The Role of Adverbs in Language Variety Identification: The Case of Portuguese Multi-word Adverbs

Izabela Meira Grein Müller, Jorge Baptista, Nuno Mamede

The Role of Adverbs in Language Variety Identification: The Case of Portuguese Multi-word Adverbs

Izabela Müller and Jorge Baptista
U. Algarve - FCHS
Campus de Gambelas, Faro (Portugal)
INESC-ID Lisboa – HLT
{belagrein,jorge.baptista}@inesc-id.pt

Nuno Mamede
U. Lisboa - IST, Lisbon (Portugal)
INESC-ID Lisboa – HLT
R. Alves Redol 9, Lisboa (Portugal)
Nuno.Mamede@inesc-id.pt

Abstract

This paper aims to assess the role of multi-word compound adverbs in distinguishing Brazilian Portuguese (PT-BR) from European Portuguese (PT-PT). For this study, a large lexicon of Portuguese multi-word adverbs (3,665) was annotated with diatopic information regarding language variety, which has not been available so far. The paper then investigates the distribution of this category in the DSL (Dialect and Similar Language) corpus of journalistic texts, representing Brazilian (PT-BR) and European Portuguese (PT-PT). Results indicate a substantial similarity between the two varieties, with a considerable overlap in the lexicon of multi-word adverbs. Additionally, specific adverbs unique to each language variety were identified. Lexical entries recognized in the corpus represent 18.2% (PT-BR) to 19.5% (PT-PT) of the lexicon, and approximately 5,700 matches in each partition. While many of the matches are spurious due to ambiguity with otherwise non-idiomatic, free strings, occurrences of adverbs marked as exclusive to one variety in texts from the other variety are rare.

1 Introduction

This study seeks to identify and contrast multi-word (compound) adverbs between the Brazilian (PT-BR) and European (PT-PT) varieties of Portuguese. Two key factors underpin this focus: Firstly, multi-word expressions often prove to be less ambiguous than single words, even when their meaning is idiomatic (non-compositional). Secondly, despite constituting a significant portion of lexicons in many languages, adverbs are frequently overlooked in Natural Language Processing, possibly due to their heterogeneous nature and lexical range. Furthermore, to the best of our knowledge, no assessment has been made until now, of the lexical distribution of language variety-specific multi-word adverbs in Portuguese. And even if such a distribution were skewed, no study seems to be available on the distribution in corpora of such language variety-specific multi-word adverbs. The goal of this paper is to provide a clear answer to both these issues.

While language variety identification can be a crucial task for dialect-sensitive NLP tasks, the main idea underlying this paper is *not* to evaluate the identification of the linguistic variety between PT-BR and PT-PT *per se* using multi-word adverbs, but to determine the impact of the adverbial expressions and the extent to which they are asymmetrical across both varieties. The aim is to achieve this through an analysis conducted on two comparable corpora of journalistic texts, one in PT-BR and the other in PT-PT, that have been used in previous DSL shared tasks (Tan et al., 2014).

This paper begins with an overview of the primary goals and the resources utilized. In Section 2, we look deeper into the notion of multi-word (compound) adverbs and discuss the ongoing research focused on developing a lexicon of multi-word adverbs in Brazilian Portuguese. Section 3 outlines the methodology used in this experiment, specifically addressing the asymmetry of adverbial constructions identified in both varieties of Portuguese. Section 4 presents the findings and their analysis. Lastly, Section 5 concludes the paper with final observations.

2 The Lexicon of Portuguese Multi-word Adverbs

Multi-word adverbs, also referred to as *compound adverbs*, are expressions composed of two or more words forming a single lexical unit with specific word combinations. While they generally adhere to syntactic rules for phrase formation, their structure is often "frozen", meaning their components cannot be rearranged, inserted, or reduced (through ellipsis), and they typically exhibit idiosyncratic constraints on morphosyntactic variation (Gross,

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 99–106

1982, 1986b; Guimier, 1996). One key characteristic of multi-word adverbs is their lack of semantic compositionality, wherein the meaning of the expression cannot be deduced from the individual meanings of its elements, resulting in an idiomatic overall meaning, e.g.:

(1) *Aprendi isso* a duras penas.[1] 'I learned that *the hard way*.'

