Annotation of fixed Multiword Expressions (MWEs) in a Portuguese Universal Dependencies (UD) treebank: Gathering candidates from three different sources

Elvis de Souza[1], **Cláudia Freitas**[2]

[1]Department of Letters – PUC-Rio
Applied Computational Intelligence Laboratory – PUC-Rio

[2]Department of Letters – PUC-Rio

elvis.desouza99@gmail.com, claudiafreitas@puc-rio.br

Abstract. *Delimiting and correctly annotating multiword expressions (MWEs) is an important task in constructing a gold standard treebank. In this paper, we applied three methods to the PetroGold corpus to identify MWE candidates. The methods include (1) leveraging expressions previously identified by the PALAVRAS annotator, (2) statistical analysis of collocations in Petrolês, a larger non-annotated corpus, and (3) a curated list of co-occurring words from the POeTiSA project. Through extensive filtering and alignment with Universal Dependencies (UD) guidelines, we revised the annotations of 2,467 MWEs in the PetroGold corpus, we tested a new annotation for the part-of-speech (POS) of the words that are part of MWEs and we provide two computationally readable resources to assist other annotators.*

1. Introduction

Multiword Expressions (MWEs) are constructions that can take many forms in a language, such as compound nouns (e.g., “guarda-chuva” [umbrella] and “óleo diesel” [diesel oil]), institutionalized phrases (e.g., “comes e bebes” [food and drinks]), or functional phrases (e.g., “apesar de” [despite], “de acordo com” [according to]). [Ramisch 2012] shows that there is no single definition for MWEs in the literature, and they lie in the gray area between lexicon and syntax, presenting a relevant problem for NLP as they are difficult to handle while being very common in both everyday communication and specialized forms of communication.

Although there is no consensus on a definition, there are some common characteristics of these expressions according to [Ramisch 2012]: (1) they are arbitrary since perfectly grammatical expressions may not be accepted in certain contexts; (2) they are institutionalized, meaning they are part of everyday communication and are accepted and understood by speakers as a conventional way of expressing something; (3) they have limited semantic variation, as they do not undergo the process of semantic compositionality like other language constructions. Therefore, certain parts of an MWE cannot be replaced by any other words or constructions, as the expression is not the result of word composition (nor can the MWE be translated word by word); (4) they have limited syntactic variation, as conventional grammatical rules may not apply to these expressions, making it difficult to determine whether they belong to a speaker’s lexicon or grammar (and often they are also extragrammatical, meaning they are unpredictable and difficult to understand for a language learner who has only learned the grammar of the language), and (5) they are heterogeneous, covering a vast number of language phenomena, each with specific linguistic characteristics, which means that NLP applications should not use a unified methodology to process them.

Delimiting and correctly annotating multiword expressions is an important task in constructing a gold standard treebank. From a machine learning perspective, it is important to ensure consistent annotation of MWEs to avoid providing ambiguous clues about which words, when combined, should be treated as a unit in certain contexts. Without indicating which MWEs will be annotated as locutions (phrasal expressions) in a *corpus*, the morphosyntactic annotation of these phenomena can become inconsistent, with variations in different occurrences, or at worst, it may be impossible to perform any morphosyntactic annotation that makes sense for certain expressions without considering them as multiword units. For example, the expression "isto é" (that is), when used as a conjunctive locution, cannot be annotated as composed of a subject (the pronoun "isto") and a linking verb without jeopardizing the syntactic annotation of the rest of the sentence.

PetroGold v3 is the third version of PetroGold [de Souza 2023], a gold standard treebank for the oil & gas domain in Brazilian Portuguese. The PetroGold was also published in version 2.11 of the Universal Dependencies project, an initiative to provide treebanks for various languages using the same annotation scheme. As a result of this latest PetroGold version, several modifications were made to the annotation of multiword expressions to align with both the guidelines of the UD project and to increase the consistency in the annotation of such expressions.

The approach taken to identify candidates for multiword expressions in Portuguese involved three different sources:

1. **PetroGold:** Expressions previously identified by the PALAVRAS annotator [Bick 2014], which are present in the Bosque-UD *corpus* [Rademaker et al. 2017], and were annotated in previous versions of PetroGold.
2. **Petrolês:** Collocations of the form [PREP (DET)? N PREP], such as "de acordo com" (according to), identified through statistical methods inspired by [Oliveira et al. 2004], reproduced in Petrolês, a much larger non-annotated *corpus* from the oil & gas domain [Cordeiro 2020].
3. **POeTiSA:** A list of words that co-occur without ambiguity, meaning they are always annotated in the same way when they appear together, compiled within the POeTiSA project [Lopes et al. 2021].

