Lexicalized Meaning Representation (LMR)

Jorge Baptista, Sónia Reis, João Dias, Pedro A. Santos

Lexicalized Meaning Representation (LMR)

Jorge Baptista[1,2], **Sónia Reis**[1,2], **João Dias**[1,2,3], **Pedro A. Santos**[2,4]
[1]U. Algarve - FCHS/FCT, [2]INESC-ID LIsboa, [3]CISCA, [4]U. Lisboa - IST
Faro/Lisboa, Portugal
jbaptis,smreis,jmdias@ualg.pt, pedro.santos@tecnico.ulisboa.pt

Abstract

This paper presents an adaptation of the Abstract Meaning Representation (AMR) framework for European Portuguese. This adaptation, referred to as Lexicalized Meaning Representation (LMR), was deemed necessary to address specific challenges posed by the grammar of the language, as well as various linguistic issues raised by the current version of AMR annotation guidelines. Some of these aspects stemmed from the use of a notation similar to AMR to represent real texts from the legal domain, enabling its use in Natural Language Processing (NLP) applications. In this context, several aspects of AMR were significantly simplified (e.g., the representation of multi-word expressions, named entities, and temporal expressions), while others were introduced, with efforts made to maintain the representation scheme as compatible as possible with standard AMR notation.

Keywords: Lexicalized Meaning Representation (LMR), Abstract Meaning Representation (AMR), Natural Language Processing (NLP), Portuguese

1. Introduction

This paper aims to contribute to the development of a theoretical and formal framework for the semantic annotation of natural language texts, facilitating the creation of tools for computational language processing. Semantic annotation of natural language texts aims to establish a representation of meaning that is valuable for developing various tools and applications (Damonte et al., 2017; Damonte and Cohen, 2018; Seno et al., 2022), particularly in Natural Language Processing (NLP). These applications include automatic sense disambiguation, machine translation, text summarization, and the generation of multilingual documents.

Various initiatives have been developed for this purpose. The Universal Networking Language (UNL) (Uchida et al., 1996)[1] provided a version of the novella *The Little Prince* (TLP) by Antoine de Saint-Exupéry (Martins, 2012) with the explicit aim of comparing representations of the same text in different languages. More recently, Abstract Meaning Representation (AMR) (Banarescu et al., 2013) has gained popularity in the NLP community. Originally proposed for English, this model aims to represent the meaning of sentences in a simplified form.

In a nutshell, each sentence's meaning is represented as a directed acyclic graph without a root. In this graph, nodes correspond to semantic predicates (operators) and their arguments, while arcs represent the semantic relations between the sentence elements. These relations, known as semantic roles, are defined in *OntoNotes* (Weischedel et al., 2013) and are associated with the arguments of (mostly) verbal predicates.

The frames of these verbal predicates form an ontology acting as a 'catalog' of meanings, serving as a reference for the various meanings of predicative elements represented in the graph. Additionally, other semantic relations are expressed by labeled arcs, linking predicates to different types of elements and circumstances, sometimes replacing textual elements that convey these relations. Grammatical elements such as auxiliary verbs, copulas, or support verbs are simply omitted. Many lexical elements are replaced either by verbs listed in *OntoNotes* or by other elements (e.g., adverbs ending in *-ly* are replaced by the morphologically associated adjectives and linked to an operator by the labeled arc :MANNER). Figure 1 illustrates the standard AMR graph representation of a simple English sentence extracted from the mentioned novella – *Draw me a sheep ...* [TLP:id=65], produced by the AMREager parser (Damonte et al., 2017)[2].

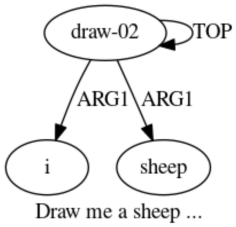

Figure 1: Standard AMR graph

The graph representing the meaning of the sentence can also be built in an equivalent PENMAN formalism (Matthiessen and Bateman, 1991). Such a PENMAN graph is shown in Figure 2 taken from the AMR annotation of the novella.[3] (The differ-

[1] http://www.unlweb.net/

[2] https://bollin.inf.ed.ac.uk/amreager.html

[3] https://amr.isi.edu/download/

5th Workshop on Designing Meaning Representations (DMR 2024) @LREC-COLING-2024, pages 101–111

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

Figure 2: AMR graph in PENMAN formalism

ences between the PENMAN graph and the AMR parser's output are deemed irrelevant for the purpose of this paper.)

Although AMR has been initially conceived for the English language and explicitly rejects the classification of inter-language (Banarescu et al., 2013), it naturally lends itself to the comparison of annotations of the same text in various languages (Xue et al., 2014). This annotation scheme was, rightly, adapted for the representation of texts, either by different annotators, or translations of the same text in different languages. Examples include annotations of the same novella in English, Chinese (Li et al., 2016), Spanish (Migueles Abraira, 2017), Turkish (Azin and Eryiğit, 2019; Oral et al., 2024), Vietnamese (Linh and Nguyen, 2019), Brazilian Portuguese (Anchiêta, 2020), and Persian (Takhshid et al., 2022).

On the other hand, the initial version of AMR aimed at describing individual sentences independently. Recently, however, the AMR guidelines were extended to the Unified Meaning Representation (UMR) formalism (Pustejovsky et al., 2019; Wein and Bonn, 2023) to encompass the annotation of sequences of sentences forming discourses (O'Gorman et al., 2018). Naturally, the original guidelines were occasionally reviewed and expanded to incorporate concepts that had not been sufficiently considered in the original proposal (Bonial et al., 2018). As recently mentioned by (Seno et al., 2022), following (Hovy and Lavid, 2010), these reformulations and extensions seek to achieve "the necessary balance between the depth of linguistic theory to be used and the stability of the annotation process", which does not prevent "critics within the community interested in this semantic representation, regarding some decisions made originally" (Seno et al., 2022, p. 51).