The example above demonstrates that the expression *a duras penas* 'the hard way' is a compound, multi-word adverb, as it adheres to the mentioned constraints regarding:[2] (a) word order: ?/°*a penas duras*; (b) reduction: **a penas*; (c) idiosyncratic restrictions on morphosyntactic variation: °*a dura pena*; and, finally, (d) limited insertions may be acceptable: *a* muito *duras penas*. In the *PtTenTenCorpus2020* (12.5 billion words) (Kilgarriff et al., 2014), (a) only 3 instances were found of the idiomatic string *a penas duras*, with the adjective in a post-nominal position; 7 other uses of this word order are all literal ('harsh penalties'), as indicated by '°'; these 10 instances are marginal, compared to 6,205 occurrences of *a duras penas*, with reversed order, all of which are idiomatic; (b) the reduced form (over 5.5 thousand occurrences) is never interpreted as an idiomatic expression, and it is usually a sub-string of a distributionally free phrase; (c) the singular form (21 instances) is always literal; and, finally, (d) only 2 instances were found with the adverb quantifier *muito* in the prenominal position, both idiomatic; and none after the noun.

Multi-word adverbial expressions have received a great deal of interest in the field of linguistic studies and have been the object of previous studies in various languages (Gross, 1996a; Di Gioia, 2001; Català, 2003; Laporte et al., 2008; Palma, 2009; Shudo et al., 2011; Català et al., 2020; Müller et al., 2022, 2023).

The lexicon of multi-word adverbs in Brazilian Portuguese is part of an ongoing study that focuses on identifying, classifying, and describing the lexical, syntactical, and semantic features of multi-word adverbs. The goal is to expand the list of European Portuguese multi-word adverbs (Palma, 2009) by incorporating Brazilian Portuguese adverbial compounds. This will result in a comprehensive lexicon, covering expressions specific to each Portuguese variety, as well as those common to both.

We adopt the theoretical-methodological framework of Lexicon-Grammar (Gross, 1975, 1981, 1996b), based on Harris (1991) Operator Transformational Grammar, along with the formal classification of compound adverbs as proposed by Gross (1986a). This classification system categorizes adverbs based on the internal sequence of their grammatical categories. Table 1 presents the current distribution of multi-word adverbs by formal classes and language variety. Currently, 68.3% of the lexicon incorporates multi-word adverbs shared between PT-PT and PT-BR. At the same time, 26.3% includes expressions exclusively found in PT-BR, and 5.5% is exclusive to PT-PT adverbial expressions.

Additionally, we apply the criteria proposed by Molinier and Levrier (2000) to classify single adverbs ending in *-ment* in French, according to their syntactic-semantic features. We believe these criteria are applicable to the description of *-mente* adverbs in Portuguese, as shown in Fernandes (2011), as well as multi-word adverbs in Portuguese (Palma, 2009; Català et al., 2020).

Molinier and Levrier (2000)'s framework outlines two primary categories of adverbs: those that modify the constituents of the sentence, and thus are considered *internal modifiers* (type M), and those that modify the entire sentence, known as *external modifiers* (type P). The authors further propose a nuanced sub-classification based on the adverbs' function and the relations they establish within a sentence, delineating various syntactic-semantic adverbial classes.

External modifiers are subdivided into three categories: (i) conjunctive adverbs (PC), (ii) disjunctive adverbs of style (PS), and (iii) disjunctive adverbs of attitude (PA). The latter category is further subdivided into four subclasses: (a) adverbs of *habit* (PAh), (b) *evaluative* adverbs (PAe), (c) *modal* adverbs (PAm), and (d) *subject-oriented* adverbs (PAs).

Adverbs that modify an internal constituent of the sentence are classified into six subclasses: (iv) *manner* adverbs (MV), (v) *subject-oriented manner* adverbs (MS), (vi) adverbs of *time* (MT), (vii) *point-of-view* adverbs (MP), (viii) *quantity* adverbs (MQ), and (ix) *focusing* adverbs (MF).