These three lists were filtered to adapt them to the very restrictive definition of multiword expressions in the UD project. We also try an alternative annotation to the part-of-speech (POS) tag of the words that compose a MWE, giving each word the POS of the expression as a whole (an annotation which is not recommended by UD), in order to assess the results of a trained model when this alternative annotation takes place.

In the end, the annotations of 2,467 MWEs in the PetroGold *corpus* were revised. We provide, along with this paper, two computationally readable resources that can be useful in other annotation projects: one that represents the union of all the multiword expressions identified using the three methods (amounting to 150 MWEs), and a subset of this list that includes only the MWEs found in the PetroGold *corpus* (totaling 112 MWEs), along with their corresponding annotations[1]. The UD guidelines to annotate MWEs, the resources we used to find candidates and the results will be presented next.

2. Universal Dependencies take on MWEs

The UD project provides three classes for annotating multiword expressions: (1) "fixed" for fixed expressions that correspond to grammaticalized expressions, which behave as functional or short adverbial words; (2) "flat" for "semi-fixed" expressions, although there is no definition for them except for a list of examples (such as personal names, dates, compound numbers, and foreign phrases); and (3) "compound" for expressions that, unlike the others, have one word functioning as a syntactic head, as in the example "apple pie" (or in our case, "óleo diesel" [diesel oil]).

UD takes an economical approach to what it considers multiword expressions. For example, a very common structure in English is that of two nominals, such as "phone book" and "Natural Resources Conservation Service," or in the PetroGold context, "planta piloto" (pilot plant) and "fase rifte" (rift phase). The guidelines indicate that when there are clear criteria in the language documentation to distinguish compound expressions, the expression can be annotated as an MWE of type "compound," where all words in the expression are annotated as dependents of the main word with the *compound* relation (Figure 1). However, when the criteria are not well established, treebanks should let go of this annotation, tagging them as regular nominal modifiers (nmod).

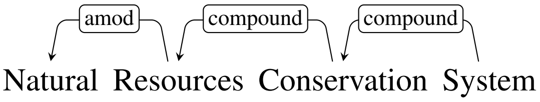

Figure 1. Possible annotation for the structure of "two nominals"

In versions 1 and 2 of PetroGold, we attempted to annotate expressions such as "óleo diesel" (diesel oil) and "meio ambiente" (environment) as *compound* since they would be useful in the oil and gas domain. However, due to the lack of a more comprehensive study on the subject and time limitations of the project, we decided to forego the *compound* label in this third version of the *corpus*, following the guidelines of UD. Instead, we opted for regular nominal modifiers (nmod) as a transparent annotation.

Another type of construction that can be considered an MWE in certain contexts is the one that contains light verbs (or support verbs), as in expressions like "dar um grito" (to scream), "ter em mente" (to have in mind), "tirar um cochilo" (to take a nap), "tomar uma decisão" (to make a decision), or "fazer vista grossa" (to turn a blind eye). In these examples, the verbs *dar/ter/tirar/tomar/fazer* (among others) are verbs "(...) with a greatly depleted meaning that, together with their complement (direct object), form a global meaning, usually corresponding to that of another verb in the language" [Neves 2000, our translation]. Not all expressions with support verbs can be replaced by other verbs in the language, as noted by [Bagno 2012], which justifies the use of the adverb "usually" by Neves. While "dar um grito" (to scream) and "tomar a decisão" (to make the decision) correspond to "gritar" (to shout) and "decidir" (to decide), respectively, "fazer questão"

[1] Available at: *https://github.com/alvelvis/mwe-petrogold-udfest.*

(to insist on) and “soltar balão” (to release a balloon) do not correspond to any verb (and are not compositional either, so these expressions would be filling a gap in the Portuguese lexicon through a phrase).

In any case, UD has defined that such expressions with support verbs should have a transparent annotation, with the noun as the object of the verb[2]. Thus, in sentence 1, “parte” (part) is the direct object of the verb “fazer” (to make), and “Projeto” (Project) is a prepositional object of the same verb, despite “fazer parte” (to be part of) being considered a multiword expression in the Portuguese language.

1. Este levantamento foi realizado em o ano de 1978, **fazendo parte** de o Projeto Aerogeofisico São Paulo – Rio de Janeiro de a CPRM.[3]

3. Methodology

3.1. Obtaining multiword expression candidates

Given that the only annotated MWEs in Portuguese UD are the *fixed* expressions (aside from compound proper names and numbers, respectively *flat:name* and *flat*), we used three methods to obtain candidates for them. The first source of fixed multiword expressions was obtained from the annotation inheritance in PetroGold. The PetroGold *corpus* was originally annotated (before the human inspection) by a model trained on the Bosque-UD *corpus*, which, in turn, is a manually corrected conversion of the PALAVRAS annotation system. Therefore, the first list of expressions analyzed consisted of expressions annotated as units in PetroGold, inherited from PALAVRAS, and subject to specific revisions in the *corpus*.

