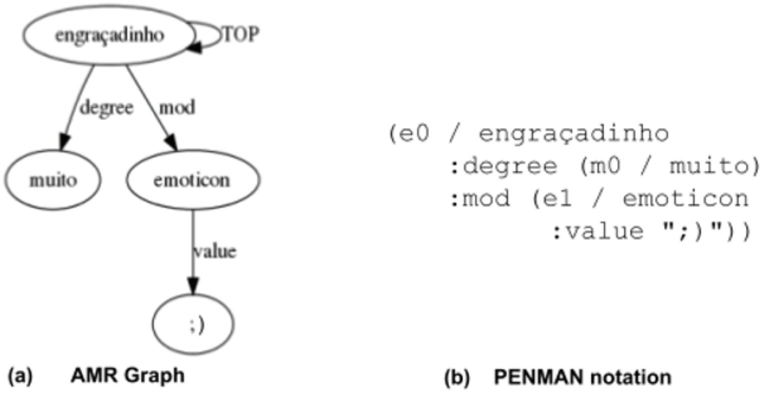
Figure 5 — Annotation of sentence 4 with a case of semantic diminutive



```
(e0 / e
    :op1 (l0 / lindo)
    :op2 (f0 / fino
      :degree (m0 / muito))
    :op3 (d0 / discreto))
```

(a) AMR Graph (b) PENMAN notation

Figure 6 — Annotation of sentence 3 with a case of pragmatic diminutive



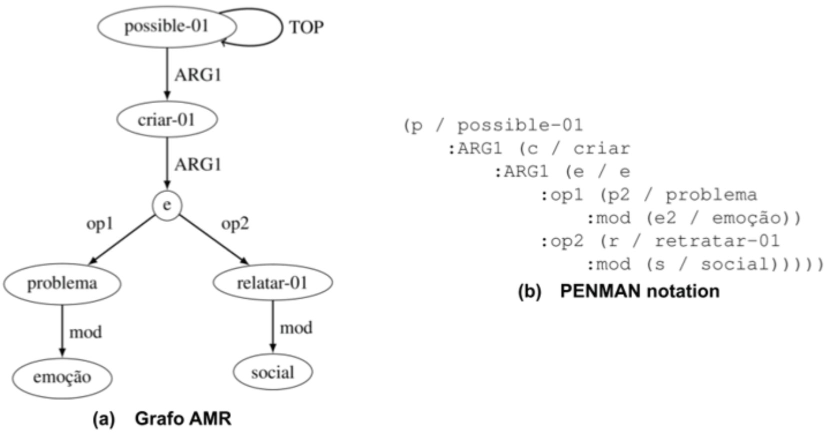
```
(e0 / engraçadinho
    :degree (m0 / muito)
    :mod (e1 / emoticon
            :value ";)"))
```

(a) AMR Graph (b) PENMAN notation

Null subject and pronoun ambiguity

In BP, as in other romance languages (e.g., Spanish), but different from English, the subject does not have to be necessarily expressed in the sentence. In the example shown in Figure 7, the subject is not present in the sentence ("*Não precisaria agir assim.*") ("[He/She/You] wouldn't have to act like this"), but is probably clear in the sentence source-text. However, the verb ("*precisaria*") indicates that the person referred to is a third person in singular, since the verb has this conjugation. In this situation, the annotation team decided to annotate the ARG0 role explicitly, even if it is not in the sentence. The reason for such a definition is that it permits the explicit identification of the ARG0 role. Thus, it would be possible to recover the agreement information. However, this decision led to another problem: the third-person pronoun ambiguity. In BP, a verb conjugated in the third person can refer to the second person in singular (“you”) or to the third person in singular (“he” or “she”), so it had also to be decided if the pronoun annotated would be the second or the third person pronoun, and, in the later case, if it is masculine or feminine. Thus, the decision was to annotate it as the third person masculine (he). The same decision was kept for the ambiguity of possessive pronouns, when *seu*/*sua* can refer both to yours and his/her.

