Proceedings Paper:
Ramisch, C., Ramisch, R., Zilio, L. et al. (2018) A corpus study of verbal multiword expressions in Brazilian Portuguese. In: Villavicencio, A., Moreira, V., Abad, A., Caseli, H., Gamallo, P., Ramisch, C., Oliveira, H.G. and Paetzold, G.H., (eds.) PROFOR2018 : Computational Processing of the Portuguese Language. PROPOR: International Conference on Computational Processing of the Portuguese Language, 24-26 Sep 2018, Canela, Brazil. Springer International Publishing, pp. 24-34. ISBN: 9783319997216. ISSN: 0302-9743. EISSN: 1611-3349.

https://doi.org/10.1007/978-3-319-99722-3_3

This is a pre-print of an article published in PROPOR 2018.
The final authenticated version is available online at: https://www.springer.com/us/book/9783319997216

A Corpus Study of Verbal Multiword Expressions in Brazilian Portuguese

Carlos Ramisch[1], Renata Ramisch[2], Leonardo Zilio[3],
Aline Villavicencio[4], Silvio Cordeiro[1]

[1] Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France first.last@lis-lab.fr
[2] Interinstitutional Center for Computational Linguistics, São Carlos, Brazil
renata.ramisch@gmail.com
[3] Université catholique de Louvain, Belgium
leonardo.zilio@uclouvain.be
[4] University of Essex, United Kingdom
alinev@gmail.com

Abstract. Verbal multiword expressions (VMWEs) such as *to* ***make ends meet*** require special attention in NLP and linguistic research, and annotated corpora are valuable resources for studying them. Corpora annotated with VMWEs in several languages, including Brazilian Portuguese, were made freely available in the PARSEME shared task. The goal of this paper is to describe and analyze this corpus in terms of the characteristics of annotated VMWEs in Brazilian Portuguese. First, we summarize and exemplify the criteria used to annotate VMWEs. Then, we analyze their frequency, average length, discontinuities and variability. We further discuss challenging constructions and borderline cases. We believe that this analysis can improve the annotated corpus and its results can be used to develop systems for automatic VMWE identification.

Keywords: multiword expressions · annotation · corpus linguistics

1 Introduction

Multiword expressions (MWEs) are groups of words presenting idiosyncratic characteristics at some level of linguistic processing [1]. Some MWEs function as verb phrases, and are thus referred to as verbal MWEs (VMWEs). Examples in Brazilian Portuguese (PT-BR) include verbal idioms (e.g. ***fazer das tripas coração*** 'make.INF of-the.FEM.PL tripes heart' ⇒ 'to do everything possible'), light-verb constructions (e.g. ***tomar*** *um* ***banho*** 'take.INF a shower') and inherently reflexive verbs (e.g. ***queixar-se*** 'complain.INF-self.3' ⇒ 'to complain').

VMWEs have been the focus of much attention, both in linguistics and in natural language processing [11,1,3,15]. From a linguistic point of view, they present restricted variability patterns, licensing phenomena such as passivization, pronominalization of components, reordering, and free PP-movement depending on the VMWE category [14,10,17,8]. Moreover, verbs (and VMWEs) tend to have rich morphological inflection paradigms, and allow many (but not all) syntactic changes [6,5]. These are often unpredictable [11], making VMWEs challenging to represent in resources and to model in applications.

For the automatic identification of VMWEs, their variability and their potential for discontinuous realizations make them hard to model, especially when put together with non-compositionality and ambiguity [3]. Indeed, VMWEs were the focus of initiatives like the PARSEME shared task[1] [15], whose goal is to foster the

[1] Editions 1.0 (2017) and 1.1 (2018): http://multiword.sourceforge.net/sharedtask2018

This is a pre-print of an article published in PROPOR 2018.
The final authenticated version is available online at: https://www.springer.com/us/book/9783319997216

development and evaluation of computational tools for VMWE identification. A by-product of this shared task was the release of freely available VMWE-annotated corpora in several languages, including PT-BR.