The second source of multiword expressions was obtained by applying statistical methods, based on [Oliveira et al. 2004], to Petrolês [Cordeiro 2020], a larger text collection, containing 330 academic documents in the field of oil & gas. [Oliveira et al. 2004] dealt with the notions of collocation and multiword expression in order to investigate the cases in which prepositional phrases are both multiword expressions (linguistic units) and collocations (words that frequently co-occur). The analyzed multiword expressions are of the form [PREP (DET)? N PREP], such as “de acordo com” (according to), where there is no determiner, and “no caso de” (in case of), with the determiner.

The method we used to identify collocates was the Likelihood-ratio, which is one of the methods used by [Oliveira et al. 2004]. It measures the probability that events that occurred together are not due to chance. Thus, two hypotheses are calculated: (i) that the words have the same probability of appearing together or separately, and (ii) that the words are more likely to appear together than separately. The metric tells us how much hypothesis (ii) is more likely than (i), and if that’s the case, the word sequence is considered a collocation [Manning and Schutze 1999]. We used the NLTK (Natural Language Toolkit) library for the Python programming language to calculate the collocates present

[2] Although they say that each language should define the criteria for annotating *compound*, the UD guidelines indicate that, for English, the “transparent” annotation is the most suitable for “light” constructions (such as “take a decision”) and adjective + noun combinations (*hot-dog*). Source: *https://universaldependencies.org/u/dep/compound.html*. Accessed on Mar. 5, 2023.

[3] Transl. “This survey was conducted in the year 1978, **as part of** the São Paulo – Rio de Janeiro Aerogeophysical Project by CPRM.”

in the *corpus* and identified which of them would qualify as prepositional phrases (not just collocations) in a sample of the results (the top 40 entries according to the algorithm's evaluation[4]). The results were then manually filtered to find the MWEs in the list.

The third and final source of multiword expressions is not a MWE list itself, compiled within the scope of the POeTiSA project, as part of a series of linguistic resources to improve the quality of POS annotation in a *corpus*. It is a list of words that, when co-occuring, are always unambiguous and should have the same POS annotation. The list was compiled by a linguist during the *corpus* annotation process[5] and the authors note that the entries in the list are not necessarily multiword expressions, so we conducted an analysis to filter out those cases that, according to our criteria, should be removed from the list.

3.2. An alternative Part-Of-Speech (POS) annotation for MWEs

UD prioritizes the annotation of part-of-speech (POS) tags based on the "base" class of the word, regardless of the context in which it is used. Therefore, according to the project, the POS annotation of the words that make up multiword expressions is not different from the annotation of the words if they were not part of an MWE. For example, in the case of "isto é" (meaning "this is"), the first element is a pronoun, and the second element is an auxiliary verb.

During the development of PetroGold, we added extra information to multiword expressions regarding the POS of the entire expression. This information is encoded in the tokens' miscellaneous attribute (tenth column). In the case of the expression "isto é" (meaning "this is"), the first token, "isto" (meaning "this"), annotated as a pronoun in UD, has the information "MWEPOS=CCONJ" in the miscellaneous column, indicating that it is a conjunctive locution. This "alternative" annotation was made with the intention of observing what would happen in syntax if the analysis of MWEs was not done literally at the POS level. To achieve this, we created a variation of the *corpus* where only the POS annotation of all words composing MWEs was replaced with the information from the MWEPOS field. Using the [Straka et al. 2016] tool, we generated two models: one for the modified version with MWEPOS and another for version 2.11 of the UD project. We compared the results using the evaluation metrics presented in [Zeman et al. 2018].

4. Results

The first method to obtain a list of MWEs was the revision of PetroGold, which had initially been annotated by a model trained on the Bosque-UD. The list of fixed MWEs in PetroGold initially contained 148 items. However, after applying the annotation criteria for expressions, only 101 items remained. The second method used was the statistical approach, which had a precision of 70%, considering that only the first 40 entries from the algorithm's results were analyzed (this number would drop significantly if expressions with lower scores were considered). And the third method, using a list compiled by

[4] The limit of 40 was established because, beyond this number, it became very difficult to find prepositional phrases through manual analysis.

[5] We would like to thank Magali Duran and the POeTiSA team for providing the list and all other resources from the project.