Figure 7 — Null subject annotation



(a) **Grafo AMR**

```
(p / possible-01
    :ARG1 (c / criar
        :ARG1 (e / e
            :op1 (p2 / problema
                :mod (e2 / emoção))
            :op2 (r / retratar-01
                :mod (s / social)))))
```

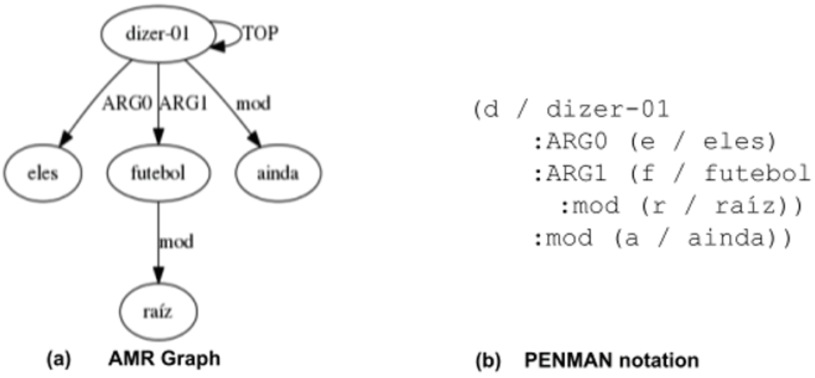
(b) **PENMAN notation**

This decision was based on the argument that, as the initial annotation focused on a journalistic corpus, it was expected to be more frequent that the null subject refers to a person about whom something is being reported. Besides, the decision for the masculine is based on the original lemmatization rules for concepts in Portuguese, that orient to lemmatize the modifiers in their masculine singular form.

Another problem arises when the verb is in plural form and the subject is indeterminable, as in “*Dirão até que é futebol raiz*” (“[They/Someone] will also say it is the old soccer”) (Figure 8). In this example, the annotators were oriented to explicitly annotate the ARG0 as “they”, but mention the fact that the sentence contains an indeterminable subject. In this case, the standard orientation of always using singular was changed in favor of the possibility to represent this phenomenon.

Lastly, there are some cases where it is not possible to identify if the pronoun is a personal one or a demonstrative one, as in “*Pode até criar problema emocional e retração social*” (“[She/He/It/You] can even create emotional problem and social withdrawal”) (Figure 9). The subject could be a person, a fact or the entire previous sentence that is being taken up, and it is impossible to recover this information without context (note that AMR considers only the sentence level for the annotation). In this case, the annotators were oriented to not explicitly annotate any pronoun.

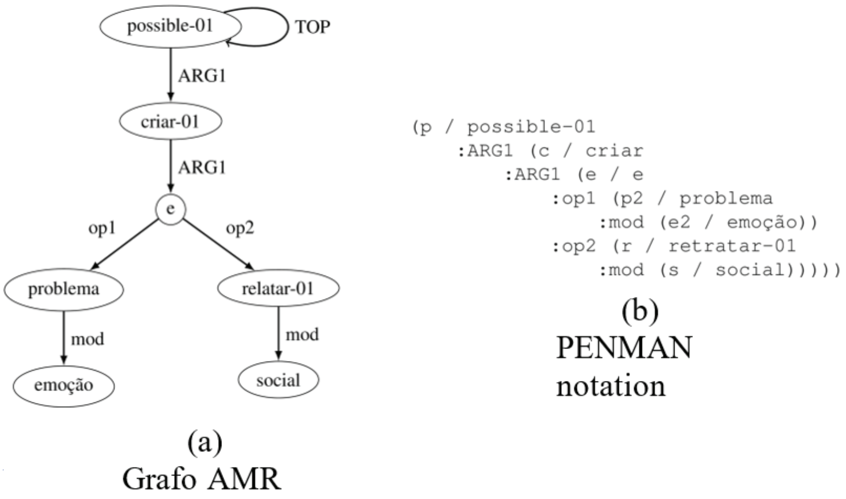
Figure 8 — Indeterminable subject explicitly annotated.



```
(d / dizer-01
    :ARG0 (e / eles)
    :ARG1 (f / futebol
      :mod (r / raíz))
    :mod (a / ainda))
```

