Hurdles in Parsing Multi-word Adverbs: Examples from Portuguese

Izabela Muller and **Jorge Baptista**
Univ. Algarve, Faro, Portugal
INESC-ID Lisboa, HLT
R. Alves Redol 9, Lisboa
belagrein@inesc-id.pt
jorge.baptista@inesc-id.pt

Nuno Mamede
Univ. Lisboa - IST
INESC-ID Lisboa, HLT
R. Alves Redol 9, Lisboa
Nuno.Mamede@tecnico.ulisboa.pt

Abstract

This paper addresses the challenges posed by multi-word adverbs in the context of natural language parsing, with a specific focus on Portuguese adverbs; e.g. *à beça* 'a lot' and *às três pancadas* 'in a hurry' or 'carelessly'. These adverbs present complex combinatorial constraints and often carry non-compositional, idiomatic meanings, making them a significant hurdle for natural language processing systems. Recognizing them as distinct lexical units at an early stage of parsing is crucial. To investigate this issue, we conducted experiments using a selection of the 300 most frequently occurring multi-word adverbs from the Portuguese TenTen2020 corpus. We employed two different parsers, one rule-based and one statistical-based, and presented the results of these experiments. The main goal of this paper is to advocate for the development of enhanced lexical resources to facilitate the accurate parsing of these expressions. This includes broader lexical coverage of these units and a more detailed syntactic description, particularly about distinguishing between sentence-external and sentence-internal modifiers. Furthermore, we provide an update on an ongoing project aimed at creating this crucial lexical resource, with a specific focus on Brazilian Portuguese. Our current dataset includes 3,500 compound adverbs, each annotated with syntactic and semantic information, and it covers two regional varieties of Portuguese, namely Brazilian and European Portuguese.

1 Introduction

In recent years, there has been a significant increase in research dedicated to the investigation and analysis of *multi-word expressions* (MWE) (Ramisch, 2015; Constant et al., 2017; Kahane et al., 2018; Savary et al., 2023; Mel'čuk, 2023), especially within Natural Language Processing (NLP).

In this study, our focus lies on *compound adverbs* (Gross, 1986). These are *lexical units* composed of multiple words that exhibit constraints on the syntactic combination of their elements and often display a degree of semantic non-compositionality. In other words, the syntactic properties and overall meaning of a compound adverb cannot be derived from the properties and meanings of its constituent elements when considered separately.

Understandably, the absence of comprehensive descriptions of MWEs may lead certain NLP systems to tokenize and parse the words in these expressions as independent lexical units, thereby assuming a compositional relationship between the elements of the expressions. Such an approach can complicate various NLP tasks (Foufi et al., 2017), such as machine translation, word-sense disambiguation, information retrieval, and more. For instance, Gonçalves et al. (2020) highlighted several limitations of existing Portuguese NLP systems when parsing sentences containing MWEs, including adverbs.

In this article, we present and illustrate some of the primary challenges involved in identifying Portuguese adverbial MWEs. We showcase these challenges through the lens of two Portuguese parsers: (a) LX-DepParser (Branco et al., 2014), a statistically-based parser, developed within the framework of Universal Dependencies and trained on the CINTIL corpus (Barreto et al., 2006); and (b) the STRING processing chain (Mamede et al., 2012), which employs a rule-based parsing module called XIP (Xerox Incremental Parser) (Ait-Mokhtar et al., 2002) along with its lexical resources for Portuguese.

This paper is structured as follows: Section 2 presents some of the main linguistic aspects about parsing multi-word adverbs. Section 3 describes the experimental setup and Section 4 presents the results and discussions. Finally, Section 5 draws some conclusions from these experiments and proposes the next steps ahead.

2 Parsing Adverbs: hurdles and challenges

Adverbs are a complex and multifarious part-of-speech. (Quirk et al., 1985; Guimier, 1996; Gross, 1996a; Larsen-Freeman and Celce-Murcia, 2016). However, extant linguistic descriptions, drawing from several theoretical perspectives, have led to a certain degree of consensus regarding the prerequisites for their proper parsing. One of these prerequisites is the distinction between the two main types of adverbs – sentence adverbs and verb modifiers; see, for example, Mørdrup (1976), among others. This is not new and has, in fact, an already long grammatical tradition, including in Portuguese linguistics (Cunha and Lindley Cintra, 1986; Costa, 2008; Bechara, 2012; Paiva Raposo, 2013). In this paper, we adhere to the Lexicon-Grammar perspective (Gross, 1996b), as developed in Gross (1986) for the syntax of adverbs, and, particularly, we adapted the syntactic-semantic classification proposed by Molinier and Levrier (2000). In particular, we consider the distinction, as made by the latter authors, between adverbs functioning as *sentence-external* or *sentence-internal modifiers*.

