



An annotation manual as a Natural Language Processing resource: the Universal Dependencies model in Portuguese

Manual de anotação como recurso de Processamento de Linguagem
Natural: o modelo *Universal Dependencies* em língua portuguesa

Magali Sanches DURAN*

Maria das Graças Volpe NUNES**

Lucelene LOPES***

Thiago Alexandre Salgueiro PARDO****

ABSTRACT: With the advances of the Natural Language Processing area, corpora are resources that have had a prominent place. More than subsidizing linguistic studies, they constitute the basis for training Machine Learning models and developing cutting-edge computational applications. In particular, there is a great need for annotated corpora, but their production requires another essential resource, the annotation manual, which instantiates the annotation model of interest for the language in question and outlines the annotation decisions that should be adopted. In this paper, we explore issues related to the development of manuals for the annotation of Brazilian Portuguese corpora according to the Universal Dependencies model, widely adopted in the field. We discuss the

RESUMO: Com o avanço da área de Processamento de Linguagem Natural (PLN), *corpora* são recursos que têm tido um lugar de destaque. Mais do que subsidiar estudos linguísticos, eles constituem as bases para o treinamento de modelos de Aprendizagem de Máquina e para o desenvolvimento de aplicações computacionais de ponta. Particularmente, há grande necessidade de *corpora* anotados, porém sua geração requer outro recurso essencial, o manual de anotação, que instancia o modelo de anotação de interesse para a língua em questão e delinea as decisões de anotação que devem ser adotadas. Neste artigo, exploramos questões relacionadas ao desenvolvimento de manuais para a anotação de *corpus* em português brasileiro segundo o modelo

* PhD in Linguistic Studies from UNESP, São José do Rio Preto and Postdoctoral Researcher at NILC. ORCID: <https://orcid.org/0000-0002-3843-4600>. magali.duran@uol.com.br

** Professor at the Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos campus. ORCID: <https://orcid.org/0000-0002-2776-6140>. gracan@icmc.usp.br

*** PhD in Computer Science from PUC-RS and Postdoctoral Researcher at NILC. ORCID: <https://orcid.org/0000-0003-0314-140X>. lucelene@gmail.com

**** Professor at the Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos campus. ORCID: <https://orcid.org/0000-0003-2111-1319>. taspardo@icmc.usp.br

evolution of NLP and the use of corpora, the fundamental issues, resources and tools related to syntactic representation, the Universal Dependencies model, and the main decisions made in the instantiation of UD guidelines in Brazilian Portuguese. For practical and didactic reasons, we divided the manual into two parts: the PoS Tag Annotation Manual (morphosyntactic annotation) and the Dependency Relations Annotation Manual. Both resulted from the process reported in this paper and are available for free access on the POeTiSA project's Web site.

internacional *Universal Dependencies*, amplamente adotado na área. Partimos da discussão da evolução do PLN e o uso de *corpora*, passamos pelas questões, recursos e ferramentas fundamentais relacionados à representação sintática, discutimos o modelo *Universal Dependencies* e apresentamos as principais decisões tomadas na instanciação de suas diretrizes no português brasileiro. Por questões práticas e de didática, dividimos o manual em duas partes: o Manual de Anotação de *PoS tags* (anotação morfossintática) e o Manual de Anotação Relações de Dependência. Ambos foram resultado do processo relatado neste artigo e estão disponíveis para livre acesso no site do projeto POeTiSA na Web.

KEYWORDS: Annotated corpora. Annotation manual. Universal Dependencies. Dependency trees. Brazilian Portuguese.

PALAVRAS-CHAVE: *Corpora* anotados. Manual de anotação. *Universal Dependencies*. Árvores de dependência. Português brasileiro.

1 Introduction

In this paper, we present the annotation guideline as a Natural Language Processing (NLP) resource, a perspective that is not widely recognized among either linguists or computer scientists. To substantiate this claim, we have divided this introduction into three subsections, beginning with a retrospective on the evolving role of the linguist in NLP (1.1), move on to the rise of annotated corpora as a valuable asset for modern NLP methods (1.2), and conclude by discussing the importance of annotation guidelines as an integral part of any annotation scheme (1.3).

1.1 Natural Language Processing: The Evolution Towards Corpora

NLP has developed significantly in recent decades, and the linguist's role in these projects has shifted accordingly. Until the 1990's, the prevailing belief was that linguistic knowledge, represented through lexicons, ontologies, grammars, rules, and

other knowledge representation formalisms, could provide all the necessary groundwork for NLP.

From the 1990's onward, a series of technological advancements and shifts in social behavior revolutionized the entire field. The advent of the internet and its interfaces led to an exponential increase in the volume of natural language documents on the Web and, consequently, has increased the demand for processing tasks: search, classification, production support, analysis and correction, information retrieval, and summarization, among many others. Hardware advancements massively boosted processing power and data storage capacity, enabling the deployment of sophisticated methods that were previously prohibitive due to their heavy time and memory demands. Finally, the more recent rise of social media has driven a surge in user-generated content, compelling NLP to handle texts that deviate significantly from formal language registers.

These shifts in the landscape, each in its own way and with its own influence, ushered NLP into the Machine Learning (ML) era (MITCHELL, 1997). With ML, machines can learn to perform tasks based on human knowledge without requiring humans to formally and explicitly structure and represent that knowledge. This proves particularly advantageous for problems where formalizing the required knowledge is difficult or prohibitively expensive. What is important for the ML is to have a sufficiently large and representative set of examples from which the algorithms can learn. In the case of NLP, these examples take the form of text or transcribed speech corpora, which, in line with the empiricist research tradition, represent a specific area of interest within the language (MANNING; SCHÜTZE, 1999; SARDINHA, 2000).