We adopted this general framework to classify and describe Portuguese adverbs. Furthermore, we introduce a new category, (x) *locative* adverbs

[1] *O Estadão*, from *PtTenTenCorpus*, id=11740543771

[2] In the examples, '*' indicates the string is unacceptable, '?' dubiously acceptable, or '°' acceptable but with literal/compositional meaning.

Class	Internal Structure	Examples	PT-PT	%	PT-BR	%	PT	%	Total	%
PC	*Prep C*	*em vão* 'in vain'	28	0.14	243	0.25	667	0.27	938	0.256
PDETC	*Prep Det C*	*pelo menos* 'at least'	57	0.29	218	0.23	522	0.21	797	0.218
PAC	*Prep Adj C*	*de má vontade* 'unwillingly'	11	0.06	46	0.05	231	0.09	288	0.079
PCA	*Prep C Adj*	*por maioria absoluta* 'by absolute majority'	22	0.11	70	0.07	268	0.11	360	0.098
PCDC	*Prep C1 de C2*	*por conta da casa* 'on the house'	21	0.11	83	0.09	207	0.08	311	0.085
PCPC	*Prep C1 Prep C2*	*da cabeça aos pés* 'head to toes'	46	0.23	105	0.11	265	0.11	416	0.114
PCONJ	*Prep C1 Conj C2*	*em verso e prosa* 'in verse and prose'	9	0.05	74	0.08	168	0.07	251	0.068
PF	frozen sub-clause	*dito isso* 'this said'	2	0.01	41	0.04	88	0.04	131	0.036
PV	*Prep V W*	*até dizer chega* lit.:'until say enough', 'a lot'	2	0.01	2	0.002	25	0.01	29	0.008
PJC	*Conj C*	*e por aí vai* 'and so on'	2	0.01	47	0.05	31	0.01	80	0.022
PACO	*<Adj>como C*	*<surdo>como uma porta* 'deaf as a door'			7	0.01	3	0.001	10	0.003
PVCO	*<V>como C*	*<trabalhar>como uma mula* 'word like a mule'			26	0.03	25	0.01	51	0.014
		Total	200	0.055	962	0.262	2,500	0.683	3,662	

Table 1: Formal classification of Portuguese multi-word adverbs. Codes for classes are conventional. Internal structure: adjective *Adj*, *C1* and *C2* lexical constants, conjunction *Conj*, Determiner *Det*, Preposition *Prep*, Verb *V*, undefined sequence of elements *W*. Distribution per variety: European Portuguese *PT-PT*, Brazilian Portuguese *PT-BR*, Common Portuguese *PT*. Zero values were removed.

(ML), which was not included in this framework before, even though it is not new to the study of adverbs. You can find more details about each category in (Müller et al., 2022, 2023).

Table 2 displays the distribution of the lexicon based on this syntactic-semantic classification across different language varieties. To the best of our knowledge, this lexicon represents the most extensive collection of multi-word adverbs available in Portuguese.

The predominant categories are *manner* adverbs (MV: 59.9%) and *time* adverbs (MT: 14.8%). Within the latter category, 76% are corresponded to *date* adverbs, indicating temporal locatives. Additionally, the recently introduced locative class (ML) from (Müller et al., 2023) accounts for 5.5%. Conjunctive adverbs (PC: 7%) and quantifying adverbs (MQ: 5.1%) are also noteworthy.

The assignment of language variety to the multi-word adverbs in the lexicon is based mostly on their distribution in the corpora, particularly (i) for PT-PT, the CETEMPúblico corpus (Rocha and Santos, 2000)[3] (ii) for PT-BR, the Corpus Brasileiro[4], with approximately 1 billion words; both (i) and (ii) are available through Linguateca[5]; and, (iii) for both variants, the Portuguese Web 2020 (ptTenTen-Corpus20) (Wagner Filho et al., 2018; Kilgarriff et al., 2004), with 12,5 billion words (PT-PT: 893.2 million words, PT-BR: 8 billion words).