Sentence-external adverbs *simultaneously* verify two conditions:
(1) The adverb can be fronted to the beginning of the sentence, and the sentence can be put in a negative form. This test/property demonstrates that the adverb is out of the scope of a negation adverb, which directly modifies the predicate, and hence it modifies the entire sentence. For example, many *conjunctive adverbs* (*a.k.a. discourse connectives*) link the current sentence to a previous utterance in the discourse. Thus, they are sentence-external modifiers, and negation has no bearing on them: *No entanto, o Pedro (não) gosta de futebol* 'However, Pedro likes/does not like football'.
(2) The adverb cannot undergo *extraction* with *ser. . .que* (*a.k.a. clefting*); this is a constituency test that only applies to sentence-internal constituents. Hence, we observe the unacceptable sequences: **É no entanto que o Pedro (não) gosta de futebol* 'It is however that Pedro likes/does not like football'.

Both properties are necessary and sufficient to classify *no entanto* 'however', and many other syntactically similar adverbs, as a sentence-external modifiers.

Conversely, adverbs that do not simultaneously satisfy both properties are considered sentence-internal modifiers, having their scope on a specific sentence constituent, typically a verb. A common scenario is when a *manner adverb* modifies a verb, as in: *O Pedro contactou telefonicamente o João* 'Pedro contacted João by phone', as its syntactical behavior sharply contrasts with that of sentence-external modifiers. In this case, we can confirm the unacceptability of the sequence when the first test is applied, while the sentence is deemed acceptable on the second test: (i) **Telefonicamente, o Pedro não contactou o João* 'By phone, Pedro did not contact João', and (ii) *Foi telefonicamente que o Pedro contactou o João* 'It was by phone that Pedro contacted João'. For a comprehensive classification of Portuguese multi-word adverbs, please refer to references (Palma, 2009; Català et al., 2020; Müller et al., 2022; Müller et al., 2023).

In addition to this broader classification, we would like to highlight a special sub-class of sentence-internal adverbs known as *focus adverbs* (Baptista and Català, 2009), such as *em especial/especialmente* 'especially'. In this case, the adverb places focus on a specific sentence constituent: *O Pedro gosta em especial/especialmente de chocolates* 'Pedro especially likes chocolates'. Consequently, the adverb cannot be extracted separately, as demonstrated by the unacceptable sequence: **É em especial/especialmente que o Pedro gosta de chocolates* 'It is especially that Pedro likes chocolates'. Instead, it can be extracted along with the constituent it focuses on, as seen in the acceptable construction *É em especial/especialmente de chocolates que o Pedro gosta* 'It is especially chocolates that Pedro likes'.

Regarding parsing, and specifically dependency parsing, the type of adverb (sentence-external/internal, focus) plays a crucial role in determining the syntactic dependencies between the adverb and the sentence's elements. Sentence-internal adverbs typically modify the sentence's verb. A simplified representation could be *contactar* 'contact' > MOD > *telefonicamente* 'by phone'. The direction of the dependency is a matter of formalism, not a significant conceptual difference. However, when it comes to sentence-external adverbial modifiers, as they modify the entire sentence rather than just one of its constituents, it is not entirely accurate to parse them as modifying the verb, or any other element within the sentence, for that matter.

Many parsers establish a ROOT node, serving as the starting point for constructing the depen-

dency graph of the entire sentence. It is possible that either this ROOT node or an analogous intermediate node, representing the entire sentence, could be considered as the point to which the sentence-external adverbial modifier may be connected. Hence, for the sentence *No entanto, o Pedro não gosta de futebol* 'However, Pedro does not like football', either ROOT > MOD > *no entanto* (or another representation with an intermediate node instead of ROOT), should be used. This is the solution proposed in this paper. On the contrary, the more common representation *gosta* > MOD > *no entanto* does not seem adequate.

In the context of focus adverbs, it appears more suitable to parse them as modifying the headword of the constituent they emphasize. For example, in the sentence *O Pedro gosta em especial/especialmente de chocolates* 'Pedro especially likes chocolates', the dependency *chocolates* > MOD > *em especial/especialmente* is considered a more appropriate representation than *gosta* > MOD > *em especial/especialmente*. An alternative representation could also place the MOD relation on the preposition introducing the prepositional phrase, but here, we consider the head of the constituent as the more fitting target for the dependency. Furthermore, the focusing nature of the modification could be made explicit in the arc label (e.g. MOD:FOCUS). This corresponds to ADVMOD:EMPH in UD. This is also the solution proposed in one of the parsers used in this paper.

Next, we outline the experiments performed to assess the lexical coverage and parsing strategies of two parsers using a set of straightforward sentences extracted from a sizable corpus.

