Still on arguments and adjuncts: the status of the indirect object and the adverbial adjunct relations in Universal Dependencies for Portuguese *

Elvis de Souza and Cláudia Freitas

Pontifical Catholic University of Rio de Janeiro
elvis.desouza99@gmail.com,
claudiafreitas@puc-rio.br

Abstract. We report the process of annotating verbal arguments and adjuncts in PetroGold, a treebank of the oil & gas domain. The corpus follows the dependencies approach of the Universal Dependencies multilingual project. The argument-adjunct distinction in UD is not a relevant one, and it is up to the contributors of each language to decide how to annotate it in some particular cases. After consulting Portuguese grammars to assist in the annotation of the adverbial adjunct and indirect object relations, we propose a semantic-discursively oriented approach, which was used in the PetroGold annotation and affected 14.8% of the sentences in the treebank. Finally, we present a visualization of the results, showing the distribution of verbs by transitivity in the corpus.

Keywords: Treebank annotation · Universal Dependencies guidelines · Portuguese grammar

1 Introduction

When syntactically annotating or revising a treebank, every word, phrase or term in a sentence must be classified. When there is not a possibility for multiple categorisation, in many cases the distinction between one class and another is not trivial. The difficulties may arise from lack of studies of the linguistic phenomenon or lack of specific annotation guidelines.

In Portuguese, the distinction between indirect objects (one of the verbal arguments) and adverbial adjuncts can be particularly difficult in some cases due to the fact that both phrases are prepositioned – in the adverbial adjunct, the difficulty only occurs when it is prepositioned – and both are dependent on the predicate head. None could say that both classes were not thoroughly studied, but the tendency of grammars is to simplify the subject, presenting prototypical sentences in which the distinction is more easily made. Our concern, however, is

* Elvis de Souza thanks the National Council for Scientific and Technological Development (CNPq) for the Masters scholarship process no. 130495/2021-2.

with real corpus sentences, such as sentences (1)-(3), found in Bosque [1], where distinguishing between verbal complement and adverbial adjunct is not a simple task.

1. Os jogadores se dividem **pelos dez quartos** do alojamento, equipados com frigobar, ar condicionado, televisão e telefone.[1]
2. Papa indica mulher **para secretaria**[2]
3. O PDT pretende reduzir os impostos federais **a quatro**.[3]

In this work, we report the process of annotating the prepositioned complements of verbs in the second version of PetroGold [9], a gold standard treebank of the oil & gas domain. The corpus contains 250,595 tokens (8,949 sentences) morphosyntactically annotated according to the multilingual annotation of the Universal Dependencies [6] project. Therefore, our starting point for studying the phenomenon is the project guidelines, discussed in section 2.1.

After noticing that the project allows each language contributors to find their own solutions to language-specific constructions such as the prepositional objects in Portuguese, we verify what Brazilian and Portuguese grammarians have said on the argument-adjunct distinction in section 2.2. Vilela and Koch [10], for instance, recognize that the argument-adjunct distinction "has deserved some reflection and a definitive conclusion has not yet been reached" (transl., p. 347). We will see inconsistencies in the criteria suggested by these and other authors as well as incentive to our proposal of a semantic-discursive criterion to differentiate prepositional objects from adverbial adjuncts.

We present our annotation proposal with the aim to increase the inter-annotator agreement without giving up meaningful linguistic information. We report the methodology used in the PetroGold annotation in section 3, and, in section 4, we carry out a study of the subcategorization of verbs in the corpus according to the results obtained.

2 A multilingual framework meets Portuguese grammarians

2.1 The core-oblique dichotomy in the Universal Dependencies framework

In the UD annotation guidelines, the argument/adjunct issue follows the same direction since the first version of the project: in view of stated difficulties which are present in a good number of languages that make up the project, UD decided to eliminate the distinction between argument and adjunct in favor of the core-oblique dichotomy.