The introduction of ML to NLP was groundbreaking, setting off a rapid pace of new breakthroughs. Deep learning (GOODFELLOW; BENGIO; COURVILLE, 2016), one of the most advanced ML models, boosted the performance of Machine Translation, the first major NLP task, and have been deployed with excellent results

across a wide array of scenarios: chatbots, speech recognition and generation, document generation, sentiment analysis, and the detection of hate speech and fake news, among many others.

In this current NLP paradigm, the linguist's role has expanded to encompass corpus design, which supplies the knowledge to the learning algorithms. There is a trade-off between the corpus provided, the ML algorithm used, and the results obtained. Consequently, the more "informed" the provided examples are, the better the achieved results. Various features can be annotated within texts, encompassing both linguistic analysis (such as morphological, syntactic and semantic features) and other analytical dimensions, like sentiment and opinion analysis.

While NLP advances on the computational front through improvements in machine learning methods, it progresses on the linguistic front through the production of increasingly logical, comprehensive, and consistent annotated corpora, thereby guiding and solidifying linguistic models and theories.

1.2 Corpora at the Heart of NLP

Corpus annotation is a science (IDE, 2017). It begins with the modeling of the phenomenon to be annotated, culminating in the definition of an annotating scheme that comprises a specific set of tags to be assigned to texts segments. Modeling phenomena for annotation is a highly demanding task, consequently, it is common practice within the NLP community to reuse successful annotation schemes.

Languages with advanced NLP resources create annotation models, which are then frequently replicated and adapted for languages with fewer human resources dedicated to model development. This is how, for example, clones and adaptations of FrameNet (BAKER; FILLMORE; LOWE, 1998) and PropBank (PALMER; GILDEA; KINGSBURY, 2005) proliferated, which modeled Fillmore's (1968) theory of semantic

roles in different ways, along with VerbNet (KIPPER-SCHULER, 2005), which modeled Levin's (1993) theory of verb classes.

Although at a given point in history the reuse of schemas was aimed solely at saving modeling effort, it did not take long to realize the great potential that using the same annotation schema could provide for NLP, in terms of facilitating cross-lingual comparison and enabling the development of multilingual NLP tools and applications.

In 2006, an NLP challenge to create a multilingual dependency parser (BUCHHOLZ; MARSI, 2006) required the organizers to adopt a reasonably flexible corpus annotation schema to annotate the thirteen languages involved in the competition. Due to its simplicity, the chosen schema was the one presented in Nivre et al. (2006). The results of training multilingual parsers from these corpora were promising, and by 2015, a "universal" syntactic dependency annotation schema was already being used for several languages. This schema, called Universal Dependencies (UD) (NIVRE et al., 2015, 2020; MARNEFFE et al., 2021), aims to be language-independent and has gained followers in various parts of the world. Over a hundred languages, from the most diverse families, already have corpora annotated in the UD schema.

The availability of corpora annotated in various languages following a single annotation schema opened up the opportunity to develop several parsers whose methods are language-independent. As examples, we can cite UDPipe (STRAKA; STRAKOVA, 2017), UDify (KONDRATYUK; STRAKA, 2019), Stanza (QI et al., 2020), and the spaCy parser (HONNIBAL; JOHNSON, 2015). These parsers, due to their strong learning capacity, even from limited data, are often used to pre-annotate new corpora in the UD format, which are subsequently reviewed by humans. Thus, in a virtuous cycle, with more data available for retraining, the parsers can be continuously improved.

The reuse of an annotation schema does not happen, however, without effort. Pustejovsky, Bunt, and Zaene (2017) distinguish what they call markables [1] from non-markables within an annotation schema. The markables are directly associated with the tagsets (explicit elements in the annotation), while the non-markables are the parts of an annotation schema expressed through guidelines that must be followed to achieve annotation consistency (implicit elements in the annotation). The markables cannot be altered, as they ensure physical comparability among the languages that use them. Without the non-markables, however, there is a risk of assigning the same tags to different phenomena, and this also compromises comparability across languages.

The non-markables are expressed through annotation guidelines, and since these guidelines make use of many examples, they are typically language-dependent. Thus, whether reusing an annotation schema designed for another language or using a "universal" schema, it is essential to instantiate the annotation instructions in the language to which the schema will be applied.

1.3 Essential resources for corpora: guidelines and manuals

The guidelines for a new language aim to make it clear to annotators how the tagset should be used, with rich exemplification covering everything from common and frequent cases to rarer and more difficult-to-annotate cases.

The UD guidelines, because they are not language-specific, leave significant gaps to be filled by the linguists who instantiate them in a given language. A forum maintained by UD and available on GitHub shows how challenging this instantiation is, as hundreds of topics are discussed and diverse opinions compete to resolve questions.

For Portuguese, three corpora have already been made available on the UD website: PUD, GSD, and Bosque-UD, the latter described in Rademaker et al. (2017). It is desirable that new corpora, larger and from new domains, be annotated following

the same model, as this will contribute to improving the performance of Portuguese parsers. However, the lack of an annotation manual hinders initiatives to produce new UD corpora in Portuguese, requiring each new project to face the challenge of instantiating the UD guidelines for the Portuguese language. Furthermore, if each project adopts different decisions to deal with the same challenges in the language, the annotated *corpora* are no longer comparable, and this may prevent them from being "combined" in the effort to train more robust *parsers*.

This was the difficulty we faced when we started a *corpus* annotation project following the UD model: to guide the annotation, we needed a manual containing explanations and examples for each tag in the UD tagsets, as well as discussions on how to resolve ambiguities that frequently raise doubts.