3 Method

3.1 Syntactic Lexicon of Compound Adverbs

Our current investigation consists of building a lexicon of multi-word (or compound) adverbs in Portuguese. The primary objective is to identify, classify, and describe compound adverbs in Brazilian Portuguese, based on their lexical-syntactic properties, following a similar study conducted by Palma (2009) for European Portuguese, and revisited by Català et al. (2020). We adopt the Lexicon-Grammar theoretical framework proposed by Gross (1986). Furthermore, we have adopted the adverbial syntactic-semantic classification proposed by Molinier and Levrier (2000) for the French, derived adverbs ending in *-ment* as the base for the linguistic description of the compound adverbs in Portuguese (Table 1). So far, approximately 3,500 adverbial expressions have been collected and described. Many of these expressions are common to both Brazilian (PT-BP) and European Portuguese (PT-PT), while some are specific to each variety.

To determine which expressions would be relevant to our study, we established the following (non mutually exclusive) criteria (Müller et al., 2023).

(1) We focus mainly on *idiomatic adverbial* constructions, that is semantically non-compositional adverbial idioms (e.g. ^{PB}*Pedro fez isso com um pé nas costas* 'Pedro did it with one foot on his back'). Such expressions exhibit a certain degree of formal internal fixedness, presenting restrictions on (i) the permutation of coordinated elements; (ii) the gender and/or number variation of its elements; (iii) the substitution of its elements with synonyms or antonyms; (iv) the insertion of free determiners or modifiers; (v) the deletion of some of the elements. For lack of space, examples can be gleaned from Müller et al. (2023).

(2) We also include *multiword adverbial* constructions morphologically, syntactically and semantically (i.e. transformationally) equivalent to a single word adverb, e.g. *geralmente* 'generally', *em geral* 'in general'.

(3) While certain adverbial constructions that allow some degree of variation in their components e.g. *a certa altura* 'at a certain point', which allows variation of the demonstrative pronouns a *esta / essa / aquela* altura 'at this / that / that point'; as these variations are, in most cases, grammatical and predictable, they are represented as *local grammars*, and only one entry is registered in the main lexicon.

(4) Some *idiomatic temporal expressions*, e.g. *no tempo das vacas gordas/magras* lit. 'in the time of the fat/thin cows' 'in the good/bad times', were included, while most temporal-denoting named entities (Maurício, 2011) were ignored.

(5) *Idiomatic fixed comparative* constructions that are unique to the Brazilian variety *como o diabo gosta* lit.'like the devil likes' 'well', since other comparative frozen constructions have already been described by Ranchhod (1991).

On the other hand, the study excludes (1) *prepositional* and *conjunctional* constructions. Some of these expressions may have adverbial value; however, they select (distributionally) free elements/complements, e.g. *ao som de_as ondas/a viola/o mar/o vento* 'to the sound of the waves/the

Class	Example	EP	BP	EP-BP	Total	%
PC (conjunctive)	*afinal de contas* 'after all'	15	93	122	230	0.07%
PS (disjunctive of style)	*com toda a franqueza* 'in all honesty'	4	27	27	58	0.02%
PA (disjunctive of attitude)	*em geral* 'in general'	2	28	35	65	0.02%
MV (manner)	*por amor à pátria* 'for love of country'	274	963	927	2164	0.62%
MS (subject-oriented manner)	*de boa fé* 'in good faith'	9	36	63	108	0.03%
MT (time)	*ao romper do dia* 'at the break of day'	55	206	251	512	0.15%
MP (point of view)	*na prática* 'in practice'	0	2	2	4	0.00%
MQ (quantitative)	*aos montes* 'in abundance'	13	90	66	169	0.05%
MF (focalizer)	*em especial* 'especially'	1	7	11	19	0.01%
ML (locative)	*nos confins do mundo* 'at the ends of the earth'	7	102	72	181	0.05%
		382	1.556	1.576	3.510	

Tabela 1: Current distribution of syntactic-semantic classes in the lexicon-grammar of Portuguese adverbs.

guitar/the sea/the wind'. (2) *adverbial* constructions associated with predicative nouns and the support verb *estar* 'to be', [estar] *com a corda no pescoço* 'with the rope around one's neck'. Most of these constructions and expressions were previously studied by Ranchhod (1990).

It is crucial to clarify that the multi-word adverb lexicon is part of our ongoing research and is still in development and, therefore, subject to further refinement and expansion. Distribution of the dataset through an appropriate repository or platform is envisaged, to ensure that it is accessible to researchers interested in this area.

The current distribution of the syntactic-semantic classes in the lexicon-grammar of Portuguese is shown in Table 1. In this Table, Px classes (top tier) correspond to sentence-external adverb modifiers, while Mx classes (bottom tier) are sentence-internal adverbs. The further subclassification of these classes was omitted here. Beyond the classification proposed by Molinier and Levrier (2000), we introduced an additional category: the *locatives* (ML). While locatives are not novel in the realm of adverbial descriptions, they were notably absent from the authors' syntactic-semantic classification scheme.