(a) AMR Graph (b) PENMAN notation

Figure 9 — Indeterminable subject not explicitly annotated



```
(p / possible-01
    :ARG1 (c / criar
        :ARG1 (e / e
            :op1 (p2 / problema
                :mod (e2 / emoção))
            :op2 (r / retratar-01
                :mod (s / social)))))
```

(b)
PENMAN
notation

(a)
Grafo AMR

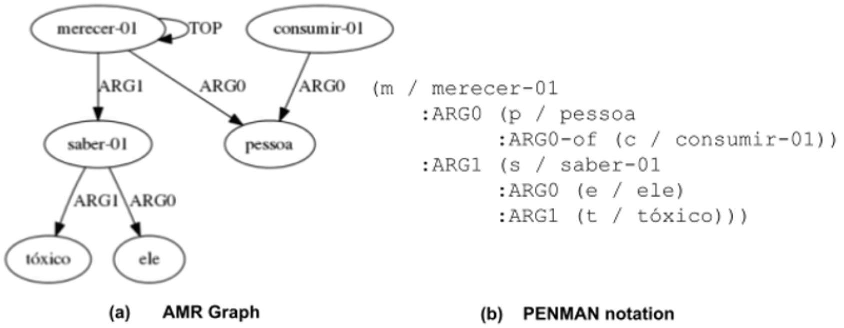
A specially challenging phenomenon was the annotation of multiword expressions (MWE). MWEs are (continuous or discontinuous) sequences of words with some degree of orthographic, morphological, syntactic or semantic idiosyncrasy with respect to what is considered general grammar rules of a language (Baldwin & Kim, 2010). Another important property of MWEs is the semantic non-compositionality, i.e., it is impossible to deduce the meaning of the whole unit based only on the meaning of its parts (Constant et al., 2017).

The original AMR guidelines define that MWEs should be represented as a unique concept that is synonym or equivalent to the MWE. One example that occurs frequently are the light-verb constructions (LVCs), such as “The girl made adjustments [adjust] to the machine”. LVCs are composed by a verb that does not add much semantics to the expression (the light verb) (Wittenberg et al., 2014), followed by a predicative noun that represents a state or an event. While many cases have indeed a unique verbal form that can replace the MWE (“make adjustments” has the equivalent full verb “adjust”), in some cases this is not possible.

In Figure 10, there is no full verb that could directly substitute the MWE “*ter direito*”, so in cases like that the team decided to find synonyms (in the example, the full verb “*merecer*” means “to deserve”). In other examples, such as in the idiom “*pagar mico*” (literally, “to pay the monkey”), that means “to completely embarrass yourself”, the best synonym in BP is also a MWE: “*passar vergonha*” (“to get embarrassed”).

The solution in these cases was finding a concept in the Verbo-Brasil repository that could represent this structure with core arguments, resulting in annotating light-verbs as full verbs and ignoring the fact that the element predicating the sentence is actually the predicative noun. This case demanded a lot of discussion every time the team faced a new MWE, because some synonyms do not express the same meaning as the composed construction. This issue arises also because, different from PropBank, Verbo-Brasil has very few MWEs and no predicative nouns as framesets. Increasing the number and the diversity of the repository would probably solve most of these problems, but this would cost time and another team of experts to improve the lexical resource before continuing the AMR annotations.