The goal of this paper is to study the characteristics of VMWEs in PT-BR using the PARSEME corpus. We describe their annotation and analyze their diversity and distribution. A deeper understanding of this complex phenomenon can inspire linguistic models and boost the development of systems to identify them automatically. In §2 and §3 we briefly discuss the criteria used to annotate VMWEs and the corpus. Our analyses are in §4 and §5, and we conclude in §6.

2 Annotation of Verbal Multiword Expressions

Our corpus was annotated according to the multilingual PARSEME guidelines v1.1, not restricted to PT-BR.[2] They define a MWE as a group of words that displays "some degree of orthographic, morphological, syntactic or semantic idiosyncrasy with respect to what is considered general grammar rules of a language" [15]. VMWEs are defined as "multiword expressions whose syntactic head in the prototypical form is a verb." VMWEs are annotated using flat annotations, where each token is tagged as being part of a VMWE or not, and where *lexicalized components* are explicitly marked, as these are the obligatory VMWE components. For instance, in *Maria* ***tomou*** *dois* ***banhos*** 'Maria took two showers', only the lexicalized components are shown in bold[3] as the determiner (*dois* 'two') can be replaced or omitted. Below, we summarize the criteria used to identify and categorize VMWEs, focusing on those that are relevant for PT-BR.

Verbal Idioms (VID) present some kind of semantic idiosyncrasy. Tests for semantic idiosyncrasies are hard to formulate, so we use flexibility tests[4] as a proxy to capture semantic idiosyncrasies. Success in *any* of these flexibility tests results in annotation as VID:

1. CRAN: The expression contains a cranberry word[5] e.g. ***foi para as cucuias*** 'went to the.FEM.PL cucuias' ⇒ 'went wrong'.
2. LEX: Replacement of a component by related words (e.g. synonyms, hyponyms, hypernyms) leads to ungrammaticality or unexpected meaning change e.g. ***quebrou*** *um* ***galho***/ *#ramo* 'broke a branch/#twig' ⇒ 'helped'.
3. MORPH: At least one of the components of the VMWE presents restricted morphological inflection with respect to general morphology, e.g. ***bateram perna***/*#pernas* 'hit.PST.3PL leg.SG/#legs.PL ' ⇒ 'they walked around'.
4. MSYNT: Morpho-syntactic changes lead to ungrammaticality or unexpected meaning change, e.g. *ela* ***eu perdi meu/#teu tempo*** 'I lost.PST.1SG my/#your time' ⇒ 'I wasted my time'.
5. SYNT: Syntactic changes are restricted, e.g. *eu* ***pisei na bola*** 'I stepped on-the ball' ⇒ 'I made a mistake' but not *#a bola na qual eu pisei* 'the ball on which I stepped'.

Light-Verb Constructions (LVC) are VMWEs composed of a light verb v and a noun n referring to an event or state. Their two sub-categories are LVC.full and LVC.cause. For LVCs, the following tests must be applied in the order specified below:

1. N-ABS: the noun n is abstract, e.g. *festa* 'party' and *prioridade* 'priority' in ***faremos*** *uma* ***festa*** 'we will throw a party' and *ele* ***dá prioridade*** *ao trabalho* 'he gives priority to his work'.

[2] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1.
[3] Boldface indicates lexicalized components for all examples throughout this paper.
[4] A *flexibility test* verifies to what extent a change usually allowed by a language's grammar also applies to the candidate to annotate.
[5] A word that does not co-occur with any other word outside the VMWE.

2. N-PRED: the noun n has at least one semantic argument, e.g. *visitas* 'visits' in ***fez visitas*** 'made visits', whose arguments are the visitor and the visitee.
3. N-SUBJ-N-ARG: v's subject is a semantic argument of n, e.g. *Maria* in *Maria* ***tomou banho*** 'Maria took a shower', which is the agent of *banho* 'shower'.
   - If test 3 passes, apply the two tests below:
     4. V-LIGHT: the verb v has light semantics, e.g. *prestar* 'to lend' in ***presta atenção*** 'lends attention' ⇒ 'pays attention'.
     5. V-REDUC: it is possible to omit v and refer to the same event/state, e.g. *o discurso da Maria* 'the speech by Maria' for *Maria* ***fez*** *um* ***discurso*** 'Maria gave a speech'. If this test passes, LVC.full is chosen.
   - If test 3 fails, apply V-SUBJ-N-CAUSE.
     6. V-SUBJ-N-CAUSE: v's subject is an external participant expressing the cause of n, e.g. *ratos* 'rats' in *ratos me* ***dão medo*** 'rats give me fear' ⇒ 'rats scare me'. If this test passes, LVC.cause is chosen.