For this study, we selected a sample of approximately 300 multi-word adverbs, previously assembled for another study (Muller et al., 2023), consisting in the most frequently expressions occurring in two corpora: (i) the CETEMPúblico corpus (Rocha and Santos, 2000)[1]; and the Corpus Brasileiro (Sardinha, 2010)[2]. These include both ambiguous adverbs, e.g., *com certeza* (PA) 'for sure', e.g. *Com certeza, o Pedro não foi à festa* 'Pedro certainly didn't go to the party'. and non-ambiguous adverbs, e.g. *no entanto* (PC), 'however'. The selection also encompasses sentence-internal modifiers, e.g., *à beça* (MQ) 'a lot', *às três pancadas* (MV) lit. 'with three strokes' 'recklessly', as well as sentence-external modifiers, e.g., *a mais das vezes* (PA) 'most of the times'[3].

It's worth noting that depending on the Portuguese variety (Brazilian or European), some adverbs may even exhibit ambiguity concerning their internal or external scope, as can be seen with *de repente* (MV) 'suddenly' (common in both PT-PT and PT-BR) or 'eventually' (PA) (only in PT-BR).

The frequency of these adverbs was then determined in the very large-sized Portuguese PtTenTen 2020 corpus (Kilgarriff et al., 2014; Wagner Filho et al., 2018), accessible via the Sketch Engine platform, and separately for each variety of the language in the corresponding partition of the corpus.

A very high Pearson correlation ($\rho = 0.978$) was found to exist between the frequency of these expressions in the corpora CETEMPúblico and Corpus Brasileiro corpora. A similarly high value ($\rho = 0.967$) was found when comparing the frequency of these expressions in the two partitions of the Portuguese TenTen 2020 corpus. Finally, when comparing the frequency of these expressions in the smaller corpora (CETEMPúblico and Corpus Brasileiro) with the corresponding frequency on each variety partition in the larger Portuguese TenTen 2020 corpus, similar and very high Pearson correlation values were found ($\rho = 0.974$ for the Portuguese corpora, and $\rho = 0.974$ for the Brazilian corpora). These correlation values indicate that the distribution of the expressions was similar not

[1] https://www.linguateca.pt/CETEMPublico/

[2] https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS

[3] The codes inside brackets indicate the syntactic-semantic class of the adverb. Refer to Table 1 for an overview.

only across language varieties but also across the different-sized corpora.

From the concordances of the multi-word adverbs, the Good Dictionary Examples extraction tool (GDEX) (Kilgarriff et al., 2008)[4] was then used to select the highest-ranking examples according to the sorting criteria of that tool. For example, for the adverb *à beça* (PT-BR, MQ) 'a lot', one finds the sentence: *Aquilo estava me divertindo à beça e eu não queria perder outras oportunidades* 'That was amusing me a lot, and I didn't want to miss other opportunities'. The examples were then edited to make them as short as possible, without affecting their overall intelligibility or the function of the adverbs in them.[5] Thus, for example, the previous sentence was shortened, by removing the second coordinated sentence, yielding: *Aquilo estava me divertindo à beça* lit.'That was amusing me a lot', 'I was having a blast'. A full stop was also inserted at the end when missing.

3.2 Parsers

This list of curated examples was then used for the experiments with the two selected Portuguese parsers. These parsers are presented below.

The LX-DepParser[6] is a model that has been trained specifically on Portuguese data, namely, the CINTIL-UDep treebank (Barreto et al., 2006)[7]. According to the parser's documentation, this treebank comprises 22,118 sentences and 250,056 word tokens.

Regarding the handling of multi-word expressions (MWE), the parser's handbook (Branco et al., 2014, p.12) appears to exclusively address proper names and 'cardinals' (i.e. cardinal numbers). The components within these expressions are connected through specific dependencies, N and CARD, respectively. However, it seems that various other multi-word expressions are also identified as MWEs. Their elements are linked by the dependency FIXED. No information about this dependency was provided in the documentation. This is illustrated in Fig. 1, corresponding to the parse of the sentence *Por enquanto, os problemas registados foram apenas pontuais* 'For now, the recorded issues have been only isolated'. In this parse tree, one finds below the sentence (center layer) the part-of-speech (PoS) of the tokenized items (e.g. DET=determiner, NOUN, VERB, ADJ=adjective, SCONJ=subordinate conjunction, and PUNCT=punctuation). Prepositions introducing phrases are marked as ADP (definition not found in the documentation). Below the words' PoS, one finds the corresponding *lemmata*. Concerning the (relevant) syntactic dependencies, the elements of the compound adverb *por enquanto* 'for now' are linked by FIXED, and another FIXED dependency links the adjective *pontuais* 'isolated' (the topmost predicative element of the sentence) to the MWE initial preposition *por* lit. 'by'. Note that another adverb, *apenas* 'only', is linked to the adjective using the ADVMOD (adverbial modifier) dependency (this dependency seems not to be explained in the documentation consulted).

To sum up, this suggests that the parser identifies adverbial MWEs but only at the dependency level, employing the labeled arc FIXED to connect their components and also to associate the topmost predicative element with the MWE expression. Other adverbs are linked by way of an ADVMOD dependency.