Figure 10 — MWE annotated by synonym

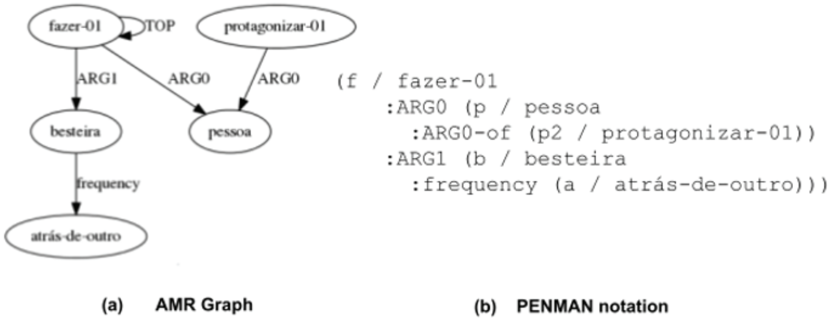


```
(m / merecer-01
    :ARG0 (p / pessoa
          :ARG0-of (c / consumir-01))
    :ARG1 (s / saber-01
          :ARG0 (e / ele)
          :ARG1 (t / tóxico)))
```

(a) AMR Graph (b) PENMAN notation

Furthermore, some MWEs are composed of non-lexical words and have adverbial meanings, as "*atrás de*" in the sentence "*A protagonista faz uma besteira atrás da outra*" ("The protagonist makes a mess after the other") (cf. Figure 11).

Figure 11 — MWE annotated with hyphens.



```
(f / fazer-01
    :ARG0 (p / pessoa
      :ARG0-of (p2 / protagonizar-01))
    :ARG1 (b / besteira
      :frequency (a / atrás-de-outro)))
```

(a) AMR Graph (b) PENMAN notation

The prepositional compound *atrás de* typically has the locative meaning of "behind", but in the MWE "*um* [noun] *atrás de outro*" (e.g., "*uma besteira atrás da outra*") it acquires the sense of a temporal sequence of things (as represented by "after", in English). In this case, the team annotated the expression as it occurs, using hyphens, as implied by some examples presented at the original AMR dictionary guidelines for the English language (e.g., "He can recite the poem by heart." is annotated with a concept `by-heart`, to indicate the manner in which the poem is recited). This leads to many concepts that have to be represented in this way, which is not a good long-term solution, since it makes room for annotators using hyphens whenever a compound arises.

The best way to deal with this phenomenon continues to be an open question, and it would be useful to analyze the frequency of MWEs in the corpus for proposing other (better) solutions for annotating them. One option would be to improve Verbo-Brasil and adding not only multiword framesets, but also the predicative nouns and their argument structure, so they could be used for the annotation as it has been made for English. This challenge highlights the importance of robust lexical resources that are not always available for under-represented languages. This could be one of the most important constraints in using AMR as an interlingua. Another possible option (that was not taken by now in the BP team) is using other lexical repositories that are specific for MWEs, such as the Parseme corpus for verbal MWEs, that is available for Portuguese, as for many other languages (Ramisch et al., 2018). This could be used at least as a consulting repository to identify when an expression is a real MWE (since this identification is not trivial), and for future improvement of Verbo-Brasil.

6. Final Remarks

This work presented and detailed two AMR annotated corpora for Brazilian Portuguese — the AMRNews and the OpiSums-PT-AMR corpora — and carried out a comparative analysis between opinions and news, highlighting important differences on the occurrence of semantic phenomena between each type of text. The released version of the AMR Corpus for Brazilian Portuguese is available at the web portal of the POeTiSA project[25]. Although the amount of AMR annotated data for Portuguese is still small (due to the hard task that AMR annotation represents), it has already subsidized NLP initiatives for the Portuguese language, as semantic parsing (Anchiêta & Pardo,

[25] https://sites.google.com/icmc.usp.br/poetisa

2018b, 2022), text generation (Sobrevilla Cabezudo & Pardo, 2022), and opinion summarization (Inácio & Pardo, 2021).