The report produced after the annotation of the Bosque-UD *corpus* (SOUZA et al., 2020) was of great help in starting our work, as it reports several of the problems faced during the annotation of the UD schema in the Portuguese language. However, such a report does not seem to intend to serve as a UD annotation manual, as it is not organized in a way that systematically addresses each tag in the UD tagsets.

Facing this gap, we undertook the construction of a UD *corpus* annotation manual, in which the UD guidelines are instantiated, encompassing design decisions and studies on highly ambiguous words that, during the annotation process, proved to be a source of doubt and disagreement among annotators. In this article, we explore in detail the issues related to the development of this manual.

The article is organized into six sections in addition to this introduction. In Section 2, we provide a brief review of *corpus* syntactic annotation for the Portuguese language and its relationship with the development of Portuguese *parsers*. We dedicate Section 3 to a discussion of the reasons that led the NLP community to adopt Tesnière's theory (1959, 2015) as a valuable framework for syntactic annotation. In Section 4, we present the annotation scheme for the UD. In Section 5, we present the sets of

morphosyntactic labels (5.1) and syntactic relations (5.2) for the UD, relating them to the morphosyntactic classes and syntactic functions of the normative grammars of Portuguese. In Section 6, we discuss which of the UD guidelines can be easily adoptable in Portuguese, which ones are problematic and required decision-making during the project based on linguistic and computational principles, and which ones required arbitrary decisions made solely to maintain consistency in annotation, but that may be changed in the future if a better solution becomes available. In Section 7, we elaborated concluding remarks and discussed future work.

2 Portuguese parsers and treebanks

For a long time, it was possible to build NLP systems that did not rely on syntactic analysis. Almost all knowledge was derived from lexical items and their relative positions (even if this, indirectly, mirrored syntax). For many simpler, more limited applications, such as early spell-checkers, this was sufficient. For other, more advanced applications, especially those involving any level of content interpretation, such as question and answer systems, syntactic analysis has proven to be essential. As work on the NLP tasks progressed, syntactic analysis became a critical issue, since a whole series of subsequent processing steps depended on its quality.

In the 1990s, statistical *parsers* marked the beginning of a new era: of *parser* generation through machine learning. To train these parsers, syntactically annotated corpora were needed, preferably reviewed by humans, as this would ensure that the model would not learn from annotation deviations. In order to facilitate machine learning, practically all annotation used a tree-based format, whether for constituents or dependencies, which is why syntactically annotated *corpora* are called *treebanks*.

The strategy was as follows: the corpora were annotated using existing rule-based parsers; they were then reviewed by humans and made available as examples for machine learning.

The first widely known and used Portuguese parser in the field of natural language processing (NLP) in both Brazil and Portugal was *Palavras* (BICK, 2000), and its first commercial version dates back to the mid-1990s. It consists of hundreds of rules written by the author that follow the *Constraint Grammar* paradigm (KARLSSON, 1990). Initially, *Palavras* only generated dependency trees, but starting in the 2000s, it also began generating constituency-based parse trees. And it was this *corpus* that was chosen to pre-annotate the *Floresta Sintá(c)tica* (AFONSO et al., 2002) *corpus*, the first successful initiative to create a Portuguese-language *treebank*.

At the time of its release, only 10% of the *Floresta Sintá(c)tica corpus* had been reviewed; however, all the deviations found in the reviewed portion were used to improve the rules of the *Palavras parser*. Since the unreviewed portion of the *Floresta corpus* was reannotated using the improved version of *Palavras*, it can be said that the review benefited the entire *treebank*. Since *Floresta Sintá(c)tica* contains fully revised sections, partially revised sections, and unrevised sections, these sections have been given distinct names.

O Bosque is the part of the *Floresta Sintá(c)tica corpus* that has been completely revised. It contains 9,368 sentences extracted from journalistic *corpora*, about half of which are in Brazilian Portuguese (CetenFolha) and half in European Portuguese (Cetempúblico), and to date it remains the only reviewed *treebank* that includes Brazilian Portuguese. *Selva*, whose tree structure has been partially reviewed, contains approximately 30,000 sentences and includes samples of spoken language as well as separate collections of literary and scientific texts. *Floresta Virgem* has approximately 96,000 unreviewed sentences. And finally, *Amazônia*, which was added later, contains approximately 275,000 unedited sentences extracted from blogs; therefore, it is a portion of the *treebank* that differs from the others because it consists of content produced by web users and features colloquial language register. (FREITAS; ROCHA; BICK, 2008).

In 2016, *Palavras* began to generate output in the form of dependencies in UD format (BICK, 2016). This format converter was applied to the *Bosque corpus*, and the result served as the basis for the manual review that led to the creation of the *Bosque-UD corpus* (RADEMAKER et al., 2017). The *Bosque-UD corpus*, available on the UD website, has been used to train dependency *parsers* that employ language-independent techniques, which had already demonstrated strong learning capabilities in other languages (such as the aforementioned UDPipe, UDify, SpaCy, and Stanza). The advantage of these *parsers* is that they are open-source (unlike *Palavras*) and can therefore be retrained by other teams using new *corpora* and across different genres.

Other *parsers* and *treebanks* have emerged in recent years, mainly through the efforts of the Portuguese linguistic community, such as the LX-Parser (SILVA et al., 2010) and the Cintil *treebank*, but *Palavras* has remained widely used. Its role in Portuguese processing is undeniable, and thanks to it, we were able to make the transition from rule-based parsers to ML-based *parsers*. Similarly, the role of the *Bosque corpus*, the first Portuguese *treebank* whose syntax tree have been fully reviewed by humans and is therefore ideal for machine learning, is undeniable.