The primary objective of this work is to establish an annotation scheme, inspired by the AMR guidelines[4], which aims to address a set of difficulties and problems encountered in the solutions adopted thus far (see Section 2 and Table 1 for an overview).

To achieve this goal, we compared available Abstract Meaning Representation (AMR) annotations from parts of Antoine de Saint-Exupéry's work *The Little Prince* in English, Spanish (Migueles Abraira, 2017)[5], and Brazilian Portuguese (Anchiêta, 2020)[6], along with a Lexicalized Meaning Representation (LMR) annotated version of the same work in European Portuguese. We occasionally consulted the original French edition of the novella to verify any changes introduced by the translators.

Our focus was on the 50 Spanish sentences translated from the English version by Migueles Abraira (2017), ensuring a 4-tuple comparison. We conducted a critical analysis of these 50 sentences, considering observed phenomena and the annotation solutions adopted, comparing the similarities and differences between the annotations. Note that the translators' choices regarding the Portuguese or Spanish sentences are not considered here. Instead, the focus is solely on the structure and meaning of the translation output and the corresponding semantic representation (AMR/LMR). Due to space constraints, this paper provides only a succinct overview highlighting the main findings.

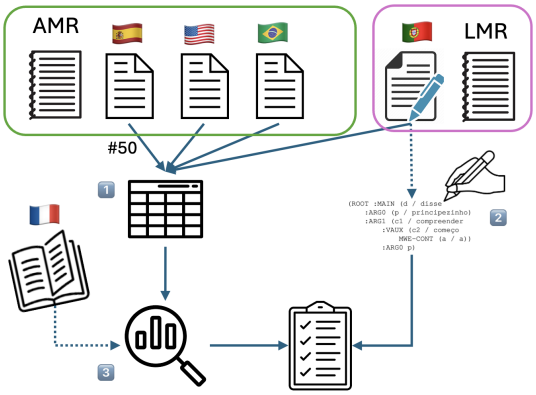

Figure 3: Comparing AMR/LMR annotations: alignment, annotation and analysis.

Figure 3 outlines the procedural stages of this study: (1) Alignment of sentences in different languages/varieties, considering translations from the original edition of the work (the English edition in the case of the Spanish version; the French edition in the case of the Portuguese translations) and resolving encountered alignment mismatches. (2) Annotation of sentences in the European Portuguese version of *The Little Prince* (Baptista, 2024b)[7], inde-

amr-bank-struct-v3.0.txt

[4]AMR 1.2.6. Specification (2019): https://github.com/amrisi/amr-guidelines/blob/master/amr.md

[5]https://github.com/ixa-ehu/amr-corpus-spanish/blob/master/es-Little-Prince-Corpus-50-AMR.txt

[6]https://github.com/rafaelanchieta/amr-br/blob/master/amr_br-v1.0.xml

[7]The European Portuguese, LMR-annotated sentences can be found at: https://gitlab.hlt.inesc-id.pt/u000803/lmr4pt/-/blob/master/public/LMR4PT_Principezinho.pdf.

pendently performed by two annotators and based on a set of LMR Guidelines autonomously developed by Baptista (2024a). These guidelines aim to: (a) adapt AMR to linguistic situations observed in Portuguese but not observed in English; (b) systematically make explicit and consistent the relation between text elements and annotation; and (c) adequately account for relevant linguistic phenomena not contemplated by the AMR framework. (3) Finally, a critical and systematic comparison of the annotations of sentences in the different languages was conducted.

2. Comparing AMR and LMR

Considering specific aspects of European Portuguese, as well as other fundamental requirements of semantic annotation, LMR introduces several extensions and reformulations of the standard AMR annotation scheme proposed by (Banarescu et al., 2013). Table 1 schematically presents the main differences between AMR and LMR annotations.

The Abstract Meaning Representation (AMR) annotation scheme is grounded in a 'catalog' of meanings derived from the verbal constructions present in *OntoNotes* (Weischedel et al., 2013). This methodology accepts both the reconstruction and suppression of textual elements, such as the insertion of pronouns in lexically unfilled syntactic positions or the replacement of conjunctions and prepositions with the semantic relations they convey. However, it does not encompass the analysis of auxiliary verbs, including copulative verbs (*Vcop*) and support verbs (*Vsup*) (or so-called *light* verbs). In fact, only some constructions with *Vsup* are considered, as most predicative nouns are assimilated into the corresponding verbal predicates (for example, [*a*] *purchase* → [*to*] *buy*). Additionally, AMR represents complex named entities (NE), particularly for denoting temporal and quantity values.

Lexicalized Meaning Representation (LMR), on the other hand, emphasizes a representation closely tied to the text, effectively constituting an annotation process where representation is directly anchored on the words of the sentences rather than merely appended to them as a whole. Furthermore, LMR strictly adheres to the principle of not replacing words in the text with arbitrary or theoretical constructs, instead anchoring relations on surface forms — the only visible elements that provide access to the meaning of the sentence[8].

[8]Certain types of zeroing, such as *appropriate* zeroing (Harris, 1976, 1991), pose serious challenges to this approach, for example, *John enjoyed* (*reading*) *the book*. These challenges must be addressed differently, though they are outside the scope of this paper.