The STRING processing pipeline (Mamede et al., 2012)[8] uses the rule-based Xerox Incremental Parser (XIP)(Ait-Mokhtar et al., 2002) as its parsing module. The parser acts on the output of the tokenizer and lemmatizer module (LexMan)(Vicente, 2013), and benefits from the system's rich, fine-grained, and large-sized, lexical resources[9]. These include an initial lexicon of 2,100 multi-word adverbs, taken from (Palma, 2009), and subsequently updated. Crucially, however, it does not yet include the larger lexicon of 3,500 multi-word adverbs, presented in 3.1. Still, as the most commonly used adverbs are frequent in both varieties of Portuguese, they had already been integrated into the STRING lexicon, before these experiments.

First, the parser splits the sentence into *chunks*. These are elementary constituents such as NP (noun phrase), AP (adjectival phrase), ADVP (adverbial phrase), and so on. It also determines their respective heads. Fig. 2 illustrates the chunking tree for the sentence mentioned earlier. The same chunking structure is presented in linear form in the final line of the sentence's parse. Notably, the multi-word adverb *por enquanto* 'for now' is correctly

[4] https://www.sketchengine.eu/guide/gdex/

[5] The list of testing examples can be retrieved from: https://string.hlt.inesc-id.pt/wiki/Compound_Adverbs

[6] https://portulanclarin.net/workbench/lx-depparser/

[7] https://hdl.handle.net/21.11129/0000-000B-D2FE-A

[8] https://string.hlt.inesc-id.pt/

[9] https://string.hlt.inesc-id.pt/w/index.php/Dictionaries

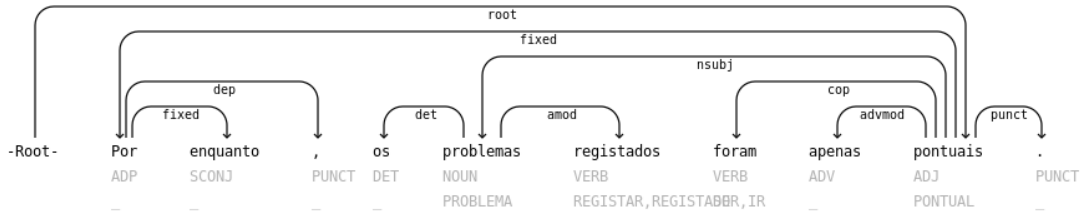
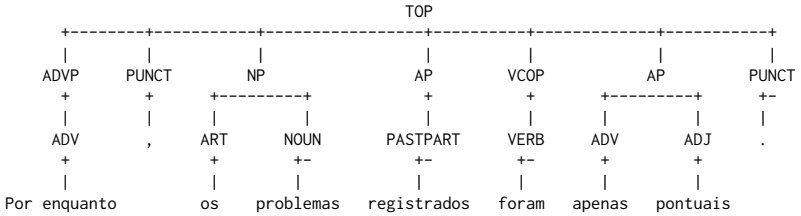


Figura 1: LX-Dep Parser: Parse tree of sentence with the compound adverb *por enquanto* 'for now'.




```
MAIN(pontuais)
VDOMAIN(foram,foram)
DETD(problemas,os)
SUBJ_PRE(foram,problemas)
PREDSUBJ(foram,pontuais)
ATTRIB(problemas,pontuais)
MOD_POST(problemas,registrados)
MOD_PRE_FOCUS(pontuais,apenas)
MOD_PRE(foram,Por enquanto)
MOD(Por enquanto , os problemas registrados foram apenas pontuais . outro,Por enquanto)
NE_T-REF-SIMULT_T-REF-ENUNC_TEMPO_T-DATE(Por enquanto)
0>TOP{ADVP{Por enquanto} , NP{os problemas} AP{registrados} VCOP{foram} AP{apenas pontuais} .}
```

Figura 2: STRING: Parse tree of sentence with the compound adverb *por enquanto* 'for now'.

identified, along with the simple focus adverb *apenas* 'only'. Both adverbs constitute the heads of their respective ADVP chunk.

Next, the parser proceeds to extract syntactic dependencies between the heads of these chunks. The extracted dependencies are listed below the parse tree. The feature _PRE on a dependency indicates that the dependent element precedes the governor, while _POST signifies the opposite linear word order.

Focusing on the relevant dependencies (MOD, modifier), we find a MOD_PRE_FOCUS relation between the adjective *pontuais* 'isolated' and the adverb *apenas* 'only', denoting the focus modifier function of this adverb on the adjective.

Furthermore, the sentence-external scope of the adverb *por enquanto* 'for now' is also represented by a MOD dependency (but without additional features). This dependency connects the entire sentence as the governor and the adverb as the dependent. The temporal aspect of this adverb is also captured through a named entity (NE) dependency, as described in (Maurício, 2011). A keen-eyed reader may have noticed an inaccurate, duplicated MOD dependency (a false positive) between this adverb and the copula verb. We will address this issue later in the discussion.