Inherently Reflexive Verbs (IRV) are composed by a verb and a reflexive clitic, but the clitic does not fulfill one of its usual roles (reflexive, reciprocal, medium-passive, etc.). A verb-clitic combination is annotated as IRV only if one of the tests below passes:

1. INHERENT: the verb never occurs without the reflexive clitic, e.g. ***se queixam*** 'self.3 complain.PRS.3PL' ⇒ 'complain' but not **queixam* and ***me abstenho*** 'self.1SG abstain.PRS.1SG' ⇒ 'I abstain' but not **abstenho*.
2. DIFF-SENSE: the reflexive and non-reflexive versions do not have the same sense, such as *ele* ***se encontra*** *na cadeia* 'he self.3 meet in prison' ⇒ 'he is in prison' but *#ele me encontra na cadeia* 'he meets me in prison'.
3. DIFF-SUBCAT: the reflexive and non-reflexive versions do not have the same subcategorization frame, e.g. *ela* ***se esqueceu*** *de Maria* 'she self.3 forgot of Maria' ⇒ 'she forgot Maria' but *ela esqueceu Maria* 'she forgot Maria'.

3 VMWE-Annotated Corpus

The corpus used in this paper is freely available at the PARSEME v1.1 repository.[6] It contains texts from two sources: 19,040 sentences coming from the informal Brazilian newspaper *Diário Gaúcho* (DG) [2] and 9,664 sentences coming from the training set of the Universal Dependencies *UD_Portuguese-GSD* v2.1 treebank (UD) [7]. DG contains running text from full documents, whereas UD contains randomly shuffled sentences from the web.

In addition to manual VMWE annotations, the corpus includes lemmas, part-of-speech (POS) tags, morphological features, and syntactic dependencies using the Universal Dependencies tagsets [7].On the DG part, POS tags and syntactic dependencies were predicted automatically. On both the UD and DG parts, lemmas and morphological features were also predicted automatically. Predictions were made using UDPipe [16] and the CoNLL-2017 shared task model [19].

Figure 1 shows two corpus excerpts. All categories are represented: LVC.full (e.g. ***tem*** *o* ***direito*** 'has the right'), VID (e.g. ***tomar posição*** 'to take position'), and IRV (e.g. ***se identificar*** 'self.3 identify' ⇒ 'to identify oneself (as)'). In the whole corpus, 1 sentence contains 5 VMWEs, 6 sentences contain 4 VMWEs, 42 sentences contain 3 VMWEs, 473 sentences contain 2 VMWEs, 4,435 sentences contain 1 VMWE and 22,947 sentences contain no VMWE annotation at all.

[6] https://gitlab.com/parseme/sharedtask-data/tree/master/1.1/PT

This is a pre-print of an article published in PROPOR 2018.
The final authenticated version is available online at: https://www.springer.com/us/book/9783319997216

(1) Da mesma forma que a imprensa **tem** o **direito** (LVC.full) de **tomar posição** (VID), [...]

(2) [...] ao **se identificar** (IRV) como policial ele teria dito 'você não sabe com quem você mexeu', e **efetuou** os **disparos** (LVC.full) na vítima [...]

Fig. 1. Two example sentences with highlighted VMWE annotations (UD-train-s7090 and UD-train-s8536). Category labels shown above, lexicalized components in bold.

Table 1. Overall corpus statistics: number of sentences, tokens, annotated VMWEs and categories in the training (train), development (dev) and test portions.