As a semantic ontology or 'catalog' of word senses, LMR relies on the *Dicionário Gramatical de Verbos do Português* [Grammatical Dictionary of Portuguese Verbs] (DGVP; Baptista and Mamede, 2020a), built on the database of the lexicon-grammar of European Portuguese verbs (ViPEr; Baptista, 2012; Baptista, 2013 (Baptista and Mamede, 2020c)), as the 'catalog' of meanings of verbal constructions. For nominal predicates, the lexicon-grammar of predicative nouns (SNIPER; Baptista and Mamede, 2020b) is used. Since both resources indicate adjectival counterparts of these verbal and nominal predicates, and in the absence of a lexicon-grammar of adjectives proper for Portuguese, the adjective is referenced to either one or the other (or both) (for simplicity, these references were not provided in this paper).

For a semantic ontology or 'catalog' of word senses, LMR relies on the *Dicionário Gramatical de Verbos do Português* (DGVP; Baptista and Mamede, 2020a), which is built on the database of the lexicon-grammar of European Portuguese verbs (ViPEr; Baptista, 2012; Baptista, 2013), serving as the 'catalog' of meanings of verbal constructions. For nominal predicates, the lexicon-grammar of predicative nouns (SNIPER; Baptista and Mamede, 2020b) is utilized. Since both resources indicate adjectival counterparts of these verbal and nominal predicates, and in the absence of a lexicon-grammar of adjectives specific to Portuguese, the adjective is referenced to either one or the other (or both) (for simplicity, these references were not provided in this paper).

One of the major differences, thus, between AMR and LMR is that LMR adopts a homologous strategy for representing the predicate-argument relations from different grammatical categories, that is, verbs, nouns and adjectives. In this way, words in the texts are represented in LMR respecting their part-of-speech, keeping the representation closer to the text. For example, the sentence TLP id=348 is represented in AMR as shown in Figure 4:

One of the major differences, therefore, between AMR and LMR is that LMR adopts a homologous strategy for representing the predicate-argument relations from different grammatical categories — verbs, nouns, and adjectives. This approach ensures that words in the texts are represented in LMR according to their part-of-speech, maintaining a representation closer to the text. For example, the sentence TLP id=348 is represented in AMR as shown in Figure 4.

In this case, the adjective *important* in under the main predicate *think*, and the copula verb *be* is ignored. In turn, in the corresponding Portuguese sentence: – *Isso não é importante?!* 'That is not important?!' [TLP id=348], the copula verb is linked to the adjective it auxiliates (Figure 5):

Abstract Meaning Representation (AMR) (Banarescu *et al*. 2013)	**Lexicalized Meaning Representation (LMR)**
A catalog of senses (semantic predicates) for verbs can be found in OntoNotes by Weischedel, R. et al. (2013). Other categories such as nouns and adjectives are represented by verbal predicates.	A catalog of senses is available in the Lexicon-Grammar of Portuguese. For verbs, references include ViPEr by Baptista (2012, 2013) and the Dictionary of Portuguese Verb Grammar by Baptista & Mamede (2020a). Predicative nouns are covered in SNIPER by Baptista & Mamede (2020b).
Directed acyclic graphs lack a root node, instead employing an arc labeled :TOP looping over the main predicative element node of the sentence.	Directed acyclic graphs feature a ROOT node, which is connected to the main predicative element (:MAIN), serving as the node to which elements with scope over the entire sentence are connected.
Reduced elements are reconstructed.	No reconstruction of reduced elements is performed.
A graph representation is appended to the entire sentence, without establishing a direct relation between the graph nodes and the text forms.	There exists an explicit relation between text forms and their representation, treating text forms as nodes of the graph.
Predicative elements in the text are replaced by verbal lemmas (especially verbs represented in OntoNotes).	Predicative elements in the text are preserved in the graph, with the association of lemmas and constructions being carried out in the post-processing phase.
Some textual elements undergo substitution, especially grammatical ones (such as conjunctions, prepositions, etc.), by the semantic relations they express.	All textual elements undergo maintenance, alongside explicit representation of the semantic relations they convey; these include conjunctions, prepositions, subordinate gerund *-ndo* '-ing' morpheme, etc.
Auxiliary verbs, copulative verbs, or support verbs (light verbs) are not considered.	All types of auxiliary verbs are considered, including verbal auxiliaries (temporal, modal, and aspectual), adjectival auxiliaries (copulative verbs), nominal auxiliaries (support verbs), and auxiliaries of passive constructions. Additionally, constructions with (causative, linking and agentive) operator verbs are also taken into account.
Multi-word expressions (MWE) of varying complexity are represented, with a sophisticated representation of named entities (NE), and particularly temporal and quantification expressions.	Very simplified representation of multi-word expressions (MWE), named entities (NE), as well as temporal and quantification expressions. MWE and NE are identified in the pre-processing phase and integrated as nodes in the LMR graph.
Intra-phrasal anaphoric relations are represented, alongside an extension of notation (O'Gorman et al., 2018) for trans-phrasal anaphoric relations through coreference chains at the text level.	Intra-phrasal anaphoric relations are represented solely between explicit elements in the text, with anaphora resolution addressed as a post-processing task (trans-phrasal anaphoric relations are not yet considered).
Verbal predicates (standard representation) and adjectival (:DOMAIN) are treated distinctly, while nominal constructions are represented by verbal constructions if present in OntoNotes.	Verbal, nominal, and adjectival predicates feature a homologous representation of argument structure, corresponding to the standard representation: *predicate* (:ARG0, :ARG1, ...).

Table 1: Summarized comparison between Abstract Meaning Representation (AMR) and Lexicalized Meaning Representation (LMR)

You think that is not important ! . [id=348]

```
(t / think-01
   :ARG0 (y / you)
   :ARG1 (t2 / that
      :ARG1-of (i / important-01
         :polarity -)))
```

Figure 4: AMR Representation of sentence id=348

Isso não é importante ?! 'That is not important?!' [id=348]

```
ROOT :MAIN (i1 / importante
     :VAUX (ser / é)
     :NEG (n / não)
     :ARG0 (i2 / isso))
     :MODE-EXCLAMATIVE)
```

Figure 5: AMR Representation of sentence id=348

When the main predicative element is the corresponding predicative noun *importância* 'importance', it appears in an equivalent support verb construction with support verb *ter* 'have', represented by LMR as shown in Figure 6.