To summarize, this indicates that the parser correctly identifies adverbial multi-word adverbs right at the lexical level, and it constructs appropriate adverbial phrase (ADVP) chunks. When the adverb serves as a sentence-internal modifier, it is linked to its governor by an appropriate MOD dependency, much like any other adverbial phrase. In the case of focus adverbs, a specific _FOCUS feature is added to that dependency. When the adverb functions as a sentence-external modifier, another MOD dependency is extracted, with the entire sentence as the governor, approximating the syntactic function of this type of modifier.

4 Results

This section presents the results of the parsing experiments using the two parsers presented above to parse the testing sentences described in 3.1.

To ensure clarity when evaluating the two parsers, we establish the following criteria:
adverb: This criterion signifies that the parser has successfully recognized the given string as a multi-word adverb. In the case of the Lx-DepParser,

	Lx-DepParser			STRING		
Result	adverb	label	target	adverb	label	target
correct	18	32	140	208	234	186
incorrect	280	266	158	89	71	77
accuracy	6%	11%	47%	70%	77%	71%

Tabela 2: Results. Comparison between 2 parsers: multi-word adverb identification, dependency label and target node.

this corresponds to the extraction of a sequence of FIXED dependencies that connect all the elements of the expression, as illustrated in Fig. 1. For the STRING parser, the entire multi-word adverb forms an ADV node, which serves as the head of an ADVP chunk, as demonstrated in Fig. 2. In the case of temporal named expressions (TIMEX), instead of a multi-word adverb, a named entity is extracted (Maurício, 2011).

label: This parameter indicates that an appropriate label has been assigned to the arc linking the adverb (or the head of the adverbial phrase) and its governor. For the Lx-DepParser, this is a FIXED dependency to the ADP node. For the STRING parser, this is a MOD dependency linked to the head of the ADVP chunk.

target: This criterion confirms that the dependency accurately connects the adverb to the designated governor node. In the case of the Lx-DepParser, this corresponds to the extraction of FIXED, an OBL, or an ADVMOD dependency between the ADP node and the main verb (irrespective of the sentence-internal/external status of the modifier). For the STRING parser, this consists of the MOD dependency linking to the main predicate, when dealing with sentence-internal modifiers, or to the root NODE, representing the entire sentence, in the case of sentence-external modifiers.

Results are presented in Table 2 and showcase the performance of each parser in the identification of multi-word adverbs and the syntactic dependencies they establish.

The LXDep parser identified 18 instances (6%) as *fixed* expressions. On the other hand, most of the remaining multi-word adverbs were parsed as a string of individual tokens, as ordinary prepositional phrases. The initial preposition is tagged as an ADP, a notation explained in the documentation as corresponding to an "adverb phrase". Often, instead of an *advmod* dependency, an *obl* (=oblique) dependency is extracted. This seems to indicate that the string of words forming the compound adverb is parsed in the same way as ordinary adverbial adjuncts.

Thus, the parser's output suggests that while the parser can recognize adverbial constructs, distinguishing between simple and compound adverbs remains a challenge to the model, probably because these multi-word frozen/idiomatic expressions have not been annotated as such in the learning corpus.

Simultaneously, 32 (11%) of the dependencies were categorized as modifiers, while 140 (47%) of the dependencies accurately established connections to the intended verbs, indicating a syntactic relationship between the elements. Furthermore, the system encountered challenges in parsing 7 sentences from the testing set. These sentences featured expressions such as: *a torto e a direito* 'left and right', *a pouco e pouco* 'little by little', *daqui a pouco* 'in a while', *de uma vez por todas* 'once and for all', *no entanto* 'however', *de maneira geral* 'in a general way', and *por um acaso* 'by chance'. We have made several experiments with this small subset of sentences, moving the adverbial expression to the front of the sentence, or inserting commas to separate it from the remaining elements of the sentence, ensuring that the overall meaning was not affected nor their authenticity. This, however, did not change the result.

The STRING parser, in turn, exhibits a contrasting performance. It successfully identified 208 (70%) compound adverbs, labeling 234 (77%) of their arcs as modifiers, including 4 out of 7 focus adverbs correctly signaled by the _FOCUS feature on the MOD dependency. Additionally, the system also demonstrated high accuracy in linking the adverb to the appropriate target verb (186 instances, 71%). However, in 39 cases, two dependencies were extracted, one targeting the main verb and another one modifying the entire sentence. This happens when the adverb is at the beginning of the sentence, usually detached by a comma. Depending on the type of adverb involved, only one

of the two analyses is correct, which corresponds either way to both having a true-positive and a false-positive.

In cases where the adverb was not detected (89 instances, constituting 30% of the total), a MOD dependency was still extracted 27 times (8.9%), and in 24 cases (7.9%), the dependency was correctly linked to the target node. The number of total false-negative results (56, 18.4%) remains significant.

