Insights into the UD Tagset: Unveiling its Intricacies

Magali Sanches Duran

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP)

magali.duran@uol.com.br

Abstract. *This opinion paper explores our inclination to draw on principles of syntactic analysis established in the grammars of our native language when using the Universal Dependencies tagset to assign dependency relation tags. Taking the Portuguese language as a case example, this study argues that a fine-grained comparison of concepts and terms used in traditional grammars of Brazilian Portuguese and those used by Universal Dependencies reveals gaps which lead to different interpretations and, ultimately, to a deviation from the envisaged universality of the dependency relations tagset.*

Resumo. *Este artigo de opinião discute a tendência que temos de projetar, no conjunto de etiquetas de relações de dependência universais, conceitos e termos da análise sintática de gramáticas de nossa própria língua. Tomando a língua portuguesa como exemplo, este estudo argumenta que uma comparação pormenorizada entre os conceitos e termos usados nas gramáticas tradicionais do português do Brasil e os usados pelas Dependências Universais revela lacunas que levam a diferentes interpretações e, em última análise, a um desvio da universalidade prevista para o conjunto de etiquetas de relações de dependência.*

1. Introduction

Using Universal Dependencies (UD) (NIVRE et al. 2020; de MARNEFFE et al, 2021) to annotate a corpus in the Portuguese language and following the discussions that have taken place in the UD Issues on GitHub over the past three years, I have come to realize that there are still some unresolved issues with the assignment of UD dependency relation tags.

For the UD user community, maintaining tagsets unchanged is of paramount importance, as many corpora have already been annotated using them. Every change in a tagset or in the guidelines on how to apply its tags requires rework or, in a worst-case scenario, renders existing corpora non-compliant with the new guidelines. In May 2022, for example, there was a change in the way reported speech is annotated in UD. Although the tagset remained the same, the treebanks had to be adjusted to comply with the new guidelines.

It is also necessary to reach a consensus regarding which phenomena should be annotated with each tag. This is where many discussions start since some phenomena are more universal than others and the point of view of those who discuss the assignment of UD tags is always influenced by the languages they annotate. Although English is widely used as a lingua franca, it can be challenging to describe phenomena that do not exist in the English language to an English-speaking community.

By applying the UD scheme to Portuguese (PT) corpus annotation, I was able to detect occurrences where we had to make decisions because they were not covered by any tag in the UD tagset. Sharing the process of understanding Universal Dependencies (UD) concepts and identifying gaps in the annotation of certain phenomena may be of interest not only to PT annotators, but also to those who currently use or plan to use the UD annotation scheme in other languages as well.

2. The brief, naive illusion that concepts are familiar

PT annotators may feel comfortable when they first come into contact with the UD tagset because many of the relation tags appear to refer to syntactic functions they are already familiar with. However, not everything is what it seems to be. For example, anyone who sees the tags **nsubj**, **obj** and **iobj** quickly associates them with “subject”, “direct object” and “indirect object”, which are familiar terms in PT grammars. Our background shapes what we see.

However, this initial mapping of traditional PT syntactic functions onto UD relations is quite misleading. Although **iobj** is the short form for "indirect object" according to the UD guidelines, it is a case of "false friend" in that it does not correspond to the concept of an indirect object in PT (that is, an object introduced by a preposition). In fact, **iobj** is a tag dedicated to a third “core”[1] element (the other two being subject - **nsubj**, and object - **obj**), almost always non-prepositional, which occurs close to the verb and can occupy the subject position in a passive voice alternation. For instance, “Mary” in the English dative construction “John gave **Mary** a book” can be rendered as “Mary was given a book” in one of the two possible passive voice constructions in English. If “Mary” is prepositioned, as in “A book was given to Mary by John”, “Mary” is no longer an **iobj**.

UD’s decision to annotate **iobj** is supported by comparative studies on "core elements" across languages of various origins (THOMPSON, 1997; ANDREWS, 2007), and it is logical in an approach that aims for universal use. The only problem is the misleading retention of the adjective "indirect" in the tag.