3 Methods

In order to assess the influence of multi-word adverbs on the two Portuguese varieties, we employed the European (PT-PT) and Brazilian (PT-BR) partitions of the Discrimination of Similar Languages (DSL) Corpus Collection (DSLCC, v.04) (Tan et al., 2014)[6]. These partitions were originally curated for the DSL task and served as the primary dataset for the shared tasks conducted as part of the NLP for Similar languages, Varieties and Dialects (VarDial) workshop (Zampieri et al., 2017). The PT-PT texts comprise 18,000 sentences with a total of 735,503 words, while the PT-BR texts also encompass 18,000 sentences and a slightly larger word count of 791,872. Table 3 shows the breakdown of the number of sentences, words and different words in each partition.

To process the corpora, we utilized the linguistic development platform Unitex (v.3.3)(Paumier et al., 2021).[7] The texts underwent pre-processing using the linguistic resources provided by the system, specifically the text segmentation tool and the simple-word dictionary. The lexicon of multi-word adverbs was also formatted into the DELA format compatible with Unitex and then applied to the corpora. For instance, consider the entry for the manner adverb *a duras penas* 'the hard way':

```
a duras penas.ADV+PAC+MV+PT+BR
```

In this format, each adverb entry consists of a string with a part-of-speech designation and a set of features, including its formal class, syntactic-semantic class, and the language varieties it pertains to (+PT and/or +BR). Adverbs not specific to a language variety are explicitly labeled with the features +NotPT (e.g., [*responder*] *de bate pronto* '(to reply) right away') or +NotBR (e.g., [*cair*] *de ratatulha* '(to fall) headlong').

[3] www.linguateca.pt/cetempublico
[4] http://corpusbrasileiro.pucsp.br/
[5] www.linguateca.pt/
[6] http://ttg.uni-saarland.de/resources/DSLCC/
[7] https://unitexgramlab.org/

Class	Examples	PT-PT	%	PT-BR	%	PT	%	Total	%
PC (conjunctive)	*afinal de contas 'after all'*	3	0.015	42	0.043	213	0.085	258	0.070
PS (disjunctive of style)	*com toda a franqueza 'in all honesty'*	1	0.005	10	0.010	54	0.022	65	0.018
PA (disjunctive of attitude)									
PAa (evaluative)	*por pura sorte 'by sheer luck'*					21	0.008	21	0.006
PAm (modal)	*com certeza 'certainly'*	1	0.005	10	0.010	25	0.010	36	0.010
PAs (subject-oriented)	*pelo meu lado 'for my part'*			4	0.004			4	0.001
PAh (habit)	*de costume 'usually'*			4	0.004	12	0.005	16	0.004
MV (manner)	*por amor à pátria 'for love of country'*	157	0.781	615	0.637	1,423	0.570	2,195	0.599
MS (subject-oriented mode)	*de boa fé 'in good faith'*	2	0.010	22	0.023	93	0.037	117	0.032
MT (time)									
MTd (date)	*a horas mortas 'at dead of night'*	24	0.119	80	0.083	307	0.123	411	0.112
MTf (frequency)	*dia sim dia não 'every other day'*	5	0.025	25	0.026	53	0.021	83	0.023
MTu (duration)	*anos a fio 'for years on end'*			12	0.012	35	0.014	47	0.013
MP (point of view)	*na prática 'in practice'*					4	0.002	4	0.001
MQ (quantifier)	*aos montes 'in abundance'*	5	0.025	60	0.063	119	0.048	185	0.051
MF (focalizer)	*em especial 'especially'*			3	0.003	17	0.007	20	0.005
ML (locative)	*nos confins do mundo 'at the ends of the earth'*	3	0.015	78	0.081	120	0.048	201	0.055
	Total	201	0.055	966	0.264	2,496	0.681	3,662	