[1] Transl. "The players are divided **into the ten rooms** of the accommodation, equipped with minibar, air conditioning, television and telephone."
[2] Transl. "Pope appoints woman **to secretary**"
[3] Transl. "PDT aims to reduce federal taxes **to four**."

Marneffe et al. [4] explain that "the core-oblique distinction has to do with the morphosyntactic encoding of dependents, not with their status as obligatory or selected by the predicate" (p. 268). Starting from the idea that some dependency relations are more equally encoded than others across languages, the core terms are those that would be less variably encoded and occur in the same way on the surface, being the subject and the bare object – when it occurs in an "unmarked" way. The criteria for defining what are marked or unmarked forms of the subject and object, as noted by Marneffe et al., are specific to each language, however, some criteria are recurrent, among which we highlight:

i Verbs usually only agree with core arguments.
ii Core arguments often appear as bare nominals while obliques are marked by adpositions or other grammatical markers.
iii Valency-changing operations such as passive, causative, and applicative are often restricted to the promotion or demotion of core arguments.

Considering the criteria, we conclude that phrases preceded by a preposition (item [ii]), when valency-changing operations are not allowed (item [iii]), cannot be core terms.

Zeman [11] notes that a simple criterion for distinguishing between core and oblique in the English treebank is the presence or absence of a preposition, a posture that could also be adopted for the Portuguese language. Thus, a verbal argument is *obj* (direct object) when it is not preceded by a preposition, it is *iobj* (indirect object) only when there is already a direct object in the sentence and this indirect object must necessarily be an oblique pronoun, as it occurs in the dative case and can be un-prepositioned, and *obl* for all other cases, both of prepositional arguments and of adverbial adjuncts.

For treebanks which previously differentiated both classes, Zeman [11] proposes a subspecification from the oblique, the *obl:arg* relation, to be used when, in addition to being prepositional, the phrase is also considered an argument of the verb. Thus, the tags change labels, but the difficulty of distinguishing the argument from the adjunct remains – the lack of consensus, in the grammatical tradition, between indirect object and adverbial adjunct, appears in the Portuguese UD between *obl* (verb-dependent, prepositional) and *obl:arg* (also verb-dependent and prepositional), the first being an adverbial adjunct and the second a verbal argument, traditionally named *indirect object*.

2.2 The argument-adjunct distinction in Portuguese grammatical literature

We consulted different Portuguese grammars about the phenomenon of prepositional phrases attached to verbs.

An essential element for Vilela and Koch [10] in the argument-adjunct distinction is the interrogation directed to the verb in order to identify those terms that "are installed in the very meaning of the predicate" (transl., p. 347). If the term answers the questions "who, which, what, where, how much, how" asked

to the verb, it is an argument; if, on the other hand, the phrase answers the questions "where, why, how, when", it is an adverbial adjunct. We see, however, that there are questions that are repeated in the two classifications (where, how and when), which are thus useless questions for distinguishing between the classes. In the sentences below, where in both (a) and (b) "Francisco" answers the question "*quem* colocou/descobriu" ("*who* put/discovered it"), "Francisco" is classified as an argument (of the subject type), but there is difficulty in classifying the phrase "na prateleira", as in both sentences the phrase answers the questions "*onde* colocou/descobriu" ("*where* they put/discovered it"), an answer that fits both the argument and the adjunct classifications, according to the authors criteria.

a O Francisco colocou a enciclopédia **na prateleira**. (transl. "Francisco put the encyclopedia **on the shelf**.")
b O Francisco descobriu a enciclopédia **na prateleira**. (transl. "Francisco discovered the encyclopedia **on the shelf**.")

In this case, the authors' "intuitions" (VILELA & KOCH [10], p. 348) would tell them that, for the verb "to put", "on the shelf" is an argument, while for the verb "to discover" it is an adjunct. The sense of intuition understood by the authors of the grammar is similar to that criticized by Borges Neto [5] in a similar context, when he provokes: "Perhaps illiterates may have ´'intuitions" about the language, linguists recall analyzes with whom they had contact" (transl., p. 69). The author suggests that this "intuition" is just a process of reaffirming the same categories by repeating analyzes already carried out by the grammatical tradition.