	Sentences	Tokens	VMWEs	VID	LVC.full	LVC.cause	IRV
train	22,017	506,773	4,430	882	2,775	84	689
dev	3,117	68,581	553	130	337	3	83
test	2,770	62,648	553	118	337	7	91
Total	27,904	638,002	5,536	1,130	3,449	94	863

Table 2. Top-5 most frequent VMWEs per category, with frequency in parentheses.

LVC.full	VID	IRV	LVC.cause
marcar gol (47)	*fazer parte* (56)	*apresentar-se* (40)	*dar acesso* (7)
ter chance (43)	*ir ao ar* (48)	*tratar-se* (37)	*causar prejuízo* (6)
fazer gol (40)	*entrar em campo* (26)	*encontrar-se* (33)	*dar continuidade* (5)
ter direito (33)	*chamar atenção* (21)	*queixar-se* (31)	*gerar emprego* (4)
ter condição (29)	*ser a vez* (17)	*referir-se* (26)	*dar origem* (4)
correr risco (28)	*ter pela frente* (15)	*esquecer-se* (25)	*colocar em risco* (4)

Table 1 contains a summary of the corpus statistics. It contains in total 27,904 sentences and 5,536 annotated VMWEs, yielding an average of about 1 VMWE every 5 sentences. The predominant category is LVC.full, which represents more than 60% of the annotations. Then, VID and IRV represent respectively around 20% and 15% of the annotations. The corpus contains only few instances of LVC.cause, representing less than 2% of the total number of VMWEs. Because of its use in a shared task, the corpus is split into 3 portions: a training set (train), a development set (dev) and a test set.

The annotation of VMWEs was performed by a team of six PT-BR native speakers, including the authors of this paper, using a dedicated annotation platform [18]. The reported inter-annotator agreement between two of the annotators on a sample of 2,000 sentences is $\kappa = 0.771$ for VMWE identification, and $\kappa = 0.964$ for categorization [15].

Table 2 shows the 5 most frequent annotated VMWEs in each category. To extract this list, we have used the lemmas of annotated VMWEs in their canonical order to neutralize alternations (e.g. passive voice, enclitic vs proclitic pronouns). Since the majority of sentences comes from the DG newspaper, many VMWEs are related to topics often published in this newspaper, such as football (e.g. *marcar/fazer gol* 'mark/make

This is a pre-print of an article published in PROPOR 2018.
The final authenticated version is available online at: https://www.springer.com/us/book/9783319997216

Table 3. Average and histogram of VMWE length (L), i.e. nb. of lexicalized items, and of gap size (G), i.e. nb. of non-lexicalized items between first/last lexicalized ones.

	Length (L)				**Gap size (G)**			
	Avg(Stdev)	**%L=2**	**%L=3**	**%L≥4**	**Avg(Stdev)**	**%G=0**	**%G=1**	**%G≥2**
VID	2.90 (±1.01)	42.04	34.16	23.81	0.42 (±0.80)	66.64	28.32	5.04
LVC	2.05 (±0.24)	95.06	4.54	0.40	1.09 (±2.03)	40.76	41.18	18.06
IRV	2.00 (±0.08)	99.30	0.58	0	0.13 (±0.45)	87.72	12.05	0.23
All	2.22 (±0.61)	84.90	9.97	5.11	0.80 (±1.72)	53.36	34.01	12.63

goal' ⇒ 'to score a goal') and television (e.g. *ir ao ar* 'go to-the air' ⇒ 'to go on air'). In the remainder of this paper, we analyze the annotated VMWEs in terms of their properties and challenging aspects [3].