In Portuguese, it seems that there are no cases of a third core element like in English dative constructions. Like other Romance languages, to the best of my knowledge, PT exclusively uses the dependency relation **iobj** to annotate dative pronouns[2] as they are not introduced by prepositions. Although these pronouns cannot assume the subject position in the passive voice, they occur near the verb and meet the criterion of givenness, a key issue in determining core elements.

[1] We found no criteria for distinguishing core arguments from other dependents in Portuguese, which is why we used the criteria adopted by the English language.

[2] In Portuguese: *me, te, lhe, se, nos, vos, lhes* (dative pronouns in English are: me, you, him, her, us, them)

The syntactic function known as “indirect object” in PT grammars is annotated as **obl** (oblique) in UD. The relation **obl** is used for both argumental and adjunct modifiers in the form of prepositional noun phrases (PP). This position of UD Guidelines is supported by psycholinguistic studies indicating that the boundary between argumental PP and adjunct PP is not well defined (see Boland & Blodgett (2006) to revisit some of them).

In the first version of UD Guidelines, there was no **obl** relation: all PP modifiers were annotated as **nmod** (nominal modifier). As of version 2, **nmod** is used exclusively for nominal modifiers of NOUN, PROPN and PRON, and the newly created **obl** is now applied to PP modifiers of VERB, ADJ and ADV, both argument and adjunct (UD guidelines do not make a distinction between arguments and adjuncts).

The problem is that, although having the same lexical realization, adjuncts modifying nominals, adjectives, adverbs, and verbs, are now annotated with different dependency relations depending on the part-of-speech (PoS) tag of the head of the dependency relation. For example: in “Rainfall in March is a problem”, “in March” is annotated as **nmod**, because its head “rainfall” is a NOUN, whereas in “It starts to rain in March”, “in March” is annotated as **obl**, because its head “to rain” is a VERB. This difference in classification comes naturally to PT annotators since PT grammars dictate that the former would be classified as “adnominal adjunct” and the latter as “adverbial adjunct”. However, when it comes to arguments, this division of relations into **nmod** and **obl** separates into two groups cases annotated as "*complemento nominal*" (noun complement) in PT. These are PP related to argument-taking nouns, adjectives and adverbs, such as “lack **of confidence**” (**nmod**), “eager **for change**” (**obl**), and “regardless **of nationality**” (**obl**), respectively.

As a result, the relation **obl** corresponds to three syntactic functions in traditional grammars of PT: “nominal complement” of adverbs and adjectives; “indirect object”, and “adverbial adjunct” in PP form. Examples of **obl** (in bold) are:

- *ansioso* ***por novidades*** [avid **for news]: obl** modifying an ADJ, corresponding to a “nominal complement” in PT;
- *independentemente* ***da hora*** [regardless **of the time] obl** modifying an ADV, corresponding to a “nominal complement” in PT;
- *reclamar* ***do barulho*** [to complain **about the noise] obl** modifying a VERB, corresponding to an “indirect object” in PT;
- *dormir de noite* [to sleep **at night**] **obl** modifying a VERB, corresponding to an “adverbial adjunct” in PT.

The relation **nmod**, on the other hand, is used to annotate any nominal modifier of a nominal (NOUN, PROPN, PRON), which corresponds to what PT traditional grammars refer to as “*adjuntos adnominais*” (“adnominal adjuncts”) and “nominal complements”. Following this rationale, **nmod** is also used to annotate what is known as a “*aposto especificativo*” (“specifying appositive”) in PT. Examples of **nmod** (in bold):

- *gosto* ***de chocolate*** *na boca* (a taste **of chocolate** in one's mouth) **nmod** corresponding to a “nominal complement” in PT;

- *gosto de chocolate* ***na boca***" (a taste of chocolate **in one's mouth**) **nmod** corresponding to an "adnominal adjunct" in PT;
- o *presidente* ***Lula*** (President **Lula)** **nmod** corresponding to a "specifying appositive" in PT.