POeTiSA, had initially 110 expressions with unambiguous word co-occurrence. After our analysis, 55 MWEs remained from this list.

The three lists – expressions found in the original annotation of PetroGold, extracted through statistical methods from Petrolês, and compiled in the POeTiSA project – have some entries in common and others that are unique, highlighting the efficiency of the strategy used to obtain each list. The diagram in figure 2 shows the number of MWEs found by each method.

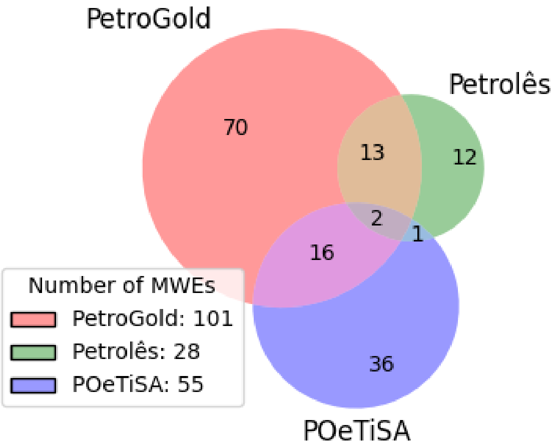

Figure 2. Quantity of multiword expressions obtained by each method

Proportionally, the PetroGold list brings the highest number of MWEs that no other method found – 69.3% of the entries are unique – followed closely by the POeTiSA list (65.45%), and far behind, the Petrolês list (42.85%). Although the statistical method was less efficient in finding MWEs, further investigation could be done in the future to determine if the search performed – for expressions of the type [PREP DET? N PREP] – and the algorithms used are the reason for the low coverage of MWEs returned by the method.

Finally, table 2 compiles the three returned lists without repetitions, containing 150 entries. These 150 entries were applied to the PetroGold *corpus* for the correction of MWEs, and the result is a total of 2,467 occurrences of fixed-type multiword expressions which are now available in the third version of PetroGold. The list containing all 112 MWEs found in PetroGold, along with their corresponding morphosyntactic annotation, can be found in the dedicated repository to this paper[6].

Regarding the part-of-speech (POS) annotation of the words that make up multiword expressions, we also tested an alternative annotation, where each word receives the POS of the entire expression. Thus, if the expression is a conjunction phrase (e.g., "isto é" meaning "that is"), both words were assigned the conjunction POS tag (CCONJ) in this alternative version.

By comparing two models trained on different datasets – the version that follows the UD guidelines (published in version 2.11 of the project) and this variation where the only modification was the POS annotation of MWEs – we observed the results in table 1.

[6] Available at: *https://github.com/alvelvis/mwe-petrogold-udfest*.

In this table, we can see that the POS[7] annotation results worsen, which is not surprising, given the ease with which a model seems to generalize that words should always receive the same POS tag, regardless of the context. However, there is an improvement in syntactic annotation that reaches 0.85 percentage point in the LAS[8] metric, which allows us to reflect on whether the annotation, for certain purposes, may be more appropriate.

UPOS	LAS
98.23% (-0.19 p.p.)	89.48% (0.85 p.p.)

Table 1. Variation in the automatic annotation when using MWEPOS

5. Concluding remarks

In conclusion, our analysis of multiword expressions (MWEs) using different methods and annotation techniques has yielded valuable insights. The results demonstrate that the PetroGold list and the POeTiSA list were the most effective in identifying unique MWEs, with proportions of 69.3% and 65.45% respectively. The Petrolês list, although lagging behind, still managed to uncover a significant number of MWEs at 42.85%. While the statistical method employed in this study proved to be less efficient in identifying MWEs, further investigation could be pursued to determine whether the search performed and the algorithms used are responsible for the lower coverage of MWEs obtained through this approach.

The compilation of the three returned lists, without repetitions, resulted in a total of 150 MWE entries. These entries were then applied to the PetroGold *corpus*, leading to the correction of 2,467 instances of fixed-type multiword expressions in the third version of PetroGold. For a comprehensive list of the 112 MWEs found in corpus, along with their corresponding morphosyntactic annotations, readers can refer to the dedicated repository associated with this article[9].

In terms of part-of-speech (POS) annotation, we also experimented with an alternative approach where each word in an MWE receives the POS tag of the entire expression. The comparison of models trained on both datasets revealed that while the POS annotation results worsened, there was an improvement of 0.85 percentage points in syntactic annotation, as measured by the LAS metric. This finding prompts further reflection on the appropriateness of this alternative annotation for specific purposes.