4 Characterization of Annotated VMWEs

Length and Discontinuities Table 3 summarizes the distribution of VMWE length and gap size. Most IRVs have exactly 2 lexicalized (L=2) adjacent (G=0) components. IRVs containing gaps (G=1) simply correspond to the non annotated intervening hyphen in proclitic uses (e.g. ***chama*** - ***se*** 'calls - self') whereas those of length 3 (L=3) or containing gaps larger than 1 (G≥2) correspond to annotation or tokenization errors (e.g. ***se auto*** - ***proclamava*** 'self auto - proclaim'). Most LVCs also have exactly 2 lexicalized components (L=2) but some include a lexicalized preposition (e.g. ***submetido a*** *um* ***tratamento*** 'subjected to a treatment'). The majority of LVCs have a gap (G=1) corresponding to a determiner. The distance[7] between the first and last lexicalized components of LVCs ranges from 0 (1,349 cases out of 3,543 LVCs) to 36 (1 case), with 9.20% having a distance of 3 or more intervening tokens (e.g. ***teve*** *há três anos a* ***ideia*** 'had three years ago the idea'). VIDs tend to be longer, with 2.9 tokens in average. The longest annotated VID contains 10 words (***está com a faca e o queijo na mão*** 'is with the knife and the cheese in-the hand' ⇒ 'is in good conditions to carry something out'). Most VIDs are continuous (G=0) but it is not uncommon to include a gap (e.g. ***cai*** *muito* ***bem*** 'falls very well' ⇒ 'comes in very handy').

Overlaps Overlapping VMWEs are rare but complex to model. Out of the 12,166 tokens belonging to a VMWE, 112 (≈1%) belong to multiple VMWEs simultaneously (overlaps). Among them, 67 are verbs, 27 are nouns and 18 belong to other POS tags. Overlaps are often caused by coordination, e.g. when a light verb is factorized for several predicative nouns (***ter***$_{1,2,3,4}$ ***ensino***$_1$ ***médio***$_1$ *completo,* ***experiência***$_2$ *em vendas, boa* ***comunicação***$_3$ *e* ***disponibilidade***$_4$ 'have completed high school, experience in sales, good communication and availability'). Noun overlaps are often due to coordination (e.g. *se vamos* ***fazer***$_1$ *ou não vamos* ***fazer***$_2$ ***sacrifícios***$_{1,2}$ 'if we will make or we will not make sacrifices') or due to relative clases (e.g. ***cometer***$_1$ *os* ***erros***$_{1,2}$ *que vinha* ***cometendo***$_2$ 'make the errors that he has been making').

Variability The 5,536 annotated VMWE tokens correspond to 2,126 unique normalized forms, with 1,244 (58.5%) of them occurring only once.[8] This raises concerns over the variability of the annotated VMWEs, which could impact the usability of this corpus when building machine learning models to automatically

[7] In number of intervening tokens.

[8] The *normalized form* of a VMWE is its sequence of lemmatized lexicalized components in lexicographic order, whereas its *surface form* is the textual sequence [8].

This is a pre-print of an article published in PROPOR 2018.
The final authenticated version is available online at: https://www.springer.com/us/book/9783319997216

Table 4. Proportion of VMWEs in dev/test corpora also present in the training corpus.

	Unseen	**Seen-identical**	**Seen-variant**
dev $\subseteq$ **train**	144/553=26%	180/553=33%	229/553=41%
test $\subseteq$ **train**	156/553=28%	164/553=30%	233/553=42%

identify VMWEs from incomplete/insufficient annotated data. Table 4 shows the coverage of the dev and test corpora with respect to the training corpus. Around 26-28% of the VMWEs in the dev/test corpora are unseen in the training data. Therefore, models learned on the training corpus will struggle to overcome 70% recall and should probably recur to external VMWE lexicons [9,4]. Among the 72-74% of seen VMWEs, most of them are actually variants, characterized by a normalized form identical to one seen in the training corpus, but with a different surface form. Hence, it is crucial to take morphological and syntactic variability into account when modeling VMWEs, otherwise $\approx$2/3 of them might be missed.

Ambiguity Human annotators and automatic VMWE identification systems need to distinguish true VMWE occurrences from literal uses and accidental co-occurrence [13]. Because of the polysemous uses of reflexive clitics in PT-BR, IRVs are quite ambiguous [12]. Examples include *dar-se* (IRV 'to happen' vs. 'to give-self'), and *formar-se* (IRV 'to graduate' vs. passive of 'to form'). This ambiguity is magnified by accidental co-occurrence due to POS-tagging errors, when the homonymous conjunction *se* 'if' is wrongly identified as a reflexive clitic. VIDs are generally less ambiguous, with some interesting examples of true ambiguity such as *fechou a porta, mas se esqueceu de trancá-la* 'closed the door, but forgot to lock it' vs. *duas escolas* ***fecharam as portas*** 'two schools have shut down'.