Notice that the role of the negation adverb is explicitly encoded and attached to the negation adverb *não* 'not'. This solution, however, is arguably equivalent to the AMR notation, though it avoids zeroing the negation adverb and anchors the negation con-

Isso não tem importância . lit.: 'That doesn't have importance'

```
ROOT :MAIN (i1 / importância
    :VSUP (t / tem)
    :NEG (n / não)
    :ARG0 (i2 / isso)))
```

Figure 6: LMR Representation: a predicative noun in a support-verb construction

struct on a textual element. Notice also that the exclamative mode of the sentence is attached to the `:ROOT` node, which is theoretically seen here as a more adequate representation (Harris, 1991) as it bears on the entire sentence. The lack of a root node in AMR forces the modality to be attached to the main predicative element (though the AMR notation, shown in Figure 4, fails to do).

A similar representation is also proposed for the corresponding verb, if it exists in the language (these triplets are not rare in Portuguese), e.g. – *Isso não importa?!* 'That [does] not matter?!':

```
(i1 / importa ... :ARG0 (i2 / isso)).
```

LMR maintains the equivalence relation between lexical elements by offering analogous representations for full verbs, predicative adjectives, and predicative nouns. It maintains notation closely tied to the text, anchoring semantic representation directly on its elements. While these paraphrastic equivalence relations (*transformational*, in the sense of Harris (1964, 1976, 1991)) should indeed be established, they are better suited for higher-order representation to minimize *ad hoc* interpretations during human annotation. Ideally, the "catalog of senses" or semantic predicates underlying the AMR/LMR notation should provide such equivalence. This is indeed the case for the works by Baptista and Mamede (2020a,c).

A notable contrast between the two schemes is that in LMR, a root node (`ROOT`) is instantiated for each sentence, with a `:MAIN` dependency linking this node to the main predicative element. This resolves a technical issue previously highlighted by Anchiêta (2020) regarding the evaluation of competing semantic representations for the same sentence. Still, this also affects the adequacy of representing elements that operate on the entire sentence, such as sentence-external adverbial modifiers, as defined by Molinier and Levrier (2000). For instance, in the sentence *But my drawing is certainly very much less charming than its model* [TLP id=52], the adverb *certainly* imparts a modality value to the entire sentence, akin to *It is certain that my drawing is very much less charming than its model*. In such cases, and unlike AMR that hinges the `:mod (c / certain)` under another node of the graph, LMR suggests representing the adverb as a modifier on the ROOT node:

```
(ROOT :MOD (c / certainly) ...
```

By closely adhering to the text and preserving the words' part of speech, LMR effectively distinguishes between the main types of adverbial constructions: sentence-external and sentence-internal adverbs. Moreover, astute readers may have observed the conjunction *but* at the sentence's outset, serving to connect it with preceding discourse in a manner akin to *conjunctive adverbs* (or *discourse connectives*). Consequently, the identical descriptive approach is employed for both scenarios.

```
(ROOT :MOD (b / but) ...
(ROOT :MOD (b / furthermore) ...
```

Furthermore, LMR incorporates auxiliary verbs, encompassing copulative and support verbs, into its analysis. This inclusion is justified by the significance attributed to these elements as integral components of textual meaning units. Indeed, Portuguese features a particularly rich system of auxiliary verbs (Baptista et al., 2010; Baptista and Crismán Pérez, 2021), particularly for expressing aspectual nuances. For instance, in the sentence: – *Começo a compreender, disse o principezinho.* 'I begin to understand, said the little prince.' [TLP id=1080], the auxiliary *começar a* 'begin to' is represented as:

```
(ROOT :MAIN (d / disse
    :ARG0 (p / principezinho)
    :ARG1 (c1 / compreender
        :VAUX (c2 / começo
            MWE-CONT (a / a))
        :ARG0 p)
```

The auxiliary construction is depicted as a multiword expression, with a `:MWE-CONT` arc linking the auxiliary verb to the preposition it introduces. This enables distinguishing its precise aspectual value from other nuanced constructions involving the same verb but with a different preposition (e.g., *começar por* for 'begin by'). The representation of modal auxiliaries is particularly relevant for legal domain texts, where deontic modality is essential. Two domain-specific relations were devised solely for this purpose, `:dever` 'must/ought' and `:poder` 'may/can', corresponding to the verbs most commonly used with that function. Besides, a similar notation was devised for all types of auxiliary verbs. In many situations, it is possible to keep LMR compatible with AMR (except for modal auxiliaries, treated as full predicates in AMR).

LMR also adopts a simplified representation both of multi-word expressions (e.g., compound nous, idioms) and of named entities (e.g. people, organizations, and places), as well as temporal and quantification expressions, delegating this task to a pre-annotation step, prior to the semantic annotation.

Other differences of detail were envisaged. For instance, in relative sub-clauses, e.g. *the girl who*

adjusted the machine, while AMR eliminates the relative pronoun:

```
(g / girl
     :ARG0-of (a / adjust-01
          :ARG1 (m / machine)))
```

LMR keeps the relative pronoun in the representation, maintaining consistency in the representation of the predicate-argument structure of sub-clause's predicate:

```
(g / girl
     :ARG0-of (a / adjust-01
          :ARG0 (w / who)))
          :ARG1 (m / machine)))
```

An aspect of language-specific adaptation is the existence of the so-called gerundive reduced subclauses. Here, we analyse the gerund morpheme (the *-ndo* '-ing' verb ending) as having a function similar to that of an adverbial subordinative conjunction, but with an underspecified semantic value. In fact, the nexus between the main clause and the gerundive subclause is often difficult to determine (cause, time). In order not to 'force' any interpretation, a generic `:NDO` is proposed (Figure 7).