While **obl** and **nmod** are broad relations, **appos** (appositional modifier) has a more restricted usage. The tag is exclusively used for relations that satisfy the following restrictions: occuring after the nominal they modify, having the same referent as the nominal they modify, and being interchangeable with the modified nominal. Examples of **appos** that meet these restrictions (in bold) are:

- *Pelé,* ***o rei do futebol****, morreu no último ano.* (Pelé, the **king** of soccer, died last year.) **appos** corresponding to an appositive in PT;
- *O rei do futebol,* ***Pelé****, morreu no último ano.* (The king of soccer, **Pelé**, died last year.) **appos** corresponding to an appositive in PT;

The modifiers family has three other members: **amod** (adjectival modifier), simple adjectives that modify nominals, as in "an **incredible** landscape"; **nummod** (numeric modifier), numbers that modify a noun indicating a quantity, as in "**three** years"; and **advmod** (adverbial modifier), simple adverbs that modify verbs, adjectives, adverbs and, to a lesser extent, even nouns, as shown in the following examples (**advmod** in bold):

- *falar* ***alto*** (to speak **loudly**) **advmod** modifying a VERB;
- ***extremamente*** *cansado* **(extremely** tired) **advmod** modifying an ADJ;
- ***somente*** *agora* **(only** now) **advmod** modifying an ADV;
- ***só*** *amigos* **(just** friends) **advmod** modifying a NOUN.

3. The search for symmetry

As can be seen, the PoS (Part-of-Speech) tag of the dependent (and sometimes the PoS tag of the head) is used as a criterion for assigning dependency relations. This approach is effective for phrasal dependents, but not for clausal dependents, obviously, because predicates, except for nominal predicates, always implicate a VERB.

Notwithstanding, when we see that a clausal subject is **csubj**, we promptly think: **csubj** is for **nsubj** just as other clause types are for other relations (**obj, iobj, obl, nmod, amod, advmod, appos**), an association similar to that of subordinate clauses in PT (subject subordinate clause, direct object subordinate clause, etc.).

This natural quest for mappings is an individual exercise, since, except for **nsubj/csubj**, the UD guidelines do not associate simple relations (**obj**, **iobj**, **nmod**, **obl**, **advmod**, **amod**, **appos**) with clausal ones (**acl**, **advcl**, **ccomp**, **xcomp**) on a one-to-one basis.

The initial shift that unveils non-symmetric mappings is the existence of two relations for clausal objects: **ccomp** and **xcomp**. The criterion for distinguishing them is related to the subject of the dependent clause. If the subject or object of the parent clause controls a null subject in the dependent clause, it is an **xcomp**[3]. Otherwise, it is a **ccomp**. In PT, the predicate of a **ccomp** dependent always implicates a finite[4] form and is introduced by a subordinating conjunction or a wh- adverb. The predicate of an **xcomp** always implicates an infinitive form, and is not introduced by subordinating conjunctions. It is relevant to say that the finite form of **ccomp** dependents and the non-finite form of **xcomp** dependents may be "assumed" by the main verb, by an auxiliary or by a copula verb (for nominal predicates), as underlined in the following examples (**ccomp** and **xcomp** dependents in bold)[5]:

- *Ele confirmou que* ***viria***. (He confirmed that he would **come**.) **ccomp**
- *Ele confirmou que havia* ***bebido***. (He confirmed that he had **drunk**.) **ccomp**
- *Ele nos disse que seria o* ***palestrante*** *convidado*. (He told us he would be the invited **speaker**.) **ccomp**
- *Ele quer* ***vir***. (He wants to **come**.) **xcomp**
- *Ele queria ter* ***vindo***. (He wished he had **come**.) **xcomp**
- *Ele pretende ser* ***professor***. (He intends to be a **teacher**.) **xcomp**

However, we encounter a problem when we use **xcomp** in PT: there are many clauses that have all the characteristics of a **xcomp,** but are introduced by a preposition, where in English an infinitive marker is used ("to"[6]):

- *Ele começou* ***a*** *andar.* (He started **to** walk.)
- *Ele esqueceu* ***de*** *fazer isso.* (He forgot **to** do this.)