We also explored the language-specific challenges that appeared during the AMR annotation process and some strategies to deal with these. As it could be seen, some of them may be better handled (diminutives). However, there are other phenomena which are hard to deal with and a deeper study has to be conducted. On the other hand, projects aiming to build unified multilingual sembanks for NLP tasks have to follow a minimum pattern by annotating similar phenomena to allow comparing them in terms of frequency and structure among different languages. In this way, annotation adaptations should be restricted to specific phenomena (and as general as possible to capture similar phenomena in similar languages), so the core idea of the AMR scheme rests true for as many languages as possible.

There are also phenomena that may lead to further research of the AMR semantic representation, such as metaphorical language, which is situated in a boundary interface between semantics and pragmatics, according to Legroski (2009). This type of phenomenon also has a degree of relation to multiword expressions, which, as we discuss throughout this paper, present a challenge for annotation.

Gender is also an interesting path for research. Migueles-Abraira (2017) includes grammatical gender annotation for their Spanish version of AMR, under the argument that it has influence on the understanding of a sentence. For example, the word "caixa" may represent two different concepts in Portuguese: box or clerk, depending on its gender (feminine or masculine, respectively). This decision, however, needs to be further discussed taking into account the lemmatization process of words into concepts (since, in Portuguese, the lemmas are commonly represented by the words' masculine forms) and which other morphological aspects (e.g., number) should be included in a semantic representation as the AMR.

Acknowledgments

The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and from the IBM Corporation.

This research also had the support of Coordination for the Improvement of Higher Education Personnel (CAPES) and the OPINANDO project[26] (PRP #668).

Conflict of interests (multiple authors)

The authors declare they have no conflict of interest.

Credit Author Statement

We, Marcio Lima Inácio, Marco Antonio Sobrevilla Cabezudo, Renata Ramisch, Ariani Di Felippo and Thiago Alexandre Salgueiro Pardo, hereby declare that we do not have any potential conflict of interest in this study. The first two authors have carried out the collection of the texts to be annotated and, alongside Renata Ramisch, performed the annotation of the copus in AMR. Marcio Inácio and Marco Cabezudo have also been responsible for the contrastive data analysis. All authors contributed to the qualitative discussion about the phenomena observed in the data and the writing of this work. Ariani Di Felippo and Thiago Pardo were responsible for the project supervision. All authors approve the final version of the manuscript and are responsible for all aspects, including the guarantee of its veracity and integrity.

References

Abend, O., & Rappoport, A. (2013). UCCA: A semantics-based grammatical annotation scheme. *Proceedings of the 10th International Conference on Computational Semantics – Long Papers*, 1–12. https://aclanthology.org/W13-0101.pdf (accessed 23 August, 2022).

Abzianidze, L., Bjerva, J., Evang, K., Haagsma, H., van Noord, R., Ludmann, P., Nguyen, D.-D., & Bos, J. (2017). The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 242–247. https://aclanthology.org/E17-2039.pdf (accessed 23 August, 2022).

Alves, E. (2006). O diminutivo no português do Brasil: funcionalidade e tipologia. *Estudos Linguísticos*, *35*, 694–701. http://www.gel.hospedagemdesites.ws/estudoslinguisticos/

[26] https://sites.google.com/icmc.usp.br/opinando/

edicoesanteriores/4publica-estudos-2006/sistema06/885.pdf (accessed 23 August, 2022).

Anchiêta, R. T., & Pardo, T. A. S. (2018a). Towards AMR-BR: A sembank for Brazilian Portuguese language. *Proceedings of the eleventh international conference on language resources and evaluation*, 974–979. https://aclanthology.org/L18-1157.pdf (accessed 23 August, 2022).

Anchiêta, R. T., & Pardo, T. A. S. (2018b). A rule-based AMR parser for Portuguese. *Proceedings of the 16th Ibero-American Conference on Artificial Intelligence*, 341–353. https://doi.org/10.1007/978-3-030-03928-8_28.