Vilela and Koch look for "supplementary criteria" to justify their intuition. They consider that by deleting an adjunct, the sentence would remain complete – according to them, one can say "Francisco discovered the encyclopedia Ø" and the sentence remains complete, but it would not be acceptable to end the other sentence in "Francisco put the encyclopedia Ø" without the place complement.

We carried out a brief exploration to verify the claim that the verb "to put" requires a place complement. We queried the corpus "todos juntos", in the AC/DC service of Linguateca[4], and it returned 313,047 occurrences of the verb "colocar" ("to put"). At the beginning of the list, we find a small number of sentences using the verb without the prototypical place complement, discrediting the authors' "intuition":

1. Para aproveitar o contra-ataque, Ramirez vai **colocar** os volantes Ney e Cristóvão exercendo uma forte marcação no meio-campo .[5]
2. Para situar nosso questionamento no modelo lógico da Política Nacional de Monitoramento e Avaliação da Atenção Básica8, é necessário **colocar** a

[4] Available at: https://linguateca.pt/ACDC. Accessed on 11 Jan. 2022.
[5] Transl. "To take advantage of the counterattack, Ramirez will **put** the midfielders Ney and Cristóvão exerting a strong marking in the midfield."

aquisição de novos conhecimentos e a melhoria do desempenho do Sistema Único de Saúde (SUS) como suas principais finalidades.[6]

Bechara [3] is careful not to call those prepositional phrases as neither prepositional objects (which is the case of "Amar **a Deus** sobre todas as coisas"/ lit. "Love **to God** over all things."), nor indirect objects ("The director wrote letters **to parents**"). He names them "relative complement", being similar to the direct object in semantic-syntactic properties, except for the presence of a preposition.

Bechara indicates that each verb is accompanied by its own preposition by what he calls "grammatical servitude". Thus, "depender de" ("to depend on"), "competir com" ("compete with") and "agregar a" ("aggregate to") are predictable, although there are exceptions: first, the case in which the norm allows the use of more than one preposition ("ela se parece ao/com o pai"/ "she resembles to/with her father"), and second, the case of linguistic variation (diatopic, diastratic and diaphasic), as with the verbs "socorrer", "contentar" and others, that can be used with or without a preposition. This position is updated by Bagno [2], who presents examples of historical change, and not just variations of Brazilian Portuguese, as in the cases of "desagradar (a) alguém" (lit. "displease (to) someone"), "desobedecer (a) algo" (lit. "disobey (to) something"), "aspirar (a) algo" (lit. "aspire (to) something"), etc.

Finally, Bechara reminds that not all scholars agree that relative complements should be considered arguments: "Taking into account exclusively the semantic aspect, many prefer to consider such terms as circumstantial or adverbial adjuncts (...)" (BECHARA [3], p. 446). As we will see in section 3, our proposal to annotate the verbal arguments and adjuncts is endorsed by this position.

3 Methodology

Our annotation of the *obl* and *obl:arg* relations is motivated by the need to both achieve internal consistency and to make the analyzes informative, distinguishing sentences (1) from (2), which will be *obl:arg* and *obl* (argument and adjunct, respectively), and equating (3) and (4), which will be *obj* and *obl:arg* (both arguments).

1. Gostar **de sorvete**. (lit. "To like **to icecream**.")
2. Viajou **de carro**. (transl. "Traveled **by car**.")
3. Assistiu **o filme**. (transl. "Watched **the movie**.")
4. Assistiu **ao filme**. (lit. "Watched **to the movie**")[7]

[6] Transl. "To place our questioning in the logical model of the National Policy for Monitoring and Evaluation of Primary Care8, it is necessary to **put** the acquisition of new knowledge and the improvement of the performance of the Unified Health System (SUS) as its main purposes."