5 Challenging and Borderline Examples

Challenging LVCs According to the guidelines, LVCs contain predicative nouns (expressing an event or state, §2). These nouns are defined as having semantic arguments, that is, the meaning of the noun is only fully specified in the presence of its arguments. During annotation, we have found some challenging predicative constructions such as ***fazer falta*** 'make lack/foul', because they are ambiguous, and it is hard to identify the arguments of the noun. In *Os dois jogadores* ***fazem falta*** *ao time* 'The two players are missed by the team', the event can be rephrased as *a falta dos jogadores ao time* 'the lack of-the players to-the team', indicating that *falta* 'lack' has 2 arguments here, so it is a LVC.full. However, in *O jogador [...]* ***fez*** *uma* ***falta*** *desnecessária* 'The player [...] made an unnecessary foul', the verbless paraphrase *a falta do jogador* 'the player's foul' indicates that *falta* 'foul' only has one argument. Nonetheless this construction is also annotated as LVC.full. To complicate things further, *falta* 'foul' may also be combined with non-light verbs such as *cobrar* 'charge' and *bater* 'hit', where *falta* refers to a *free kick*. Both are annotated as VIDs.

Causative LVCs The guidelines distinguish full LVCs (LVC.full) from causative ones (LVC.cause). The corpus includes unexpected causative VMWEs, like ***trazer riscos*** 'bring risks' and ***levar à criação*** 'lead to-the creation'. Verbs like *trazer* are unexpected to form causative relations, but this is the fourth most frequent causative verb among the ones we annotated. One of the examples is *A ausência do sexo também* ***traz*** *uma forte* ***angústia*** 'Lack of sex also causes strong anguish' which we annotated as a LVC.cause. Since the LVC category is the most frequent one PT-BR, the specific tests in the guidelines and the mistakes found during pilot annotations helped the annotators to be consistent in annotating challenging cases like the ones exposed in this section.

Challenging IRVs The guidelines emphasize the difference between true IRVs and free constructions formed by a full verb combined with a reflexive clitic (§2). While it is relatively easy to identify IRVs that do not exist without the clitic, IRVs that bear a different meaning without the clitic posed some challenges to the annotation team. In particular, verbs like *encontrar-se* 'find-self' can be fully ambiguous in isolated sentences. For instance, in *A banda se encontrava novamente em São Paulo.* 'The band found itself again in São Paulo.' ⇒ 'The band met/was again in São Paulo', it is impossible to know, without access to a larger context, if the members of the band met each other, or if the information is solely that they were there. Another difficult case is *adaptar-se* 'adapt-self' that should not be annotated as IRV according to the provided tests. While the construction **A mãe adapta o filho à escola.* 'The mother adapts the son to-the school' is ungrammatical, the following one is perfectly admissible: *O escritor adapta o livro ao público.* 'The writer adapts the book to-the public'. Since the guidelines do not mention the semantic attributes of the arguments (e.g. +human), this example does not fit the definition of IRVs, even if it could be interesting to annotate it.

Underrepresented Categories: MVC, VPC and IAV Some VMWE categories described in the guidelines are underrepresented in PT-BR, namely verb-particle constructions (VPC), multi-verb constructions (MVC) and inherently adpositional verbs (IAV). The latter was optional and was not annotated in PT-BR. Only two possible cases of MVC were found in the corpus: *querer dizer* 'want know' ⇒ 'to mean' and *ouvir falar* 'hear talk' ⇒ 'to hear (about)'. Because they are extremely infrequent, both were annotated as VID, with the former being among the top-10 most frequent annotated VIDs. As for VPCs, there is only one (borderline) example of this category, namely *jogar fora* 'throw away' ⇒ 'throw away'. Since it is difficult to prove that *fora* 'away' ⇒ 'away' works as a particle in this case (as opposed to an adverb), and this is the only potential example of VPC in the corpus, it was annotated as VID.