Although there are currently three *corpora* available in UD format on the UD website, only *Bosque* consists of sentences originally written in Portuguese. The PUD *corpus* contains only 1,000 sentences and is a Portuguese version of the parallel *corpus* produced for the CoNLL 2017 competition for multilingual *parsers*. The GSD *corpus* contains 12,078 sentences in Brazilian Portuguese, resulting from the conversion to UD of the annotations from the Google *corpus*, a parallel *corpus* of translations into various languages.

When we compare the size and variety of genres in the *treebanks* we have with those available in languages that are better supported by NLP resources, we see that there is much work to be done in Portuguese, and especially in Brazilian Portuguese.

3 The NLP and the syntactic structure dependency

Dependency-based parse trees competed with the constituency-based parse trees for a while in the context of the NLP. However, when syntax trees started being used to make automatic language comparisons, dependency-based syntax trees were found to be more adequate. This is mainly due to the fact that the dependency-based structures allow for syntactic relations between words, regardless of the position which they occupy in the sentence, unlike the constituency-based method, more rigid in regards to the order of its constituents.

This aspect of dependency-based structures is key when relating configurational languages (such as Portuguese, in which the order of the constituents is relevant, due to the fact that the syntactic category is largely based on the sentence's word order) to non-configurational languages (such as Latin, in which the constituent's order is free, due to the fact that the syntactic role of words is designated by declensions).

Furthermore, when discussing machine learning techniques, the dependency structure allows for a more lexical study (since it's based on tokens, instead of syntagmas), which has shown to be advantageous in various applications of NLP.

For those reasons, dependency grammars (inspired by Tesnière (1959)), were revalued as a promising NLP model.

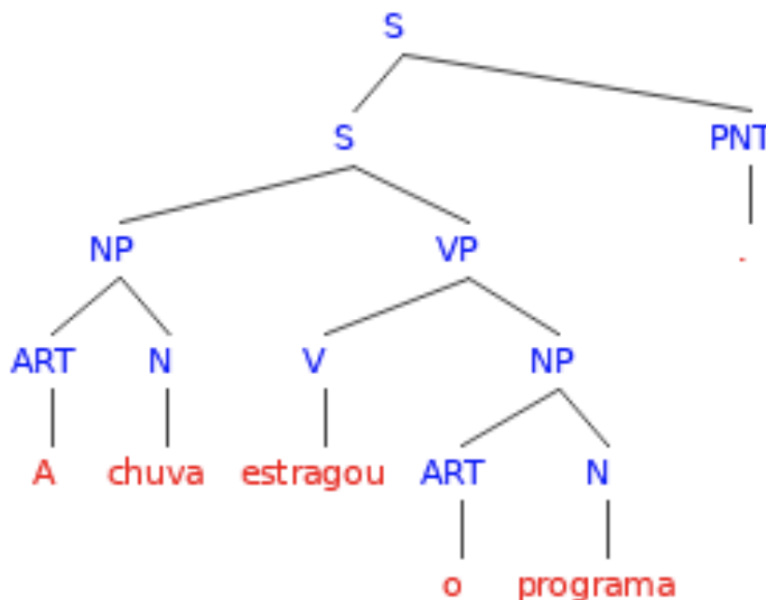
Tesnière (1959) rejected the dualist view of the syntactic structure in subject and predicate. In his view, the verb is the root of the syntactic structure and subject and object have a subordinate role. He realized that a sentence is not only comprised of its constituent words (explicit elements), but also implicit elements between those words, relations that maintained a hierarchy (a ruler and a subordinate) and a vector (that shows which element is ruling the interaction). In order to make these relations explicit, Tesnière (1959) introduced the idea of representing syntax with a tree, which

he called *stemma*, in which all elements of a sentence and its dependencies connecting those elements are represented in a single structure.

The dependency trees are the core of the UD annotation scheme, whose ultimate goal is to benefit *corpora* annotation which will serve as material for the development of multilingual NLP applications. The UD initiative, thus, incited the dissemination of the late Tesnière's work in English (TESNIÈRE¹, 2015).

In order to illustrate this, we present an example of syntactic representation through a constituency-based *parse* tree (Figure 1) and one with a dependency-based *parse* tree (Figure 2). In comparing both trees, it is possible to notice that the second one is simpler, as it contains less nodes in total, creating a more direct relationship between tokens.

Figure 1 — An analysis of “A chuva estragou o programa” utilizing a constituency-based parse tree.

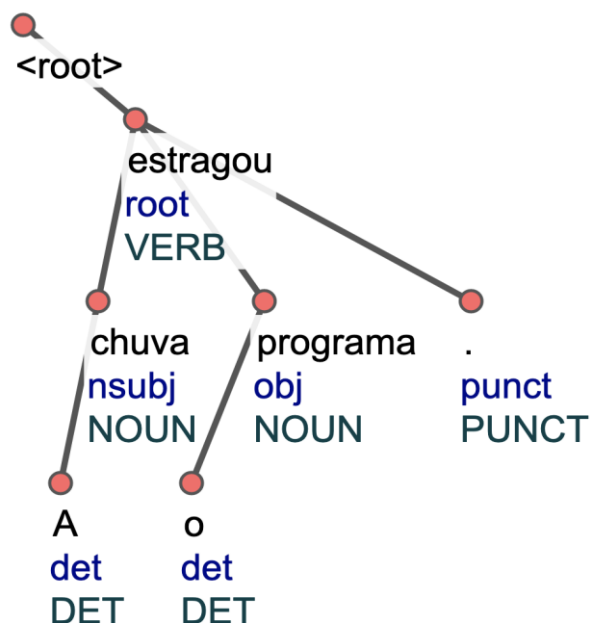


Source: syntax tree generated using the parser LX-Parser².

¹ 2015's English translation is commented in the light of the theory's current application.