[7] As noted by a reviewer, we could consider the preposition a pleonastic element, as the sentence admits passive alternation. However, this criterion is not absolute. The verb "gostar de" (lit. "to like to") can admit the passive alternation in informal

Our strategy is to look at the meaning of the prepositional phrase in the sentence – if its meaning is the meaning traditionally associated with an adverb (time, place, manner, finality, causality, conformity), we annotate it as *obl* and, in the absence of adverbial semantics, it is an *obl:arg*. Thus, we shift the syntactic focus on the demand made by the verb to semantic-contextual features of the noun phrase associated with it.

The corpus is composed of 20,210 verb occurrences (1,080 different lemmas), being very expensive to analyze them case by case. We bootstrapped the annotation from Stanza [8] and transformed it to our proposal using the established semantic-discursive criterion. Our strategy was conducted three steps:

718 verbs are associated with the preposition "em" (lit. "in")
371 verbs are associated with the preposition "com" (lit. "with")
307 verbs are associated with the preposition "a" (lit. "to")
305 verbs are associated with the preposition "para" (lit. "for")
250 verbs are associated with the preposition "de" (lit. "of")

Table 1. 5 first prepositions that are most associated to verbs in PetroGold

i In a spreadsheet, we list all the verbs that, indirectly, are associated with prepositions (in the dependency model, the preposition depends on a noun, which is dependent on the verb). We organized the spreadsheet by preposition, and a sample of the five most popular prepositions among verbs can be viewed in the Table 1. Four annotators were responsible for evaluating whether, for each combination of verb + preposition, the prepositional phrase could be an argument of the verb. The only focus of this step is to separate verbs that can have an argument from those that never do, because while any predicate can have an adverbial adjunct, not all can have an argument. When the annotators could not think of an argument for the verb + preposition combination, they looked at the occurrences in the corpus to make sure there was not one. This step is intermediate, and its goal is to facilitate the corpus review process, in order to minimize the number of occurrences that will be reviewed.

ii Automatic changes are performed in the corpus using the data from the spreadsheet reviewed by the annotators. Thus, if a combination of verb + preposition, such as "acarretar em" ("result in"), appeared in the spreadsheet as possibly having an argument, all occurrences of "acarretar em"

register, although being typically an indirect transitive verb, as well as possibly any other verb. This way, it is reasonable to equate sentences (1) and (4), since both have prepositional phrases which are arguments of the verb (*obl:arg*), regardless of passive alternation possibility: "A nova empreitada do Linkedin permitirá que os produtores de conteúdo vejam quantas vezes um texto **foi gostado**, comentado e compartilhado."

("result in") became an argument (*obl:arg*), regardless of whether they are correct, like the underlined words in the following sentence:

* Segundo Souza (2009), a estabilidade conferida às emulsões devido à presença dos agentes emulsionantes naturais **acarreta**, **em geral**, **em um incremento** significativo na sua viscosidade[8]

iii We contrast the automated changes made in step (ii) with the original parsing. At this point, each annotator is guided to cases where spreadsheet and parser diverged. For example, in the previous sentence, the parser tagged "geral" ("general") as an adverbial adjunct, but the spreadsheet signaled "acarretar em" ("result in") as a verb with an argument. Annotators, aid by a specific tool to contrast two analyses[9], should just select the correct one.

The goal of the strategy was to reduce the time needed to correct the arguments and adverbial adjuncts with prepositioned phrases, as we only verified the occurrences in which there was a discrepancy between the spreadsheet annotation and the parser annotation.

We provide the spreadsheet[10] which we used to indicate what verbs, when related to which prepositions, can have the (prepositioned) noun as their complement. The spreadsheet includes all verb lemmas that relate to prepositions in the corpus, with a noun example for each entry. It should be noted that our objective with the spreadsheet is not to provide the community with any kind of definitive list of the transitivity of verbs, since it played a small part of a bigger strategy to correct the annotation of difficult sentences. However, from the point of view of linguistic description, it may be interesting to obtain a list of verbs and prepositions, and it is still possible to rearrange it, in alphabetical order, obtaining a list of all the prepositions that relate to each of the verbs in the corpus instead of distributing verbs by preposition as we have provided.