Since the UD guidelines draw on English-based models and most examples are provided in English, the problem of an **xcomp** introduced by a preposition rarely arises because the infinitive marker "to" is the most usual form. However, when a preposition is used in English, the question arises: is it a marker of an **xcomp**? Ex:

- I'm relying on you to **come**.
- He complained about not being **invited**.

A pending question is: Is there a clausal equivalent to **obl**? That doesn't seem to be the case. Moreover, it seems that there is no one-to-one mapping between phrasal and clausal dependents in UD. And this is not a problem as the ambiguity existing between arguments and adjuncts in phrasal dependents is not present in clausal dependents. Therefore, if a sentence exhibits all the characteristics of an **xcomp**, it should, in my opinion, be annotated as an **xcomp**, regardless of whether it is introduced by a preposition or not.

[3] The name and the concept **xcomp** is borrowed from Lexical-Functional Grammar (LFG) (BRESNAN, 1982). Curiously, **xcomp** is used in LFG to distinguish complements from adjuncts, a distinction that UD rejects.

[4] The finite form can be expressed by auxiliary verbs or by copula verbs that modify the predicate, not necessarily by the predicate itself.

[5] Although this is not a criterion for distinguishing **ccomp** from **xcomp** in UD, this characteristic in PT contributes to having less confusion between these two tags in a confusion matrix.

[6] The English infinitive marker "to" is not always translated by a preposition in PT: *Ele quer fazer isso.* (He wants to do this.) *Ele pretende viajar*. (He intends to travel.).

If there is no clausal equivalent for **obl**, what about adverbial adjuncts and arguments of adjectives and adverbs? In PT, arguments of nouns, adjectives and adverbs are classified as "nominal complements", while their clausal correlatives are classified as "*orações completivas nominais*" ("nominal complement clauses"). As UD only explicitly annotates clausal modifiers for nouns as **acl** (adnominal clauses), there is a gap in terms of clausal (argument) modifiers of adjectives and adverbs. We have decided to fill this gap by utilizing our background, specifically by expanding the use of **acl** to clausal modifiers of adjectives and adverbs. This is an advantage from an NLP perspective, because noun, adjective and adverb complement clauses in PT have the same form: they are introduced by a preposition and always implicate an infinitive form[7] or a subjunctive inflected form

Again, in some cases, constructions with prepositions in PT are translated as constructions with prepositions into English, though in others, that is not the case. When an infinitive form in PT is translated as a gerund in English, a preposition may occur. However, when the infinitive in PT is translated as infinitives in English, the preposition is not used because in English the infinitive marker "to" and prepositions do not co-occur.

- *vontade* ***de*** *<u>viajar</u>* (desire **to** <u>travel</u>)
- *medo* ***de*** *<u>ser demitido</u>* (fear **of** <u>being fired</u>)
- *ansioso* ***para*** *<u>viajar</u>* (eager **to** <u>travel</u>)
- *temeroso* ***de*** *<u>ser demitido</u>* (afraid **of** <u>being fired</u>)
- *independentemente* ***de*** *<u>ter dinheiro</u>* (regardless **of** <u>having money</u>)

The lack of symmetric mappings between phrasal and clausal dependents can also be observed with regard to **amod**. In its clausal form, it corresponds to adnominal clauses - **acl (**the same relation used to annotate the clausal version of **nmod**) and to relative adnominal clauses - **acl:relcl**.

Another question that arises is: What is the clausal version of adverbial adjuncts, annotated with **advmod** or **obl**, depending on whether their PoS tag is an ADV or a NOUN? It seems that they are all covered by the **advcl** dependency relation. According to the UD guidelines[8], **advcl** can modify any predicate, whether it is verbal or nominal. The traditional semantic labels of adverbial clauses, such as temporal, consequence, conditional, and purpose, are mentioned in the guidelines as examples. According to the UD guidelines, the dependent of an **advcl** "must be clausal (or else it is an **advmod**)."

Finally, there are, in PT, clauses classified as "*orações apositivas*" ("appositive clauses"), in which the dependent of the relation is a clause, and "*aposto de oração*" ("apposition of clause"), in which the head of the relation is a clause.