² <https://portulanclarin.net/workbench/lx-parser/>

Figure 2 — An analysis of “A chuva estragou o programa” utilizing a dependency-based parse tree.



Source: syntactic tree in UD format generated using the UDPipe LINDAT/CLARIN³.

4 The UD model

The UD annotation scheme is a lexicalist one, which means that all labels are directly associated to tokens. In a dependency relation, only one token is the head and only one token is the dependent. All tokens, with no exceptions, are assigned a morphosyntactic label and engage in at least one dependency relation. UD defines that only content words (nouns, pronouns, verbs, adjectives, adverbs and numbers) can be the head of relations (those can also be subordinate in relations). On the other hand, functional words (auxiliaries, prepositions, determinants and conjunctions), symbols and punctuation could only be subordinate in relations.

The corpus to be annotated must be previously processed in order to: 1) separate the text in sentences (since each sentence is annotated individually) and 2) tokenize, which includes separating punctuations attached to words (except periods

³ <http://lindat.mff.cuni.cz/services/udpipe/run.php>

used in abbreviations, periods that separate groups of thousands, decimal periods, and hyphenated compounds), separate verbs from *pronomes enclíticos* (pronouns attached to the end) and *mesoclíticos* (attached to the middle) eliminating its respective hyphens and separating the tokens that constitute contractions (e.g. *deste* = *de este*, *comigo* = *com mim*, *naquela* = *em aquela*). Hyphenated compounds are usually accepted as language tokens, but tokens that incorporate multiple words, such as *panela de pressão* (pressure cooker) and *energia solar* (solar energy) are annotated as independent tokens. UD does not work with the concept of compositional phrases at the morphosyntactic level, and at the syntactic level, there are three dependency relations in order to annotate compositional phrases that present no syntactic relations between its constituent tokens.

The UD annotation scheme follows a format established at the Conference on Computational Natural Language Learning (CoNLL) and is, thus, called CoNLL-U (the U stands for "universal"). Various annotation tools with a user-friendly interface are available for annotation in that format ⁴, making the project creation process uncomplicated. It is important, however, to understand the format's constituent parts. Although its goal is the annotation of syntax dependency relations, CoNLL-U allows for a logging of much more information. This is due to the fact that, the richer the annotation, the more attributes can contribute to the distinction between labels in machine learning.

The CoNLL-U format is comprised of a 10-column table with information associated to tokens in each lines. The most important columns for the annotator are the forth, called UPOS (filled with one of the 17 morphosyntactic UD labels), and the eight, called DepRel (filled with one of the 37 syntactic dependency relation UD labels). The remaining column contain: position identifier for each of the tokens, the form, lemma, morphological attributes (e.g. inflection), the head of the dependency relation

⁴ <https://universaldependencies.org/tools.html>

and three optional column. This format, while restricting content for most columns, formatting the annotation in accordance to the UD approach, allocates some space for annotation of language-specific characteristics (the fifth column, XPOS, is reserved for language-specific morphosyntactic labels, and the tenth column, MISC, is meant to be freely used). Table 1 shows an example of a CoNLL-U formatted sentence.

Table 1 - example of a CoNLL-U formatted sentence.

# text = O estúdio também divulgou novo poster de o filme .									
id.	forma	lema	UPOS	XPOS	Atributos morfológicos	head	DelRel	Deps	MISC
1	O	o	DET	—	Definite=Def Gender=Masc Number=Sing PronType=Art	2	det	—	—
2	estúdio	estúdio	NOUN	—	Gender=Masc Number=Sing	4	nsubj	—	—
3	também	também	ADV	—	—	4	advmod	—	—
4	divulgou	divulgar	VERB	—	Mood=Ind Number=Sing Person=3 Tense=Past VerbForm=Fin	0	root	—	—
5	novo	novo	ADJ	—	Gender=Masc Number=Sing	6	amod	—	—
6	poster	poster	NOUN	—	Foreign=Yes Gender=Masc Number=Sing	4	obj	—	—
7	de	de	ADP	—	—	9	case	—	—
8	o	o	DET	—	Definite=Def Gender=Masc Number=Sing PronType=Art	9	det	—	—
9	filme	filme	NOUN	—	Gender=Masc Number=Sing	6	nmod	—	—
10	.	.	PUNCT	—	—	4	punct	—	SpacesAfter=\s\s\n

5 The Relationship Between the UD Framework and Prescriptive Grammar

Looking only at the set of 17 morphosyntactic labels (part-of-speech tags, or simply PoS tags) and 37 syntactic relations in Universal Dependencies (UD), one might rely on the knowledge acquired through the study of prescriptive grammars and venture to assign a large portion of these labels to the words in a text, since the names of the labels often coincide with terms used in the Brazilian Grammatical Nomenclature (HENRIQUES, 2009). However, the UD guidelines show us that interpreting these labels cannot always be done straightforwardly on the basis of prior grammatical knowledge alone. To illustrate how UD labeling concepts differ from the concepts largely spread by the prescriptive grammars, we will discuss morphosyntactic labels and dependency-relation labels in two separate sections (just as we addressed the guidelines for assigning these two sets of labels in two separate manuals).

Although presented in different sections, it is important to highlight that there is a high correlation between morphosyntactic and syntactic-relation labels in UD. In

Section 5.2 we will reveal how this correlation is translated into a set of restrictions that syntactic relations impose on morphosyntactic labels.

We stress that this article exemplifies only some of the issues covered by the guidelines included in our annotation manuals. The manuals were developed from initial annotation experiences and, because they are managed by the same individuals who annotate the corpus, they undergo periodic revisions in order to reflect new facts observed in the corpus that had not yet been contemplated in the guidelines⁵.