Metaphors The concept of metaphors was relevant in the context of the PARSEME shared task, due to the fact that verbal metaphors are not always VMWEs. The distinction between these two categories is, as defined in the guidelines, "a relatively unstudied and open question". The guidelines suggest marking debatable examples and discussing them within the community. Given the characteristics of the corpus (newspaper and web texts) metaphors are rare. One of the most remarkable examples is the following: *o consumidor automaticamente pisa no freio e reduz as compras* 'the consumer automatically steps on-the brake and reduces the purchases'. A closer look shows that it is perfectly acceptable to exchange between *freio* 'brake' and *acelerador* 'accelerator' and keep the idea of the metaphor by opposition. Therefore, this possibility of changing the noun indicates that the construction is a regular metaphor, and not a VMWE.

Collocations The guidelines define collocations as "combinations of words whose idiosyncrasy is purely statistical". While this definition is debated by several authors, the annotated VMWEs follow the definition provided in the guidelines. For instance, *Renata [...] está quase realizando um sonho.* 'Renata [...] is almost fulfilling a dream' could be considered as a collocation or an LVC. The corpus provides evidence that it is only a collocation: the sentence *o presidente eleito [...] admitiu realizar um sonho de seu pai.* 'The president-elect admitted he is fulfilling his father's dream', shows the possibility of someone else fulfilling someone's dream. Furthermore, both verb and noun allow several other arguments, like *realizar um desejo/uma tentativa* 'to make a wish/attempt' and *ter/carregar um sonho* 'to have/carry a dream'. The distinction between collocations and VMWEs requires special attention and linguistic analysis, in order to restrict the annotation only to the target constructions.

6 Conclusions and Perspectives

In this paper we discussed the Brazilian Portuguese PARSEME corpus containing VMWE annotations. We described the annotation guidelines and process, and analyzed the corpus in terms of the diversity and distribution of the annotated expressions, along with their linguistic characterization. This analysis can be used as a basis for refining the annotation protocol to better tailor VMWEs. Moreover, this work can provide a foundation for NLP tasks and applications that target precise modeling of lexical, syntactic and semantic characteristics of these expressions. This includes their automatic identification in corpora, for which syntactic variation and discontinuities in their realization create challenges for current approaches. The application of our findings to enhance the quality of the annotated corpus and to aid the development of automatic VMWE identification methods is part of our goal for future work.

Acknowledgement

We would like to thank Helena Caseli for her participation as an annotator. We would also like to thank the PARSEME shared task organizers, especially Agata Savary and Veronika Vincze. This work was supported by the IC1207 PARSEME COST action[9] and by the PARSEME-FR project (ANR-14-CERA-0001).[10]

References

1. Baldwin, T., Kim, S.N.: Multiword expressions. In: Indurkhya, N., Damerau, F.J. (eds.) Handbook of Natural Language Processing, Second edition, pp. 267–292. CRC Press, Boca Raton (2010)
2. Bocorny Finatto, M.J., Scarton, C.E., Rocha, A., Aluísio, S.M.: Características do jornalismo popular: avaliação da inteligibilidade e auxílio à descrição do gênero. In: VIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. pp. 30–39. Sociedade Brasileira de Computação, Cuiabá, MT, Brazil (2011)
3. Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., Todirascu, A.: Multiword expression processing: A survey. Computational Linguistics **43**(4), 837–892 (2017). https://doi.org/10.1162/COLI_a_00302, https://doi.org/10.1162/COLI_a_00302
4. Constant, M., Nivre, J.: A transition-based system for joint lexical and syntactic analysis. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 161–171. Association for Computational Linguistics (August 2016), http://www.aclweb.org/anthology/P16-1016
5. Fotopoulou, A., Markantonatou, S., Giouli, V.: Encoding MWEs in a conceptual lexicon. In: Proceedings of the 10th Workshop on Multiword Expressions . pp. 43–47. MWE '14, Association for Computational Linguistics (2014)
6. Nissim, M., Zaninello, A.: Modeling the internal variability of multiword expressions through a pattern-based method. ACM TSLP Special Issue on MWEs **10**(2) (2013)
7. Nivre, J., de Marneffe, M.C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., Zeman, D.: Universal dependencies v1: A multilingual treebank collection. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation . pp. 1659–1666. LREC 2016, European Language Resources Association (ELRA) (May 2016)
8. Pasquer, C.: Expressions polylexicales verbales : étude de la variabilité en corpus. In: Actes de la 18e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues *(TALN-RÉCITAL 2017)* (2017)