- *Ele só quer <u>isso</u>: que você* ***venha***. (He wants only <u>this</u>: that you **come**.) "appositive clause" in PT;

[7] Some nouns and adjectives also allow clausal complements in the form of finite clauses. In this case, the verb takes the subjunctive mood. Ex: *Eu tenho medo de que você se fira.* (I am afraid of you hurting yourself.)

[8] https://universaldependencies.org/guidelines.html

- *Ele propôs irmos de carro, **proposta** que ninguém aceitou.* (He proposed to go by car, a **proposal** that nobody accepted.) “apposition of clause” in PT, resumptive clause in English;
- *Ele propôs irmos de carro, **o** que ninguém aceitou.* (literally: He proposed to go by car, which nobody **accepted**.) “apposition of clause” in PT, summative clause in English.

In the UD tagset for dependency relations there is no corresponding clause to **appos**, and **appos** does not allow for either the head or the dependent to be a clause. Therefore, the decision of how to annotate them is up to each language or annotation project. In PT we advocate annotating most of them as **parataxis**, except for one case in which the referent of a clause is another clause, as if it were a relative clause with a clausal antecedent. In this case, the relative clause that modifies another clause could be well represented by **advcl:recl**, since it is adjunctive and modifies a predicate, similar to other **advcl**:

- *Ele esqueceu a chave em casa, o que o **fez** se atrasar.* (He forgot his key at home, which **made** him late.)

4. Conclusion

By contrasting the UD tagset of dependency relations with the PT set of syntactic functions, we were able to find areas of isomorphism (coincident terms and concepts) and anisomorphism (non-coincident terms and concepts) between the two. This can provide valuable insights for annotators working with different languages who currently use or plan to use the UD tagset. It helps them recognize areas of divergence and prevent the misapplication of concepts from their native language grammars to UD annotation.

This exercise is also beneficial for highlighting gaps in assigning tags to language phenomena. Some examples include:

- clauses introduced by prepositions that clearly function as an **xcomp,** but are not addressed in the UD guidelines;
- clauses introduced by prepositions that serve as complements of argument-taking adjectives and adverbs;
- appositive clauses.

By explicitly defining how to annotate clauses like these, UD would prevent each language or project from filling in the gaps according to its own interpretation. This is crucial for maintaining the universality of the UD tagset.

Acknowledgements:

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI - http://c4ai.inova.usp.br/), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation. The author thanks the support from the Ministry of Science, Technology and Innovation, with resources of Law N. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU

01245.010222/2022-44. The author is also profoundly grateful for the two anonymous reviewers' comments and suggestions.

References

Andrews, Avery. (2007). The major functions of the noun phrase. In T. Shopen (Ed.), Language typology and syntactic description (pp. 62-154). Cambridge: Cambridge University Press.

Boland, Julie E.; Blodgett, Allison. (2006) Argument Status and PP-Attachment. Journal of Psycholinguistic Research, 35, pages 385–403. DOI 10.1007/s10936-006-9021-z

Bresnan Joan. (1982) The Mental Representation of Grammatical Relations. MIT Press, Cambridge, Massachusetts. https://doi.org/10.2307/414493

de Marneffe, Marie-Catherine; Manning, Christopher D.; Nivre, Joakim; Zeman, Daniel. (2021). Universal Dependencies. Computational Linguistics, 47(2):255–308. https://aclanthology.org/2021.cl-2.11

Nivre, Joakim; de Marneffe, Marie-Catherine; Ginter, Filip; Hajič, Jan; Manning, Christopher D.; Pyysalo, Sampo; Schuster, Sebastian; Tyers, Francis; Zeman, Daniel. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020), pages 4034–4043, Marseille, France. European Language Resources Association. https://aclanthology.org/2020.lrec-1.497

Thompson, S. A. (1997). Discourse motivations for the core-oblique distinction as a language universal. In Akio Kamio (editor), Directions in Functional Linguistics, 36, pages 59–82. John Benjamins. https://doi.org/10.1075/slcs.36.06tho