Table 2: Syntactic-semantic classification of Portuguese multi-word adverbs. Codes for classes are conventional. Sub-classes of PA and MT are presented. Distribution per variety: European Portuguese *PT-PT*, Brazilian Portuguese *PT-BR*, Common Portuguese *PT*. Zero values were removed.

```
de bate pronto.ADV+PCA+MV+NotPT+BR
de ratatulha.ADV+PAC+MV+PT+NotBR
```

For the classification of adverbs according to the language variety, two linguists, native speakers of each variety, manually, separately and systematically annotated the lexicon entries, deciding whether they belonged to each other variety. Additionally, we also relied on corpus consultation *PtTenTen20* partitions of each language variety and controlled web search using domains **.pt** and **.br** to verify the occurrence of the adverbs in each variety. In a second moment, aspects of lexical variation (prepositions, determiners) were checked. Foremost, in the case of adverbs signaled to be common to both varieties, false-friends were detected by the authors, by elicitation of the meaning of those expressions. To this end, we also resource to these adverbs' use in real examples drawn from corpora, when the meaning was not clear or was apparently different from the expected meaning in one of the varieties – e.g. *toda vida* 'all life' as a locative (ML) adverb in PT-BR and not as a durative time adverb (MTd); or todo (o) dia 'all day' as a durative MTd in PT-PT instead of a frequency MTf adverb in PT-BR. As seen in these examples, it is often only at the syntactic-semantic classification that such differences arise.

(2) *É só chegar no hotel e seguir reto* toda a vida 'Just get to the hotel and go straight on ahead/'til the end(lit:. all [your] life) '

This approach allowed us to extract all instances of matched adverbs, particularly those with the `+NotBR` feature from the PT-BR partition of the corpus, and conversely, all adverbs marked as `+NotPT` from the PT-PT partition. In the following section, we present and discuss our findings.

4 Results

	DSLCC corpus	
	PT-PT	PT-BR
Sentences	18,000	18,000
Words	735,503	791,872
Different words	42,190	47,914
Adv lexical entries	715	668
PT-BR entries	629 (87.9%)	620 (92.8%)
NotPT entries	74 (10.3%)	46 (6.9%)
NotBR entries	12 (1.7%)	2 (0.3%)
Adv matches	5,695	5,700
NotPT/BR matches	517	2

Table 3: DLSCC Corpus: European (PT-PT) and Brazilian Portuguese (PT-BR) partitions. Results from lexical analysis.

From applying the lexicon of multi-word adverbs to each partition of the DSLCC corpus, the following results emerged, as depicted in Table 3. Although the word count in the PT-BR partition is marginally higher (+7.66%), the number of distinct lexical entries is slightly smaller (-7.04%).

Considering the Brazilian Portuguese (PT-BR) partition, the number of lexical entries found in the corpus (668) represents 18.23% of entries of the multi-word adverbs lexicon. These can be divided into exclusively Brazilian entries (46; 6.9%), exclusively European (2; 0.3%) and entries common to both varieties (620; 92.1%).

Moving now to the European Portuguese par-

tition, the number of lexical entries found in the corpus comprises 715 adverbial entries, comparable to the size of the PT-BR corpus. Among these entries, 629 (87.9%) are common to both Brazilian and European Portuguese, while 74 (10.3%) are exclusive to European Portuguese, and 12 (1.7%) are not found in Brazilian Portuguese.

This breakdown illustrates the substantial lexical overlap between the multi-word adverbs of the two varieties. European Portuguese contains a slightly higher proportion of unique adverbs than Brazilian Portuguese. This overlap tends to make the use of adverbs a less-than-optimal linguistic device for the DSL task. In fact, as it will be seen from the observations made below, this overlap is even greater, as some entries, marked as exclusive from one variant (`+NotPT`), do occur in the PT-PT partition.

	ptTenTen2020 Corpus partition	
adverb	**PT-PT**	**PT-BR**
ao domicílio	254 284 377 41*10^{-6}	4 4 478 384*10^{-6}
a domicílio	7 873 842*10^{-6}	2,597 324 195 295*10^{-6}
	n = 893 179 245	*n* = 8 010 603 604

Table 4: Distribution of the multi-word adverbs *ao domicílio*/*a domicílio* 'to the domicile' in combination with the verbs *entregar* and *distribuir* 'deliver' in the ptTenTen2020 corpus; number of occurrences and ratio per million words; *n* is the number of tokens per each partition of the corpus.