5.1 UD Morphosyntactic Labels

The standard UD morphosyntactic labels are:

- For open classes: ADJ (adjectives), ADV (adverbs), INTJ (interjections), NOUN (nouns), PROPEN (proper nouns), and VERB (verbs);
- For closed classes: For closed classes: ADP (adpositions), AUX (auxiliary verbs), CCONJ (coordinating conjunctions), DET (determiners), NUM (cardinal numerals), PART (particles), PRON (pronouns), and SCONJ (subordinating conjunctions);
- Other tokens that do not fit into open or closed classes: PUNCT (punctuation), SYM (symbols), and X (other cases, including onomatopoeias and foreign-language words).

The main differences between UD morphosyntactic labels and the morphosyntactic classes of prescriptive grammar for Brazilian Portuguese are as follows:

- DET: UD operates with the concept of determiner and groups under this label articles and non-nominal pronouns (demonstratives and possessives);

⁵ The version of the manual showcased in this article is based on the UD guidelines effective in April 2022. In the event of publishing of new UD guidelines they shall be incorporated into future versions of the manual.

- NUM: only cardinal numerals should be annotated as NUM, whereas ordinal numerals should be annotated as ADJ;
- PRON: only nominal pronouns are annotated with this label, while other pronouns, provided that they modify a nominal, should be annotated as DET;
- NOUN: only common nouns are annotated under this label, since proper names are annotated as PROPEN;
- VERB: only full verbs capable of being classified as verbal predicates should receive this label;
- AUX: auxiliary verbs and “highly grammaticalized” copular verbs (verbs without significant semantic content) should receive this label;
- PROPEN: a label intended for proper names, provided that they do not coincide with common lexical items of the language.

In our PoS-tag annotation manual (morphosyntactic labels), each label is treated in a separate section containing definitions, examples, lexical inventories (in the case of closed classes), and disambiguation guidelines concerning other labels commonly confused with it in certain contexts. Ambiguity between morphosyntactic labels is, to a large extent, language-dependent, and is therefore not addressed in the general UD guidelines. For example, the word *menos* (less), when preceding a noun with the meaning “a smaller quantity of,” is tagged as ADJ because it modifies a NOUN. However, when the word *menos* modifies a VERB, an ADJ, or an ADV, with the meaning “with less intensity” or “to a lesser degree,” it is tagged as ADV. There are, however, contexts in which *menos* occurs between a VERB and a NOUN, potentially confusing an inexperienced annotator. Example:

O plano favoreceu menos classes D e E do que classe média.

(The plan benefited lower-income classes D and E less than the middle class.)

(Here, *menos* (less) is an ADV modifying the verb *favorecer* (to benefit): “benefited classes D and E to a lesser degree.”)

O plano atual favoreceu menos pessoas que o plano anterior.

(The current plan benefited fewer people than the previous plan.)

(Here, *menos* (fewer) is an ADJ modifying the noun *pessoas* (people): “a smaller number of people.”)

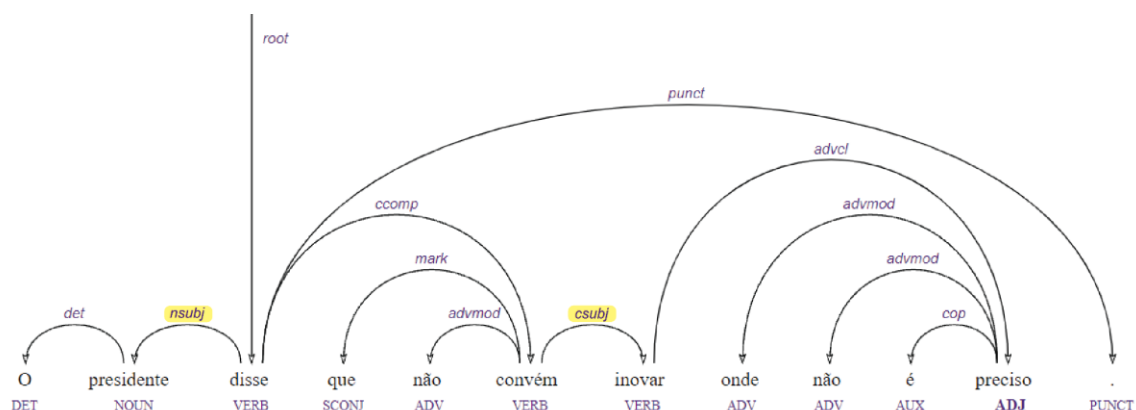
5.2 UD Dependency Relations

UD includes a set of 37 labels for marking dependency relations⁶, and the mismatch between these relations and the components of syntactic analysis described in prescriptive grammars is even greater than that observed for the set of morphosyntactic labels. In the following, we discuss the main differences between the UD syntactic annotation scheme and the syntactic functions found in prescriptive grammars.

UD has separate relations for syntactic functions filled by clauses (both verbal and nominal predicates) and the same syntactic functions filled by words serving as phrase heads. Thus, for example, a subject may be labeled *nsubj* if it is nominal, or *csubj* if it consists of a clause, as can be seen in Figure 3.

Figure 3 — Example of **nsubj** and **csubj**.