4 Results

We have made available the modifications related to arguments and verbal adjuncts in version 2 of PetroGold [9]. As a result, 1,488 tokens were modified in the corpus, which corresponds to 14.8% of sentences being modified.

In Figure 1 we present the distribution of lemmas by the frequency they occur with an argument. In the figure, we also classify as arguments the object clauses, annotated as *xcomp* and *ccomp* in UD, a position also defended by Przepiórkowski and Patejuk [7].

We removed from the analysis all verbs in the participle form and verbs with the expletive pronoun "se" dependent on them. In sentences with participles,

[8] Transl. "According to Souza (2009), the stability conferred on emulsions due to the presence of natural emulsifying agents **results**, **in general**, **in a significant increase** in their viscosity"

[9] Available at: https://github.com/alvelvis/conllu-merge-resolver. Accessed 21 Feb. 2022.

[10] Available at https://petroles.puc-rio.ai, along with PetroGold v2.

it is difficult to automatically distinguish which verbs do not accept complement ("[isso] ocorre Ø"/"[this] occurs Ø", sentence (1)) and which could accept it ("[alguma reação] hidrolisou [a poliacrilamida]"/"[some reaction] hydrolysed [a polyacrylamide]", sentence (2)). In sentences with "se", there is also doubt about the presence or not of an object: sometimes the verb actually works as an intransitive ("[algo] se sobressai Ø"/"[something] stands out Ø", sentence (3)), sometimes the verb could be interpreted as accepting a complement ("[algum fenômeno natural] assentou [as rochas]"/"[some natural phenomenon] based [the rocks]", sentence (4)). As we still do not have a systematic study of these cases, we prefer to leave them aside for the moment.

1. Isso pode ter **ocorrido** devido o clorofórmio extrair também o tensoativo.[11]
2. Viscosidade vs. taxa de cisalhamento de poliacrilamida **hidrolisada**.[12]
3. Estas fontes **se sobressaem** no mapa de amplitude do sinal analítico referido acima.[13]
4. As rochas da Bacia Sanfranciscana **assentam-se**, em discordância erosiva e angular, sobre rochas paleoproterozóicas do embasamento.[14]

As a result of the elimination of these types of verbs from the analysis, the study counted with 9,653 verbal occurrences that are distributed in 719 lemmas, 66% of the total verbal lemmas in the corpus. We can see how many verb lemmas are never accompanied by object, how many are accompanied by objects less than 30% of the time, between 30% and 70%, more than 70% of times, and how many are always accompanied by objects (*obj*, *iobj*, *obl:arg*, *xcomp* and *ccomp*).

The vast majority of verbs in PetroGold are always followed by an argument. Secondly, we have verbs that most often have an argument, thirdly we have those that never have an argument, then those that are exactly in the middle, not trending towards neither transitivity nor intransitivity, and finally, those that almost never have an argument.

This slice of lemmas that are in the middle, between "never" and "always", corresponds to 25.8% of the lemmas in the corpus. That is, a quarter of the verbal lemmas are exactly halfway between intransitivity and transitivity. For all these cases, it cannot be said, on the one hand, that when they lack a complement the sentence is incomplete, and, on the other hand, it cannot be said that the verb does not allow a complement without making a considerable mistake with that statement. This type of statistical information that we obtained escapes a categorical description of verbal subcategorization, and it is only possible because we have annotated the adjunct-argument distinction in a way that avoided transitivity as an intrinsic property of verbs.

[11] Transl. "This could have **happened** due to the chloroform extracting the surfactant as well."

[12] Tranl. "Viscosity vs. shear rate of **hydrolyzed** polyacrylamide."

[13] Transl. "These sources **stand out** in the analytical signal amplitude map referred to above."

[14] Transl. "The rocks of the Sanfranciscana Basin are **based**, in erosive and angular unconformities, on Paleoproterozoic rocks of the basement."