The search in the BR corpus for entries with the `+NotBR` tag resulted in only 2 cases, which are illustrative of the phenomena found. Table 4 shows the distribution of the locative adverb *ao domicílio* / *a domicílio* 'to (the) domicile' in each partition of the ptTenTen2020 corpus in combination with the most frequently co-occurring verbs *entregar* and *distribuir* 'deliver', allowing for a 0 to 5-word window, in the ptTenTen2020 corpus.

From the data in this table, the expression *a_o domicílio* 'to_the domicile' (with the article *o* 'the') is deemed as predominantly used in PT-PT. In fact, in PT-BR, the corresponding expression is *a domicílio* 'to domicile', which lacks the article *o* 'the'. The single, spurious occurrence of this adverb constitutes a case of ambiguity:

(3) *Em relação à filiação partidária e* ao domicílio *eleitoral, a comissão manteve a legislação atual.* 'Regarding party affiliation and electoral domicile, the commission maintained the current legislation.´

The second case was *de facto*, 'in fact', which is the PT-PT orthographic form, while in PT-BR the correct spelling is *de fato*. The distribution of the two spellings in the same corpus, when the string is followed by a comma (usually a non-ambiguous context of the multi-word adverb), is shown in Table 5. This single occurrence suggests a spelling error. However, its analysis reveals another level of ambiguity:

(4) [. . .] *O governo de facto,* [. . .] *rechaça a volta do líder deposto ao poder.* 'The de facto government, [. . .], rejects the ousted leader's return to power.'

In this case, the *de facto*/*de fato* adverb is being used here as an adjectival modifier of *governo* 'government', and its meaning 'de facto', as shown in the translation, is that of a manner-like modifier. This is a clear contrast with the modal (PAm) value 'in fact', typically associated with the adverb.

	ptTenTen2020 Corpus partition	
adverb	**PT-PT**	**PT-BR**
de facto	50,270 56 282 095 99*10^{-6}	3,644 454 897 056*10^{-6}
de fato	1,878 2 102 601 477*10^{-6}	252,109 31 471 910 54*10^{-6}
	n = 893 179 245	n = 8 010 603 604

Table 5: Distribution of the multi-word adverbs *de facto* / *de fato* 'in fact'.

Besides, the distribution of the spellings shows that the distinction between the two varieties is often not a clear-cut divide. In this particular case, the adaptation to the orthographic reform[8] may have raised some level of uncertainty among language users.

The number of instances of `+NotPT` adverbs found in the PT-PT partition of the DSLCC corpus is significantly higher (517). For lack of space, only a few different cases will be mentioned here, to illustrate the general phenomena found.

Some cases correspond to real distinct expressions in each variety. For example, the adverb (*pagar*) *às prestações* [PT-PT]/ *à prestação* [PT-BR]/ 'in installments' is used with the plural form in PT-PT and in the singular PT-BR. All 4 matches

[8] http://www.portaldalinguaportuguesa.org/acordo.php

of the Brazilian adverb *à prestação* are spurious, and correspond to free prepositional phrases:

(5) [. . .] *aplicado* à prestação *de contas* [. . .]; *O acesso* à prestação *exige a assinatura prévia* [. . .]; [. . .] *quanto* à prestação *dos cuidados assistenciais* [. . .]; [. . .] *estar atento* à prestação *dos jogadores* [. . .]
'[. . .] applied to the rendering of accounts [. . .]; Access to the service requires a prior subscription [. . .]; [. . .] regarding the provision of assistance care [. . .]; [. . .] pay attention to the performance of the players [. . .]'

Another problem results from many compound adverbs being very short strings, and therefore highly ambiguous with other word combinations, including other multi-word expressions. Examples of such ambiguous, `+NotPT` strings are *à toda* 'full speed', *às avançadas* 'to the advanced', *de primeira* 'firstly', *na maior* 'confortably', *por cima* 'above', *por detrás* 'from behind'. Finally, several expressions have been marked as `+NotPT` but are, in fact, common to both varieties, e.g. *de há muito* [*tempo*] 'a long [time] ago'.