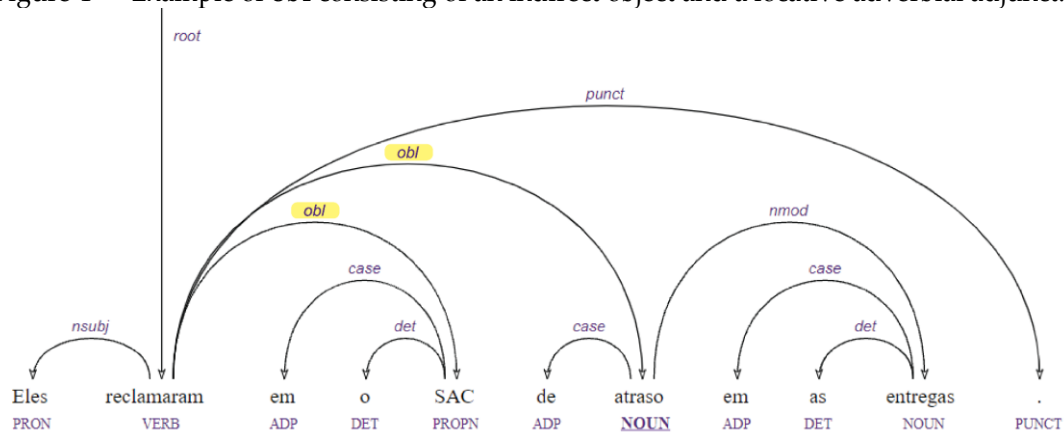
⁶ A complete chart of UD dependency relations is available in <https://universaldependencies.org/u/dep/index.html>



Source: syntactic tree in UD format produced using the Arborator-Grew software⁷.

In UD, adverbial adjuncts whose head is a NOUN (mostly introduced by prepositions) are not distinguished from preposition-introduced indirect objects, since both are considered nominal verbal modifiers and are annotated using the relation **obl** (illustrated in Figure 4).

Figure 4 — Example of **obl** consisting of an indirect object and a locative adverbial adjunct.

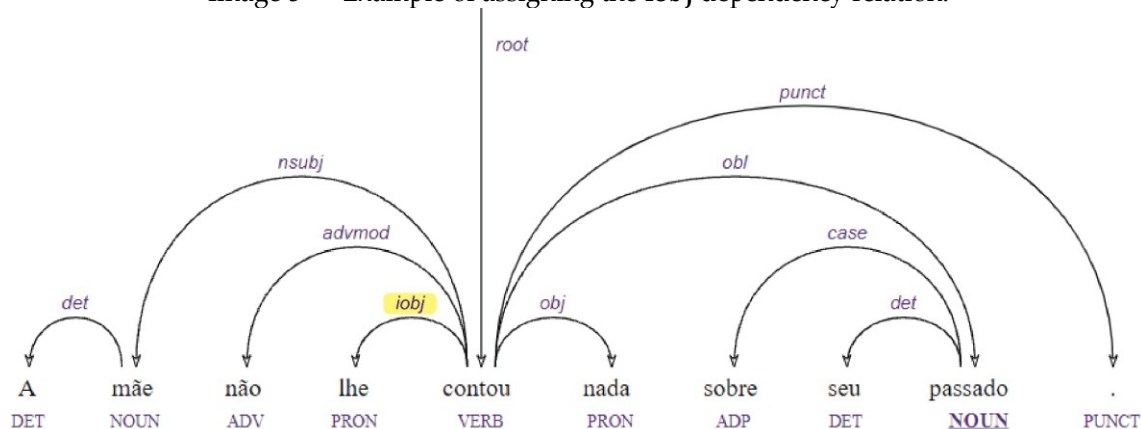


Source: syntactic tree in UD format generated using Arborator-Grew software.

⁷ <https://arborator.icmc.usp.br/#/>

Perhaps one of the easiest UD relations to misinterpret is *iobj* (indirect object). This relation should only be assigned to non-prepositional indirect objects, which is counterintuitive if we consider the meaning of “indirect” in the name of this syntactic function. This is a remnant of the origins of these relations in English-language annotation, since English allows two non-prepositional objects⁸, both of which may be promoted to subject position in the passive voice construction. In Portuguese, the relation ***iobj*** should only be used to annotate indirect objects realized as clitic pronouns (me, te, se, lhe, nos, vos, lhes), as shown in Figure 5.

Image 5 — Example of assigning the ***iobj*** dependency relation.



Source: syntactic tree in UD format generated using Arborator-Grew software.

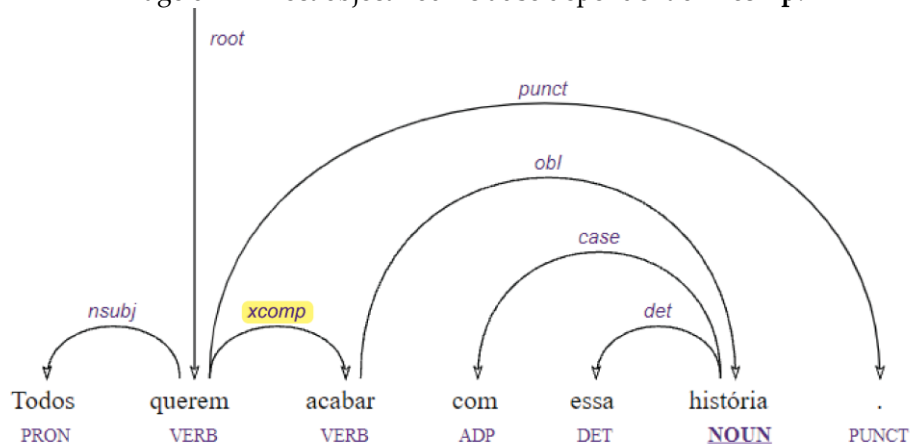
In the UD scheme, direct objects are annotated with the “***obj***” relation, but if they are expressed through a clause, they get two different types of annotation. Direct object noun clauses that present null subjects⁹, governed by the subject or object of the main clause, are annotated as dependent on the “***xcomp***” relation (Image 6), e those

⁸ Two non-prepositional objects may occur on English dative verbs. In the phrase “John sent Mary a message”, both objects can function as subjects of the passive voice: “A message was sent to Mary” and “Mary was sent a message”.