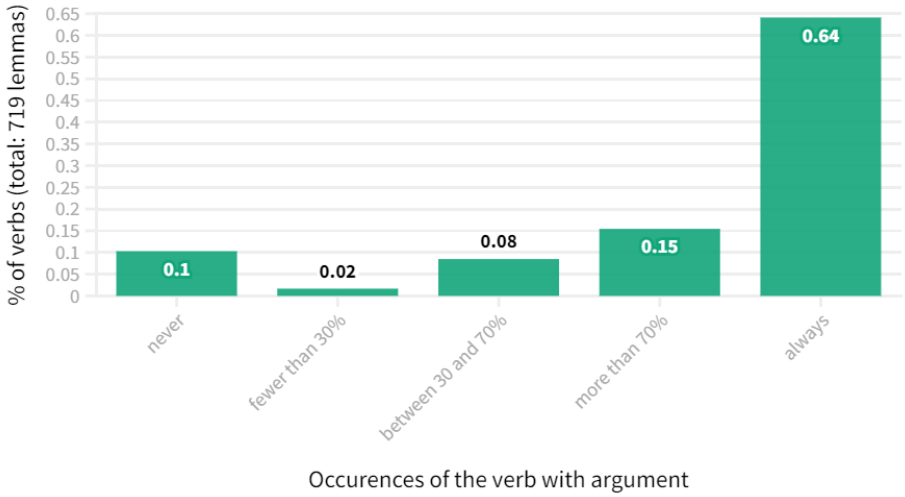

Fig. 1. Distribution of verbal lemmas in PetroGold by the frequency in which they occur with or without argument

5 Concluding remarks

This paper takes up the subject of verbal arguments and adjuncts with an empirical approach. First, we studied the status of the indirect objects and adverbial adjuncts in the Universal Dependencies guidelines, where we have seen enough arguments disfavoring this kind of distinction, while still leaving space for each treebank to discuss if and how they will annotate particular cases. Portuguese grammars brought many different criteria to establish the boundaries between both classes, but we saw they are insufficient when confronted with real language data. Then, we proposed a semantic-dicursive criterion, presented our annotation methodology and showed the results, which affected 14.8% of sentences in PetroGold and are featured in its second version.

References

1. Afonso, S., Bick, E., Haber, R., Santos, D.: Floresta sintá(c)tica: a treebank for Portuguese. In: Rodrigues, M.G., Araujo, C.P.S. (eds.) Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002). pp. 1698–1703. ELRA, Paris (29-31 de Maio 2002), http://www.linguateca.pt/documentos/AfonsoetalLREC2002.pdf
2. Bagno, M.: Gramática pedagógica do português brasileiro. Parábola Ed. (2012)
3. Bechara, E.: Moderna gramática portuguesa. Nova Fronteira (2012)
4. De Marneffe, M.C., Manning, C.D., Nivre, J., Zeman, D.: Universal dependencies. Computational linguistics **47**(2), 255–308 (2021)

5. Neto, J.B.: Morfologia: conceitos e métodos. Colóquios linguísticos e literários: enfoques epistemológicos, metodológicos e descritivos. Teresina: Edufpi pp. 53–72 (2011)
6. Nivre, J., De Marneffe, M.C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al.: Universal dependencies v1: A multilingual treebank collection. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 1659–1666 (2016)
7. Przepiórkowski, A., Patejuk, A.: Arguments and adjuncts in universal dependencies. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 3837–3852 (2018)
8. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A Python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (2020), https://nlp.stanford.edu/pubs/qi2020stanza.pdf
9. de Souza, E., Freitas, C.: Polishing the gold – how much revision do we need in treebanks? In: Proceedings of the I Universal Dependencies Brazilian Festival (UDFest-BR) (2022)
10. Vilela, Mário; Koch, I.V.: Gramática da língua portuguesa: Gramática da Palavra, Gramática da Frase, Gramática do Texto/Discurso. Almedina (2001)
11. Zeman, D.: Core arguments in universal dependencies. In: Proceedings of the fourth international conference on dependency linguistics (DepLing 2017). pp. 287–296 (2017)