Overall, this study contributes to our understanding of MWE identification and annotation techniques, opening avenues for future research to enhance the effectiveness and coverage of MWE detection methods while considering the appropriate POS annotation strategies.

Acknowledgments

The authors express their gratitude to CNPq (National Council for Scientific and Technological Development, grant #130495/2021-2), FAPERJ (Carlos Chagas Filho Foundation

[7] POS: Universal Part-Of-Speech, which scores when the tagger finds the correct POS tag.

[8] LAS: Labeled-Attachment Score, which scores when the parser finds the correct attachment for the dependency and the correct label for the relation between governor and dependent.

[9] Available at: *https://github.com/alvelvis/mwe-petrogold-udfest*.

a a medida em que	além de isto	em a verdade	por conseguinte
a a medida que	além de o mais	em direção a	por exemplo
a a toa	além de o que	em face de	por exemplos
a as vezes	além de o quê	em função de	por fim
a cargo de	aos poucos	em geral	por mais que
a despeito de	apesar de	em o caso de	por meio de
a exemplo de	apesar de que	em o entanto	por muito que
a favor de	assim como	em o tocante a	por o menos
a fim de	assim por diante	em razão de	por parte de
a fim de que	assim que	em relação a	por pouco que
a longo de	assim sendo	em relação as	por sua vez
a medida que	até que	em relação á	por vezes
a menos que	bem como	em seguida	por volta de
a não ser que	cada vez mais	em separado	pouco a pouco
a o certo	caso contrário	em termos de	quanto a
a o contrário de	cerca de	em torno de	quanto mais
a o invés de	com base em	em vez de	se bem que
a o largo de	com isso	em vão	sem mais nem menos
a o longo de	com relação a	enquanto que	sem que
a o menos	com vistas a	isto é	sempre que
a o menos que	como relação a	junto a	sendo assim
a o passo que	como também	já que	sendo que
a o ponto de que	de acordo com	logo que	tais como
a o que	de acordos com	mais de o que nunca	tal como
a o todo	de agora em diante	mesmo assim	tanto quanto
a o vivo	de aí	mesmo que	tanto que
a partir de	de forma a	nada que	toda vez que
a ponto de	de forma que	nem a o menos	tudo quanto
a principio	de maneira a	nem mesmo	um a um
a princípio	de maneira que	nem sequer	um por um
a priori	de modo a	não obstante	um pouco
a respeito de	de modo que	não que	uma vez em
a seguir	de o que	não só	uma vez que
a título de	de sorte que	ou seja	visto que
ainda mais que	de tal forma que	para que	volta e meia
ainda que	desde que	para tal	é que
além de	devido a	pelo menos	
além de isso	em a faixa de	por causa de	

Table 2. Final list of multiword expressions obtained by all methods

for Research Support of the State of Rio de Janeiro, grant #E-26/202.433/2022), and ANP (National Agency of Petroleum, Natural Gas, and Biofuels, Brazil), associated with the investment of resources from the Clauses of R,D&I, through a Cooperation Agreement between Petrobras and PUC-Rio, for the provided support, without which this work could not have been accomplished.

References

Bagno, M. (2012). *Gramática pedagógica do português brasileiro*. Parábola Ed.

Bick, E. (2014). PALAVRAS, a constraint grammar-based parsing system for Portuguese. *Working with Portuguese corpora*, pages 279–302.

Cordeiro, F. C. (2020). Petrolês-como construir um corpus especializado em óleo e gás em português. *PUC-Rio, Rio de Janeiro, RJ-Brasil: PUC-Rio*.

de Souza, E. (2023). *Construção e avaliação de um treebank padrão ouro*. Mestrado, PUC-Rio.

Lopes, L., Duran, M. S., and Pardo, T. A. (2021). Universal dependencies-based pos tagging refinement through linguistic resources. In *Brazilian Conference on Intelligent Systems*, pages 601–615. Springer.

Manning, C. and Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Neves, M. H. d. M. (2000). *Gramática de usos do português*. Unesp.

Oliveira, C., Nogueira, C., and Garrao, M. (2004). Locution or collocation: comparing linguistic and statistical methods for recognising complex prepositions. In *Anais do 2º Workshop em Tecnologia da Informação e da Linguagem Humana*.

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and De Paiva, V. (2017). Universal dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206.

Ramisch, C. (2012). A generic framework for multiword expressions treatment: from acquisition to applications. In *Proceedings of the ACL 2012 Student Research Workshop*, Jeju, Republic of Korea. ACL. `https://aclweb.org/anthology/W12-3311`.

Straka, M., Hajic, J., and Straková, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Zeman, D., Hajic, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 1–21.