After a manual inspection, it was ascertained that out of 517 matches, (i) 112 (21.6%) were true-positives, that is, the multi-word adverbs were found in the PT-PT corpus though marked as `+NotPT`, hence the assignment of those expressions to a single variety needs careful revision; (ii) 405 matches (78.3%) were false-positives, that is, the matched string did not correspond to the multi-word adverb in the lexicon applied to the corpus. From these, however, 110 instances (21.5%) made part of longer multi-word expressions:

(6) *O IVA "super-reduzido", dos bens* de primeira *necessidade, irá permanecer em 4%.*
'The "super-reduced" VAT on basic necessities will remain at 4%.'

In these case, the compound noun *bens de primeira necessidade* 'basic necessities' overlaps the compound adverb *de primeira* [PT-BR] 'to start with'. recognizing the longer multi-word expression would have prevented these false-positive cases.

It should also be mentioned that many instances identified in both partitions of the corpus and signaled as belonging to the common Portuguese (`+PT+BR`) are, in fact, also spurious (false-positive) cases, for the same reasons as explained above. That is, the system identifies a sequence of words that resembles a dictionary-listed expression, but that does not align with the intended compound adverb. This discrepancy highlights the potential for ambiguity inherent in NLP processing, and requires deeper linguistic analysis of the ambiguous strings' context to improve precision.

Both partitions of the corpus are currently being annotated to delimit the targeted multi-word adverbial forms and tag them with their POS, formal and semantic class, as well as the language variety assignment. The goal is to build a reference corpus annotated for this category, aiming at improving parsing accuracy[9].

5 Conclusion

This paper introduces a lexicon of multi-word (MW, or compound) adverbs in Portuguese, examining their lexical distribution across Brazilian (PT-BR) and European (PT-PT) varieties. From a strictly lexical perspective, the majority of the lexicon pertains to Common Portuguese (68.1%), with exclusively Brazilian compound adverbs (26.4%) outnumbering those exclusive to the European variety (5.5%). However, these preliminary figures may require revision following the experiments conducted in this study.

This lexicon was utilized to annotate the European (PT-PT) and Brazilian (PT-BR) segments of a comparable corpus sourced from the Discrimination of Similar Languages (DSL) Corpus Collection (DSLCC, v.04) (Tan et al., 2014). The count of distinct adverb entries discovered in the corpus (PT-PT: 715 / PT-BR: 668), as well as the number of matches (PT-PT: 5,695 / PT-BR: 5,700), exhibits remarkable similarity.

The proportion of lexical entries attributed to Common Portuguese is notably high and comparable across both corpus partitions (PT-PT: 629 (87.9%) / PT-BR: 620 (92.8%)), although slightly larger in PT-BR. Conversely, the count of lexical entries exclusively associated with each variety in their respective partitions is relatively small (PT-PT: 74 (10.3%) / PT-BR: 46 (6.9%)), with a slightly higher proportion observed for European Portuguese entries.

On the contrary, the number of MW adverbs labelled as *not* belonging to either variety (`+NotPT` and `+NotBR`) and found within their respective partitions is arguably negligible (PT-PT: 12 (1.7%) / PT-BR: 2 (0.3%)), albeit marginally higher in PT-PT.

[9]The corpus of annotated sentences, and the list of matched MW adverbs' can be found in the link below, under a Creative Commons license: https://string.hlt.inesc-id.pt/wiki/Compound_Adverbs

Nevertheless, the frequency of such occurrences in each partition exhibits significant asymmetry.