For instance, consider the sentence *Ao fim e ao cabo, uma imagem pode desencadear sentidos* 'After all, an image can trigger meanings'. The multi-word adverb *ao fim e ao cabo* (PC) 'in the end/after all' was not identified, resulting in the sequence being chunked as two coordinated prepositional phrases (PP). However, no dependency was extracted from either of the PPs.

In 8 cases, instead of the adverb, the system only captured a temporal named entity (NE). These are: *a cada instante* 'at every moment', *a seu tempo* 'in due time', *ao anoitecer* 'at nightfall', *ao entardecer* 'at dusk', *no dia anterior* 'on the previous day', *no último minuto* 'in the last minute', *nos dias de hoje* 'in today's times', *num determinado momento* 'at a specific moment'. This result is tied to the approach taken by Maurício (2011), wherein the system is configured to conduct regular tokenization without forming compound words in the case of named entities (NE) denoting temporal expressions. This strategy aims to facilitate a standardized (or normalized) representation of the temporal values conveyed by these expressions.

While this approach appears suitable for expressions like *a cada instante* 'every time' or *nos dias de hoje* 'these days', the interpretations of *a seu tempo* 'on its own time' and *no último minuto* 'at the last minute' are potentially ambiguous, even if the idiomatic sense is the preferred one. On the other hand, the sequences *no dia anterior* 'on the previous day' and *num determinado momento* 'at a certain moment' are indeed compositional.

Even in cases involving temporal-denoting named entities, only 2 dependencies, with *a seu tempo* and *nos dias de hoje*, were not accurately labeled, and 5 dependencies failed to target the appropriate node. Considering these cases within Table 2, STRING's overall accuracy would exhibit a marginal 0.6% improvement.

5 Conclusion and future work

This study provides a comprehensive examination of the complexities inherent to natural language parsing when confronted with multi-word adverbs, specifically in Portuguese. The primary contribution of this research lies in the development of a computational lexicon comprising 3,500 compound (multi-word) adverbial expressions, predominantly idiomatic, and enriched with syntactic-semantic information. This information encompasses their sentence-internal/external modifying functions, coupled with diatopic information specifying their prevalent usage in either the Brazilian or the European Portuguese varieties.

Our experiments, which utilized the 300 most frequently occurring multi-word adverbs from the Portuguese TenTen2020 corpus, indicate that recognizing these adverbs as distinct lexical units early in the parsing process is essential for the effectiveness of NLP systems.

The comparison between the LXDepParser and the STRING parser provided different insights into the approaches to parsing multi-word expressions. The poorer results of the first, against the better performance of the second, seem to confirm the position statement of Savary et al. (2019), that "without lexicons, multiword expression identification will never fly".

The results, however, demonstrated that even for STRING there is still room for improvement, and we hope that our ongoing project to develop a comprehensive lexical resource of multi-word adverbs for Brazilian Portuguese, currently with 3,500 entries, may contribute to improving the processing of these adverbial expressions.

Future work involves the integration of the Brazilian Portuguese multi-word adverbial expressions into the lexicon of the STRING system; providing information on the distribution of the entries in each variety, taken from the Portuguese TenTen 2020 corpus; and revising and correcting part of the parser's rule system to improve the accuracy of the syntactic dependency extraction module.

6 Acknowledgements

Research for this paper has been partially supported by national funds from Fundação para a Ciência e a Tecnologia, under project ref. UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020). Izabela Müller has also received support from the U. Algarve, through the Language Sciences PhD program.

References

S. Ait-Mokhtar, J. Chanod, and C. Roux. 2002. Robustness beyond shallowness: incremental dependency parsing. *Natural Language Engineering*, 8(2/3):121–144.

Jorge Baptista and Dolors Català. 2009. Disambiguation of focus adverbs in Portuguese and Spanish. In *ISMTCL - International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages*, pages 31–37, Université de Franche-Comté, Besançon, France. ISMTCL.

Florbela Barreto, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Bacelar do Nascimento, Filipe Nunes, and João Ricardo Silva. 2006. Open resources and tools for the shallow processing of Portuguese: The TagShare project. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Evanildo Bechara. 2012. *Moderna gramática portuguesa*. Nova Fronteira.

António Branco, Sérgio Castro, João Silva, and Francisco Costa. 2014. CINTIL DepBank Handbook: Design options for the representation of grammatical dependencies. Technical report, University of Lisbon, Faculty of Sciences, Department of Informatics.

Dolors Català, Jorge Baptista, and Cristina Palma. 2020. Problèmes formels concernant la traduction des adverbes composés (espagnol/portugais). *Langue(s) & Parole*, 5:67–82.

Mathieu Constant, G. Eryigit, Joana Monti, L. van der Plas, Carlos Ramisch, Michael Rosner, and A. Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

João Costa. 2008. *O advérbio em português europeu*. Colibri, Lisboa.