O vaidoso recomeçou a agradecer, tirando o chapéu.
'The vain person started to thank again, tipping his hat.' [TLP id=620]

```
(ROOT :MAIN (a / agradecer
    :VAUX (r / recomeçou
        :MWE-CONT (a / a))
    :ARG0 (v / vaidoso)
    :NDO (t / tirando
        :ARG0 v
        :ARG1 (c / chapéu))))
```

Figure 7: Gerundive subclauses and :NDO

In the Brazilian Portuguese annotation of the same construction, one finds either the `:subevent-of` relation[9], or `:manner`, or even an `:arg2-of` (id=344). AMR deals with English similar gerundive sub-clauses (for example, id=631) in the same way as with relative subclauses, v.g. *"I admire you," said the little prince, shrugging his shoulders slightly, ...*:

```
(s / say-01
   :ARG0 (p / prince :mod (l / little)
      :ARG0-of (s2 / shrug-01
      :ARG1 (s3 / shoulder :part-of p)
      :degree (s4 / slight))) ...
```

This is not, arguably, a representation exactly equivalent to the meaning that the gerund subordinate operator *-ing* introduces in the sentence (two simultaneous actions). In fact, the equivalent relative clause would be: *The prince* that shrugged his shoulders *said "I admire you"* ...

[9] https://www.isi.edu/~ulf/amr/lib/amr-dict.html#:subevent

On the other hand, the gerund bound morpheme is, in fact, present in the sentence, and in spite of not being able to "detach" it from the base (or host) verb, its value, vague as it is, is made explicit with the notation `:NDO`.

These methodological differences between AMR and LMR result from partly distinct approaches in the semantic representation of texts: although each presents its specific advantages and challenges, LMR distinguishes itself by seeking to reconcile the precision of semantic representation and fidelity to the underlying text, suggesting a potentially more precise approach in semantic analysis.

3. Contrastive analysis

To illustrate the systematic contrastive analysis of the notations of *The Little Prince* in the four languages here considered, we present a case study by commenting on the following sentence with id=300:

FR: *J'étais très soucieux car ma panne commençait de m'apparaître comme très grave, et l'eau à boire qui s'épuisait me faisait craindre le pire.*

In the English version, this sentence is split into two (id=299 and id=300), which we present below.

EN: *I was very much worried, for it was becoming clear to me that the breakdown of my plane was extremely serious. And I had so little drinking-water left that I had to fear for the worst.*

In the case of the Spanish translation (Migueles Abraira, 2017), which faithfully follows the English version, only the AMR representation of the second sentence is available.

ES: *Y me quedaba tan poca agua potable que me temía lo peor.* [SP id=15]

For Brazilian Portuguese (Anchiêta, 2020), which was based on the French version of the text, we find a very losely translated equivalent sentence:

BR: *Minha pane começava parecer demasiado grave, e em, breve já não teria água para beber ...*

Finally, for European Portuguese, the translator faithfully follows the French original:

PT: *Estava bastante inquieto, pois a avaria começava a parecer grave, e a pouca água que restava para beber fazia-me temer o pior.*

We start the analysis by commenting the standard AMR representation, made for the English version (Figure 8).

The first observation is the replacement of the causal subordinated conjunction *for* by the abstract construct `cause-01`. This construct takes the following arguments: as `:ARG0`, the causal subordinate clause (*it was becoming clear to me that* ...); and as `:ARG1` the main clause (*I was very much worried*).

I was very much worried, for it was becoming clear to me that the breakdown of my plane was extremely serious. And I had so little drinking-water left that I had to fear for the worst. [EN id=299.300]

```
(c2 / cause-01
  :ARG0 (c / clear-06
    :ARG1 (s / serious-02
      :ARG1 (b / break-down-12
        :ARG1 (p / plane
          :poss i))
        :degree (e / extreme))
    :ARG2 (i / i))
  :ARG1 (w / worry-01
    :ARG1 i
    :quant (m / much
    :degree (v / very))))

(a / and
  :op1 (h3 / have-degree-91
    :ARG1 (w / water
      :purpose (d / drink-01)
  :ARG1-of (l2 / leave-17)
  :ARG1-of (h / have-03
    :ARG0 (i / i)))
  :ARG2 (l / little)
  :ARG3 (s / so)
  :ARG6 (o / obligate-01
    :ARG1 i
    :ARG2 (f / fear-01
      :ARG0 i
      :ARG1 (t / thing
        :ARG1-of (h2 / have-degree-91
          :ARG2 (b / bad-07)
          :ARG3 (m / most)))))))
```

Figure 8: English AMR representation of sentence id=299.300

In the case of the adjectival construction of `clear-06`, where a subject clause is extraposed, the subject is linked by an `:ARG1` arc, as indicated in the directives[10]. However, the 3-argument frame of `clear-06` (a verb?) had been defined with a "cause" role for its `:ARG0` (?), which is now expressed by an independent node `cause-01`.

On the other hand, representing the construction of `serious-02` as a predicate with only one argument – *something is serious* – raises difficulties in justifying the semantic relation of `:ARG1` to the subject (`break-down-12`) of this adjective. In the Ontonotes[11], `serious-02` does not even have an `ARG0` role. This highlights how the association of adjectival predicates with verbal lemmas may not be entirely appropriate. The notation of these arguments as `:ARG1` is more of an artifact of the Ontonotes representation scheme than a regular (and generalizable) configuration between semantic predicates and their arguments.