Anchiêta, R. T., & Pardo, T. A. S. (2022). Abstract meaning representation parsing for the Brazilian Portuguese language. *Proceedings of the International Conference on Computational Processing of Portuguese*, 429–434. https://doi.org/10.11606/T.55.2020.tde-29072020-120805.

Baldwin, T., & Kim, S. N. (2010). Multiword expressions. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of Natural Language Processing*, 2nd ed., (pp. 267–292). CRC Press.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., & Schneider, N. (2013). Abstract meaning representation for sembanking. *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 178–186. https://aclanthology.org/W13-2322.pdf (accessed 23 August, 2022).

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics*, *27*(1), 3–23. https://doi.org/10.2307/3315487

Basile, V., Bos, J., Evang, K., & Venhuizen, N. (2012). Developing a large semantically annotated corpus. *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 3196–3200. http://www.lrec-conf.org/proceedings/lrec2012/pdf/534_Paper.pdf

Bertaglia, T. F. C., & Nunes, M. das G. V. (2016). Exploring word embeddings for unsupervised textual user-generated content normalization. *Proceedings of the 2nd Workshop on Noisy User-generated Text*, 112–120. https://aclanthology.org/W16-3916.pdf (accessed 23 August, 2022).

Cai, S., & Knight, K. (2013). Smatch: An evaluation metric for semantic feature structures. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 748–752. https://aclanthology.org/P13-2131.pdf (accessed 23 August, 2022).

Cambria, E., Poria, S., Bisio, F., Bajpai, R., & Chaturvedi, I. (2015). The CLSA model: A novel framework for concept-level sentiment analysis. *Proceedings of the Computational linguistics and intelligent text processing conference*, 3–22. https://doi.org/10.1007/978-3-319-18117-2_1.

López Condori, R. E., Pardo, T. A. S., Avanço, L. V., Filho, P., Bokan, A., Cardoso, P., Dias, M., Nóbrega, F., Sobrevilla Cabezudo, M. A., Souza, J., Zacarias, A., Seno, E., & Di Felippo, A. (2015). A qualitative analysis of a corpus of opinion summaries based on aspects. *Proceedings of the 9th Linguistic Annotation Workshop*, 62–71. http://dx.doi.org/10.3115/v1/W15-1607.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, *43*(4), 837–892. https://doi.org/10.1162/COLI_a_00302.

Damonte, M., & Cohen, S. B. (2018). Cross-lingual abstract meaning representation parsing. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1146–1155. https://doi.org/10.18653/v1/N18-1104.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *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)*, 4171–4186. https://aclanthology.org/N19-1423.pdf (accessed 23 August, 2022).

Duran, M. S., & Aluísio, S. M. (2015). Automatic generation of a lexical resource to support semantic role labeling in Portuguese. *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, 216–221. https://aclanthology.org/S15-1026.pdf (accessed 23 August, 2022).

Duran, M. S., & Aluísio, S. M. (2012). Propbank-Br: A Brazilian treebank annotated with semantic role labels. *Proceedings of the Eighth International Conference on Language Resources and Evaluation* , 1862–1867. http://www.lrec-conf.org/proceedings/lrec2012/pdf/272_Paper.pdf (accessed 23 August, 2022).

Freitas, C., Motta, E., Milidiú, R. L., & César, J. (2014). Sparkling vampire... LOL! Annotating opinions in a book review corpus. *New Language Technologies and Linguistic Research*, 128–146. https://www.researchgate.net/publication/271836545_Sparkling_Vampire_lol_Annotating_Opinions_in_a_Book_Review_Corpus (accessed 23 August, 2022).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* (1st ed.). MIT Press.

Hermjakob, U. (2013). *AMR editor: A tool to build abstract meaning representations*. https://amr.isi.edu/editor.html (accessed 23 August, 2022).

Inácio, M. L., & Pardo, T. A. S. (2021). Semantic-based opinion summarization. *Proceedings of Recent Advances in Natural Language Processing*, 624–633. https://doi.org/10.11606/D.55.2021.tde-13092021-141741.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition* (2nd ed.). Prentice Hall.