Celso Cunha and Luís Filipe Lindley Cintra. 1986. *Nova Gramática do Português Contemporâneo*. Lisboa: Edições João Sá da Costa. (3^a ed.).

Vasiliki Foufi, Luka Nerima, and Eric Wehrli. 2017. Parsing and MWE detection: Fips at the PARSEME shared task. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 54–59.

Matilde Gonçalves, Luísa Coheur, Jorge Baptista, and Ana Mineiro. 2020. Avaliação de recursos computacionais para o português. *Linguamática*, 12(2):51–68.

Gaston Gross. 1996a. *Les expressions figées en français: noms composés et autres locutions*. Editions Ophrys.

Maurice Gross. 1986. *Grammaire transformationnelle du français: 3 - Syntaxe de l'adverbe*. ASSTRIL, Paris.

Maurice Gross. 1996b. Lexicon-Grammar. In Keith Brown and Jim Miller, editors, *Concise Encyclopedia of Syntactic Theories*, pages 244–259. Pergamon, Cambridge.

Claude Guimier. 1996. *Les adverbes du français: le cas des adverbes en* -ment. Editions Ophrys.

Sylvain Kahane, Kim Gerdes, and Marine Courtin. 2018. Multi-word annotation in syntactic treebanks: Propositions for universal dependencies. In *16th international conference on Treebanks and Linguistic Theories (TLT)*.

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlỳ. 2008. Gdex: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress*, volume 1, pages 425–432. Universitat Pompeu Fabra Barcelona.

Adam Kilgarriff, Miloš Jakubíček, Jan Pomikálek, Tony Berber Sardinha, and Pen Whitelock. 2014. PtTenTen: A Corpus for Portuguese Lexicography. *Working with Portuguese Corpora*, pages 111–30.

Diane Larsen-Freeman and Marianne Celce-Murcia. 2016. The grammar book. *Form, meaning and use for English language teachers*, 3.

Nuno Mamede, Jorge Baptista, Cláudio Diniz, and Vera Cabarrão. 2012. STRING - A Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In *Computational Processing of the Portuguese Language*, volume PROPOR 2012 Demo Session, page s/p., Coimbra, Portugal. PROPOR, PROPOR.

Andreia Maurício. 2011. Identificação, classificação e normalização de expressões temporais. Master's thesis, Universidade Técnica de Lisboa - Instituto Superior Técnico, Lisboa.

Igor Mel'čuk. 2023. *General phraseology: Theory and practice*. John Benjamins.

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

Ole Mørdrup. 1976. Sur la classification des adverbes en *-ment*. *Revue romane*, 11(2):317–333.

Izabela Muller, Jorge Baptista, and Nuno Mamede. 2023. Differentiating Brazilian and European Portuguese Multiword Adverbs. Paper presented to the 39th National Meeting of the Portuguese Linguistics Association (APL), Covilhã, Portugal, October, 2023.

Izabela Müller, Nuno Mamede, and Jorge Baptista. 2022. Bootstrapping a Lexicon of Multiword Adverbs for Brazilian Portuguese. In *International Conference on Computational and Corpus-Based Phraseology*, pages 160–174. Springer.

Izabela Müller, Nuno Mamede, and Jorge Baptista. 2023. Advérbios Compostos do Português do Brasil. *Revista da Associação Portuguesa de Linguística*, 1(10):230–250.

Eduardo Paiva Raposo. 2013. Advérbio e sintagma adverbial. In Eduardo Paiva Raposo et al., editor, *Gramática do português*, volume 2, pages 1569–1675. Fundação Calouste Gulbenkian / Academia das Ciências de Lisboa.

Cristina Palma. 2009. Estudo contrastivo português-espanhol de expressões fixas adverbiais. Master's thesis, Universidade do Algarve, Faculdade de Ciências Humanas e Sociais, Faro, Portugal.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Carlos Ramisch. 2015. *Multiword Expressions Acquisition: A Generic and Open Framework*. Springer.

Elisabete Ranchhod. 1990. *Sintaxe dos predicados nominais com* estar. Instituto Nacional de Investigação Científica (INIC), Lisboa.

Elisabete Ranchhod. 1991. Frozen adverbs – Comparative forms *Como C* in Portuguese. *Linguisticae Investigationes*, XV(1):141–170.

Paulo Alexandre Rocha and Diana Santos. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. *quot; In Maria das Graças Volpe Nunes (ed) V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)(Atibaia SP 19-22 de Novembro de 2000) São Paulo: ICMC/USP*.

Tony Berber Sardinha. 2010. Corpus brasileiro. *Informática*, 708:0–1.

Agata Savary, Silvio Ricardo Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91. Association for Computational Linguistics.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023. Parseme meets universal dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, 9(1).

Alexandre Vicente. 2013. LexMan – um Segmentador e Analisador Morfológico com Transdutores. Master's thesis, Universidade de Lisboa - Instituto Superior Técnico, Lisboa.

Jorge A Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC Corpus: a New Open Resource for Brazilian Portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.