In the case of the adjectival predicate `worry-01` (*worried*), such perplexity does not arise. Its predicative structure could effectively be described by the corresponding verbal construction, given its classification as a so-called 'psychological' verb (class 04, (Baptista and Mamede, 2020a)). This would correspond to the structure *something cause somebody to worry* = *something worries somebody*. In this construction, the verb exhibits a *causative* subject and an *experiencer* complement, filled by a human noun, here represented by the pronoun *I*, to which the `:ARG1` relation could correspond.

Regarding the second sentence, the AMR annotation relies on an abstract conceptualization of predicates such as `have-degree-91`[12], which is associated with adjectival constructions expressing gradable predicates, and `have-03`[13], corresponding to the full verb *have* in the sense of "possession". However, interpreting the representation of this sentence, simplified below, remains challenging:

```
h3 / have-degree-91
  :ARG1 (w / water
    :ARG1-OF (h / have-03
      :ARG0 (i /i)))
```

This configuration does not match the sentence we are analyzing: we encounter the verb *have* with the object *water*, quantified by *so little*. Moreover, the second verb *have* (`have-03`) typically represents the meaning associated with 'possession', making the presence of both operators appear redundant, at the very least.

In the sentence *I had to fear*, the modal auxiliary *have* is replaced by the operator `obligate-01`. However, this replacement ignores the nature of the modal auxiliary, which, being transparent to the selection restrictions of the main verb *fear*, should have the same subject as this verb. Consequently, the operator appears with its subject marked as an `:ARG1`, a consequence of the substitution of the auxiliary by `obligate-01`.

Lastly, the expression *fear for the worst* is represented in a manner that attempts to analyze its idiomatic value, rather than recognizing its non-compositional semantics, which is already lexicalized.

Regarding the sentence in Spanish, corresponding only to the second sentence of the English version (id=300), the notation closely follows the standard AMR representation, as usual (Figure 9).

The conjunction *y* (and) is used here to connect the current sentence to the previous one. However,

[10] https://amr.isi.edu/doc/amr-dict.html#:domain

[11] https://propbank.github.io/v3.4.0/frames/serious.html#serious.02

[12] https://amr.isi.edu/doc/amr-dict.html#have-degree-91

[13] https://propbank.github.io/v3.4.0/frames/have.html#have.03

Y me quedaba tan poca agua potable que me temía lo peor. [SP id=15]

```
(y2 / y
  :op1 (c / causar
    :ARG0 (q / quedar
      :ARG1 (a / agua
        :mod (p / potable)
        :mod (p2 / poco
          :grado (t / tan)))
      :ARG2 (y / yo))
    :ARG1 (t / temer
      :ARG0 y
      :ARG1 (m / malo
        :grado (m2 / máximo)))))
```

Figure 9: Spanish AMR representation of sentence SP id=15

this conjunction is treated like any other coordination situation. Since there is no second coordinated element, only one conjunctive operator `:OP1` is given The operator `:OP1` should connect the conjunction to the first member of the coordination. No second member of the coordination exists, since it is the entire sentence that is being put in relation to a previous discourse. Now, accepting this to be the function of *y* (as well as that of *and*, in the English version), the first member of the coordination should be the previous sentence. As AMR does not currently handle this type of cross-sentential relations (but see (O'Gorman et al., 2018)), any notation would always be incomplete. Nevertheless, the choice of `:OP1` seems somewhat ambiguous.

Another interesting aspect is the simplification (and closer adherence to the text) of the representation of the constituent *tan poca agua potable* 'so little drinking water', an argument of *quedar* 'to be left', which is based on the words of the text and does not resort to the type of constructs seen in standard AMR. Nevertheless, we analyze this *quedar* construction as a predicate with two arguments, where *agua* 'water' should correspond to the `:ARG0`, while the first-person dative pronoun *me* corresponds to an `:ARG1`.

Finally, as in English, the annotator intended to represent the expression *lo peor* 'the worst', making it corresponds to elements that are not present in the text (*malo máximo*).

Now, let's examine the analysis of the translation in Brazilian Portuguese, comparing it with the original French version. In this sentence, the translator omitted the main clause, with the predicate *soucieux* 'worried' and the causal conjunction *car* 'for' that links it to the rest of the sentence. Similarly, there was a profound transformation of the second subordinate clause under *car*: *et l'eau à boire qui s'épuisait me faisait craindre le pire* is translated as *e em, breve já não teria água para beber...* The construction with the operator-verb *faire* 'to make' disappears, as well as the construction of the verb *s'épuiser* 'to run out/exhaust'. The idiomatic expression *craindre le pire* 'to fear the worst' also disappears. In this case, this creative translation does not allow for a direct comparison between the annotation solutions adopted among the different languages, but only a generic comment on the AMR representation produced (Figure 10).

Minha pane começava parecer demasiado grave, e em, breve já não teria água para beber ... [BR id=299;300]

```
(c / começar-01
  :ARG0 (p / pane
    :poss (m / minha))
    :ARG1 (p1 / parecer-01
      :ARG2 (g / grave
        :degree (d / demasiado)))
    :cause (t / ter-01 :polarity -
      :ARG0 (e / eu)
      :ARG1 (a / água)))
```

Figure 10: Brazilian Portuguese AMR representation of sentence id=300

Let's start by noting the treatment of *começar* 'begin', here an auxiliary verb of *parecer* 'seem', as well as the verb *parecer* itself, that are represented as full verbs. It is difficult to entertain the idea of *começar* and *parecer* as full verbs, deviating from the more conventional analysis as a copulative verbs in an adjectival construction. The relation (`:ARG2`) between this verb *parecer* and the adjective *grave* 'serious' presents an even greater challenge to comprehension.