Kusner, M. J., Sun, Y., Kolkin, N. I., & Weinberger, K. Q. (2015). From word embeddings to document distances. *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, 957–966. https://proceedings.mlr.press/v37/kusnerb15.pdf (accessed 23 August, 2022).

Legroski, M. (2009). Definindo metáfora. *Revista Polidisciplinar Eletrônica da Faculdade Guairacá*, *1*(2), 15–31. http://www.revistavoos.com.br/seer/index.php/voos/article/viewFile/42/02_Vol2_VOOS2009_CL1 (accessed 23 August, 2022).

Linh, H., & Nguyen, H. (2019). A case study on meaning representation for Vietnamese. *Proceedings of the First International Workshop on Designing Meaning Representations*, 148–153. https://doi.org/10.18653/v1/W19-3317.

Migueles-Abraira, N. (2017). *A study towards Spanish abstract meaning representation* [Master thesis]. Universidad del País Vasco. https://addi.ehu.es/bitstream/handle/10810/22056/NMA-MScThesis-June2017.pdf?sequence=5 (accessed 23 August, 2022).

Migueles-Abraira, N., Agerri, R., & Diaz de Ilarraza, A. (2018). Annotating abstract meaning representations for Spanish. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 3074–3078. https://aclanthology.org/L18-1486.pdf (accessed 23 August, 2022).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations*, 1–12. https://doi.org/10.48550/arXiv.1301.3781

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*(1), 71–106. https://doi.org/10.1162/0891201053630264.

Ramisch, C., Ramisch, R., Zilio, L., Villavicencio, A., & Cordeiro, S. (2018). A corpus study of verbal multiword expressions in Brazilian Portuguese. *Proceedings of the International Conference on Computational Processing of the Portuguese Language*, 24–34. https://doi.org/10.1007/978-3-319-99722-3_3.

Sobrevilla Cabezudo, M. A., & Pardo, T. A. S. (2019). Towards a general abstract meaning representation corpus for Brazilian Portuguese. *Proceedings of the 13th Linguistic Annotation Workshop*, 236–244. https://doi.org/10.18653/v1/W19-4028.

Sobrevilla Cabezudo, M. A., & Pardo, T. A. S. (2022). Low-resource AMR-to-text generation: A study on Brazilian Portuguese. *Procesamiento del Lenguaje Natural*, *68*, 85–97. https://repositorio.usp.br/directbitstream/a9595448-e887-4a55-9e7b-4789667ee3e0/3070380.pdf (accessed 23 August, 2022).

Yampolskiy, R.V. (2013). Turing Test as a Defining Feature of AI-Completeness. In X. S. Yang (Ed.), *Artificial intelligence, evolutionary computation and metaheuristics* (pp. 3-17). Springer.

White, A. S., Reisinger, D., Sakaguchi, K., Vieira, T., Zhang, S., Rudinger, R., Rawlins, K., & Van Durme, B. (2016). Universal decompositional semantics on universal dependencies. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1713–1723. https://aclanthology.org/D16-1177.pdf (accessed 23 August, 2022).

Wittenberg, E., Jackendoff, R., Kuperberg, G., Paczynski, M., Snedeker, J., Wiese, H., & Wittenberg, E. (2014). The processing and representation of light verb constructions. In A. Bachrach, I. Roy & L. Stockall (Eds.), *Structuring the argument: Multidisciplinary research on verb argument structure* (pp. 61–80). John Benjamins. https://doi.org/10.1075/lfab.10

Xue, N., Bojar, O., Hajič, J., Palmer, M., Urešová, Z., & Zhang, X. (2014). Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 1765–1772. http://www.lrec-conf.org/proceedings/lrec2014/pdf/384_Paper.pdf (accessed 23 August, 2022).

Recebido em: 06.08.2021
Aprovado em: 23.08.2022