[9] http://www.parseme.eu
[10] http://parsemefr.lif.univ-mrs.fr/

9. Riedl, M., Biemann, C.: Impact of MWE resources on multiword recognition. In: Proceedings of the 12th Workshop on Multiword Expressions . pp. 107–111. MWE '16, Association for Computational Linguistics (2016), http://anthology.aclweb.org/W16-1816
10. Rosén, V., Losnegaard, G.S., De Smedt, K., Bejček, E., Savary, A., Przepiórkowski, A., Osenova, P., Barbu Mitetelu, V.: A survey of multiword expressions in treebanks. In: Proceedings of the 14th International Workshop on Treebanks & Linguistic Theories Cconference (12 2015), https://hal.archives-ouvertes.fr/hal-01226001
11. Sag, I.A., Baldwin, T., Bond, F., Copestake, A.A., Flickinger, D.: Multiword expressions: A pain in the neck for NLP. In: Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing. CICLing '02, vol. 2276/2010, pp. 1–15. Springer-Verlag, London, UK (2002)
12. Sanches Duran, M., Scarton, C.E., Aluísio, S.M., Ramisch, C.: Identifying Pronominal Verbs: Towards Automatic Disambiguation of the Clitic 'se' in Portuguese. In: Proceedings of the 9th Workshop on Multiword Expressions. pp. 93–100. Association for Computational Linguistics, Atlanta, Georgia, USA (Jun 2013), http://www.aclweb.org/anthology/W13-1014
13. Savary, A., Cordeiro, S.R.: Literal readings of multiword expressions: as scarce as hen's teeth. In: Proceedings of the 16th Workshop on Treebanks and Linguistic Theories (TLT 16). Prague, Czech Republic (2018)
14. Savary, A., Jacquemin, C.: Reducing Information Variation in Text, Lecture Notes in Artificial Intelligence, vol. 2705, pp. 145–181. Springer, Berlin, Heidelberg (2003)
15. Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., Doucet, A.: The PARSEME Shared Task on automatic identification of verbal multiword expressions. In: Proceedings of the 13th Workshop on Multiword Expressions. pp. 31–47. MWE '17, Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/W17-1704, http://aclanthology.coli.uni-saarland.de/pdf/W/W17/W17-1704.pdf
16. Straka, M., Straková, J.: Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 88–99. Association for Computational Linguistics, Vancouver, Canada (August 2017)
17. Tutin, A.: Comparing morphological and syntactic variations of support verb constructions and verbal full phrasemes in French: a corpus based study. In: PARSEME COST Action. Relieving the pain in the neck in natural language processing: 7th final general meeting. Dubrovnik, Croatia (2016)
18. van Gompel, M., van der Sloot, K., Reynaert, M., van den Bosch, A.: FoLiA in practice: the infrastructure of a linguistic annotation format pp. 71–81 (2017). https://doi.org/https://doi.org/10.5334/bbi.6
19. Zeman, D., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinkova, S., Hajic jr., J., Hlavacova, J., Kettnerová, V., Uresova, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C.D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.C., Sanguinetti, M., Simi, M., Kanayama, H., dePaiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, c., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H.F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonca, G., Lando, T., Nitisaroj, R., Li, J.: Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 1–19. Association for Computational Linguistics, Vancouver, Canada (August 2017), http://www.aclweb.org/anthology/K/K17/K17-3001.pdf