As previously mentioned, the principle of distributional transparency of auxiliaries regarding the selection restrictions imposed by the elements they 'modify' (Baptista et al., 2010; Baptista and Crismán Pérez, 2021) suggests an analysis in which *grave* 'serious' functions as the main predicative element of this clause, with *pane* 'breakdown' as its :DOMAIN, as follows, while consistency with AMR guidelines would lead to eliminate both copula verbs:

```
(g / grave :DOMAIN (p / pane))
```

The second interesting aspect is that the coordinative conjunction *e* 'and' has been removed and replaced bay a causal relation, as denoted by the operator `:CAUSE`. While not implausible, this interpretation seems unmotivated. Finally, note the suppression of the temporal adverbial phrase *em breve* 'soon', without any apparent reason.

Finally, let's look at the translation in European Portuguese and the proposal for its annotation in LMR (Figure 11). This translation is much more 'faithful' to the original French version, only taking the liberty to modify *l'eau à boire qui s'épuisait* into

Estava bastante inquieto, pois a avaria começava a parecer grave, e a pouca água que restava para beber faziame temer o pior. [PT id=300]

```
ROOT :MAIN (i / inquieto
  :ARG0 m
  :VAUX (e / estava)
  :DEGREE (b1 / bastante)
  :CAUSE (p1 / pois
    :OP2 (e / e
      :COORD1 (g / grave
        :ARG0 (a / avaria)
        :VAUX (p2 / parecer
          :VAUX (c / começava
            MWE_CONT (a / a))))
      :COORD2 (f / fazia
        :CAUSE (a / água
          :QUANT (p3 / pouca)
          :ARG0-OF (r / restava
            :ARG0 (q / que)
            :PURPOSE (p4 / para
              :OP2 (b2 / beber)))
        :VOPC (top / temer_o_pior
          :ARG0 (m / me))))))
```

Figure 11: European Portuguese LMR representation of sentence id=300

a pouca água que restava para beber. This modification alters the dependency of the verb *beber* 'to drink' and inserts the quantifier *pouca* 'little' associated with the use of the verb *restar* 'to be left'. This sentence allows us to present several interesting aspects of the LMR annotation scheme. Firstly, the use of the `:OP2` operator, 'repurposed' from standard AMR to link the conjunctions *pois* 'for' and *para* 'to' to the sentences they introduce. Since the precise semantic value these conjunctions convey are (mostly) lexically determined, LMR keeps the conjunctions and the link they establish between the main clause and the sub-clause. Notice that standard AMR notation simply abstract away from the conjunction proper.

A second aspect is the explicit representation of coordination relations using the `:COORD1` and `:COORD2` operators, rather than the generic `:OP1` and `:OP2` in standard AMR. These `:COORD` operators fulfill the same function, maintaining close parallelism between the two notations.

We also analyze the verb *parecer* 'seem' following a fairly traditional approach, as a copulative verb, i.e. an auxiliary of the adjective *grave* 'serious' and the recursive auxiliary verb chain *começar a parecer* 'begin to seem'.

Another notable aspect is the treatment of relative clauses. These are connected by linking the antecedent of the relative pronoun to the verb of the relative clause via an 'inverted' `ARGn-OF` relation, where 'n' denotes the semantic relationship of this element in the base clause of the relative. Subsequently, this relation is reiterated, without inversion, between the verb of the relative clause and the relative pronoun.

Lastly, we introduce the concept of the *causative operator-verb* (*Vopc*; (Gross, 1981), (Baptista, 2005, 202 ff.)). This concept entails an operator applied to a sentence, augmenting it with an additional argument, and establishing a causal relation between this extra argument and the base sentence. In our example, the verb *fazer* (to make) fulfills this function: *A água fazia*/Vopc # *eu temia o pior* (The water made/I feared the worst). For such operators, LMR suggests delineating two relations: firstly, `:CAUSE`, connecting the operator-verb to its subject; secondly, the relation `:VOPC`, linking the operator-verb to the embedded sentence. Notice also the recognized idiomatic verbal expression *temer o pior* (to fear the worst) as a single node (Galvão et al., 2019b,a).

4. Conclusion

Throughout this article, we have underscored the challenges inherent in implementing standard AMR directives and have explored the potential of the LMR annotation proposal. It is evident that discrepancies arise not only from variations in original versions or translator choices but also from inconsistencies in applying AMR directives (particularly pronounced in translations into Spanish and Brazilian Portuguese). LMR's approach, which anchors directly to the text, offers a promising solution by providing a representation that is closer to the text and less susceptible to the inherent inconsistencies in the process of abstracting the meaning of a text.

In our future endeavors, we intend to expand the annotated texts in LMR, completing the annotation of *O Principezinho* (The Little Prince) and incorporating texts from various genres and domains, including more legal texts.

We plan to develop tools to facilitate faster and more efficient annotation implementation, including: (a) a lemmatizer to associate text forms with lemmas and unique identifiers in the lexicon-grammar; (b) a tool for constructing LMR graphs, which instantiate argument positions of predicative elements and mark positions for anaphora resolution, ensuring formal consistency; (c) a tool for converting graphs into graphical or PENMAN format to facilitate interpretation; (d) a tool for comparing annotations and assessing agreement among annotators, and subsequently, across translations in different languages. With a more extensive corpus, our objective is to develop an LMR parser for automatic representation generation, with the potential for several NLP applications.

5. Acknowledgments

This work was partially supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia FCT (Reference: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020) and by European funds through the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837).

6. Bibliographical References

Rafael Anchiêta. 2020. *Abstract Meaning Representation Parsing for the Brazilian Portuguese Language*. Ph.D. thesis, Universidade de São Paulo.