In Brazilian texts, only two instances of `+NotBR` MW adverbs were identified. One of them (e.g., *ao domicílio* / *a domicílio* 'to (the) domicile') presents a case of ambiguity, as the phrase forming the MW adverb can also exist as a free sequence of words. The other instance is a misspelled word (*facto* / *fato* 'fact'), likely resulting from some uncertainty in applying the orthographic reform of the Portuguese language. By consulting the extensive corpus of *PtTenTen2020* (Wagner Filho et al., 2018; Kilgarriff et al., 2004), it was possible to determine: (i) the asymmetric distribution of each variant form and their true association with the PT-PT or PT-BR partition; and (ii) the clear-cut distribution of orthographic variants, alongside some ambiguity due to the imperfect application of the Portuguese orthographic reform.

Regarding the `+NotPT` adverbs found in the Portuguese partition of the corpus, surprisingly, a considerable portion (21.6%) were confirmed as true-positive instances of adverbs inaccurately marked as exclusive to the Brazilian variety, necessitating reassignment to European Portuguese. However, the majority of remaining instances (78.3%) were false positives, stemming from the ambiguity of the strings forming the multi-word adverb with other word combinations. Among these, 21.5% were even part of another multi-word expression (such as compound nouns or verbal idioms). Hence, there remains ample room for improvement in accurately identifying multi-word adverbs, particularly concerning their potential overlap with other, longer multi-word expressions, either previously identified or concurrently present.

In conclusion, multi-word adverbs in Common Portuguese constitute a significant portion of this lexical class (68%), representing the majority of all adverb entries discovered in comparable corpora (ranging from 87.9% to 92.8%). However, their sparse distribution in the corpus renders this segment of the language lexicon sub-optimal for the task of distinguishing dialects and similar languages.

In the near future, we aim to provide the two corpus partitions annotated with the newly identified multi-word adverbs. We believe that such a resource could then be utilized to enhance other dialect-sensitive natural language processing tasks.

References

Dolors Català. 2003. *Les adverbs composés: approches contrastives en linguistique appliquée.* Ph.D. thesis, Universitat Autònoma de Barcelona, Barcelona, Spain.

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

M. Di Gioia. 2001. *Avverbi idiomatici dell'italiano. Analisi lessico-grammaticale*. l'Harmattan Italia, Torino.

Gaia Fernandes. 2011. Automatic Disambiguation of *-mente* ending Adverbs in Brazilian Portuguese. Master's thesis, Universidade do Algarve and Universitat Autònoma de Barcelona, Faculdade de Ciências Humanas e Sociais, Faro, Portugal.

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

Maurice Gross. 1975. *Méthodes en syntaxe*. Hermann, Paris.

Maurice Gross. 1981. Les bases empiriques de la notion de prédicat sémantique. *Languages*, 1(63):7–52.

Maurice Gross. 1982. Une classification des phrases figées du français. *Revue québécoise de linguistique*, 11(2):151–185.

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

Maurice Gross. 1986b. Lexicon-grammar. The representation of compound words. In *COLING 1986 Volume 1: The 11th International Conference on Computational Linguistics*.

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. *French adverbs: the case of en* -ment *adverbs*. Editions Ophrys.

Zellig Sabbettai Harris. 1991. *Theory of Language and Information: a Mathematical Approach*. Clarendon Press, Oxford.

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.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, and David Tugwell. 2004. The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*, pages 105–116.

Éric Laporte, Takuya Nakamura, and Stavroula Voyatzi. 2008. A French corpus annotated for multiword expressions with adverbial function. In *Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop*, pages 48–51.

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

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.

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.

Sébastien Paumier, Wolfgang Flury, Franz Guenthner, Eric Laporte, Friederike Malchok, Clemens Marschner, Claude Martineau, Cristian Martínez, Denis Maurel, Sebastian Nagel, et al. 2021. *UNITEX 3.3 User Manual*. Université de Paris Est /Institut Gaspard Monge.

Paulo A. Rocha and Diana Santos. 2000. CETEM-Público: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In *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*. ICMC/USP.

Kosho Shudo, Akira Kurahone, and Toshifumi Tanabe. 2011. A comprehensive dictionary of multiword expressions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 161–170.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15, Reykjavik, Iceland.

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 11th International Conference on Language Resources and Evaluation (LREC 2018).*, pages 4339–4344.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the vardial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.