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.

Jorge Baptista. 2005. *Sintaxe dos Predicados Nominais com* ser de. Fundação para a Ciência e a Tecnologia & Fundação Calouste Gulbenkian, Lisboa.

Jorge Baptista. 2012. ViPEr: A Lexicon-Grammar of European Portuguese Verbs. In *31e Colloque International sur le Lexique et la Grammaire*, pages 10–16.

Jorge Baptista. 2013. Viper: uma base de dados de construções léxico-sintáticas de verbos do português europeu. *Actas do XXVIII Encontro da APL-Textos Selecionados*, pages 111–129.

Jorge Baptista. 2024a. Lexical Meaning Representation - Guidelines. Technical report, University of Algarve/INESC-ID Lisboa. https://gitlab.hlt.inesc-id.pt/u000803/lmr4pt/.

Jorge Baptista. 2024b. LMR4PT - Principezinho. Technical report, University of Algarve/INESC-ID Lisboa. https://gitlab.hlt.inesc-id.pt/u000803/lmr4pt/.

Jorge Baptista and Rafael Crismán Pérez. 2021. Auxiliary verb constructions in Portuguese and Spanish: a comparative study and its applications as second languages. *Revista de Lenguas Modernas*, 34:39–57.

Jorge Baptista and Nuno Mamede. 2020a. *Dicionário gramatical de verbos do português*. Universidade do Algarve.

Jorge Baptista and Nuno Mamede. 2020b. Syntactic Transformations in Rule-Based Parsing of Support Verb Constructions: Examples from European Portuguese. In *9th Symposium on Languages, Applications and Technologies (SLATE 2020)*, pages 11:1–11:14. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Jorge Baptista and Nuno Mamede. 2020c. ViPEr v.1. Portulan-CLARIN-PT repository hosted at Research Infrastructure for the Science and Technology of Language, Handle: https://hdl.handle.net/21.11129/0000-000D-F91E-A.

Jorge; Baptista, Nuno; Mamede, and Fernando Gomes. 2010. Auxiliary verbs and verbal chains in European Portuguese. In *Computational Processing of the Portuguese Language*, number 6001 in Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence, pages 110–119, Berlin. PROPOR 2010, Springer.

Claire Bonial, Bianca Badarau, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Tim O'Gorman, Martha Palmer, and Nathan Schneider. 2018. Abstract Meaning Representation of constructions: The more we include, the better the representation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Marco Damonte, Shay B. Cohen, and Giorgio Satta. 2017. An incremental parser for Abstract Meaning Representation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 536–546, Valencia, Spain. Association for Computational Linguistics.

Ana Galvão, Jorge Baptista, and Nuno Mamede. 2019a. New developments on processing European Portuguese verbal idioms. In *12th Symposium in Information and Human Language Technology*, pages 229–238, Salvador, BA (Brazil).

Ana Galvão, Jorge Baptista, and Nuno Mamede. 2019b. Processing European Portuguese Verbal Idioms: From the Lexicon-Grammar to a Rule-based Parser. In *Computational and Corpus-based Phraseology. Proceedings of the Third International Conference EUROPHRAS 2019*, pages 70–77, Malaga (Spain). Tradulex.

Maurice Gross. 1981. Les bases empiriques de la notion de prédicat sémantique. *Langages*, 15(63):7–52.

Zellig Harris. 1964. The elementary transformations. In Henry Hiz, editor, *Papers on Syntax*, pages 211–235. D. Reidel Pub. Co.

Zellig Sabettai Harris. 1976. *Notes du Cours de Syntaxe*. Seuil, Paris. (edited by Maurice Gross).

Zellig Sabettai Harris. 1991. *A Theory of Language and Information. A Mathematical Approach*. Clarendon Press, Oxford.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation*, 22(1):13–36.

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.

Ronaldo Martins. 2012. Le Petit Prince in UNL. In *LREC*, pages 3201–3204. Citeseer.

Christian MIM Matthiessen and John A Bateman. 1991. Text generation and systemic-functional linguistics: experiences from English and Japanese. *Pinter Publishers*.

Noelia Migueles Abraira. 2017. *A Study Towards Spanish Abstract Meaning Representation*. University of the Basque Country. (Master thesis).

Christian Molinier and Françoise Levrier. 2000. *Grammaire des adverbes: description des formes en* -ment. Droz, Genève.

Tim O'Gorman, Michael Regan, Kira Griffitt, Ulf Hermjakob, Kevin Knight, and Martha Palmer. 2018. AMR beyond the sentence: the multi-sentence AMR corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3693–3702, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Elif Oral, Ali Acar, and Gülşen Eryiğit. 2024. Abstract Meaning Representation of Turkish. *Natural Language Engineering*, 30(1):171–200.

James Pustejovsky, Ken Lai, and Nianwen Xue. 2019. Modeling quantification and scope in Abstract Meaning Representations. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 28–33, Florence, Italy. Association for Computational Linguistics.

Eloize Seno, Helena Caseli, Marcio Inácio, Rafael Anchiêta, and Renata Ramisch. 2022. XPTA: um parser AMR para o português baseado em uma abordagem entre línguas. *Linguamática*, 14(1):49–68.

Reza Takhshid, Razieh Shojaei, Zahra Azin, and Mohammad Bahrani. 2022. Persian abstract meaning representation.

Hiroshi Uchida, M Zhu, and T Della Senta. 1996. Unl: Universal networking language–an electronic language for communication, understanding, and collaboration. *Tokyo: UNU/IAS/UNL Center*.

Shira Wein and Julia Bonn. 2023. Comparing UMR and cross-lingual adaptations of AMR. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 23–33, Nancy, France. Association for Computational Linguistics.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23:170.

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).