⁹ For a more in-depth discussion of the concepts of the null subject and *xcomp*, we recommend reading Bresnan (1982).

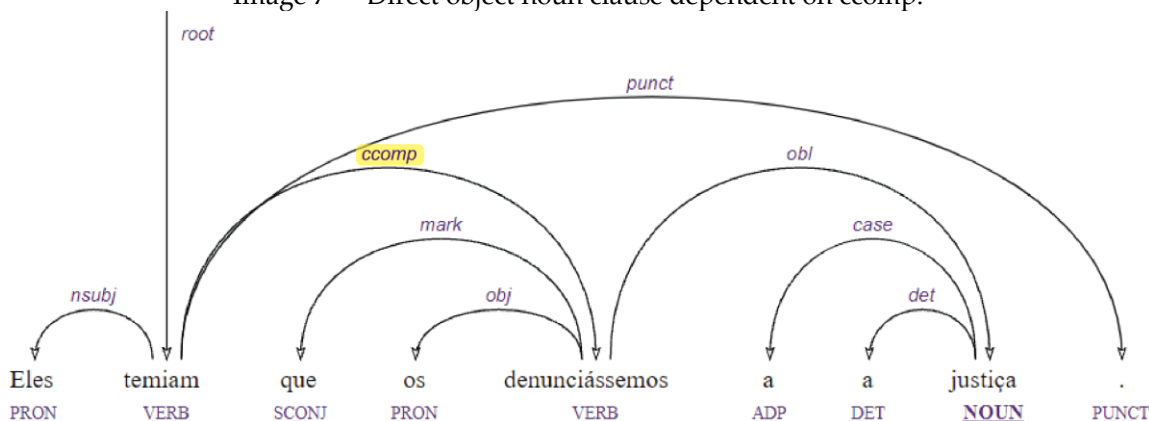
with realizable subjects (even if elliptical), are annotated with the “**ccomp**” relation (Image 7).

Image 6 — Direct object noun clause dependent on **xcomp**.



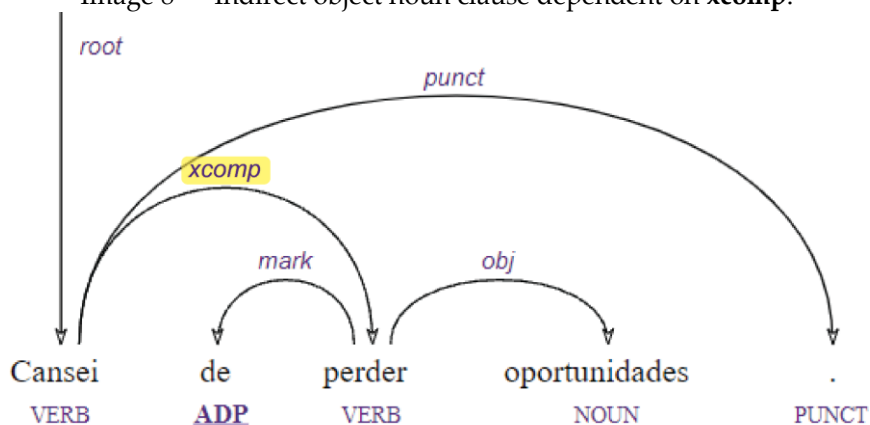
Source: syntactic tree in UD format generated using Arborator-Grew software.

Image 7 — Direct object noun clause dependent on **ccomp**.

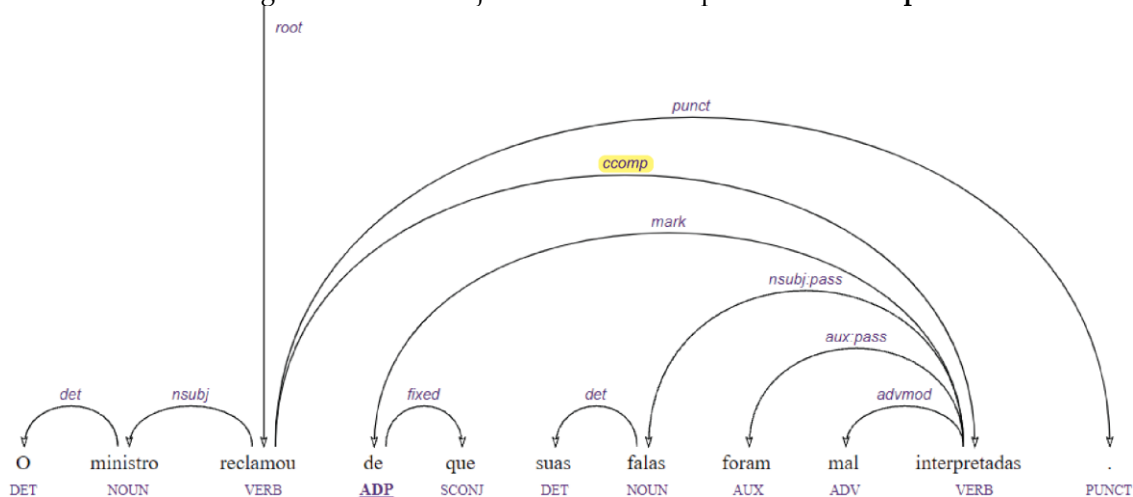


Source: syntactic tree in UD format generated using the Arborator-Grew software.

The same thing applies to the indirect objects realized in the form of clauses (indirect object noun clause): if the subordinated clause presents a null subject governed by the subject or object of the main clause, it will be dependent on **xcomp** (Image 8) and, if it presents a realizable subject, it will be dependent on **ccomp** (Image 9).

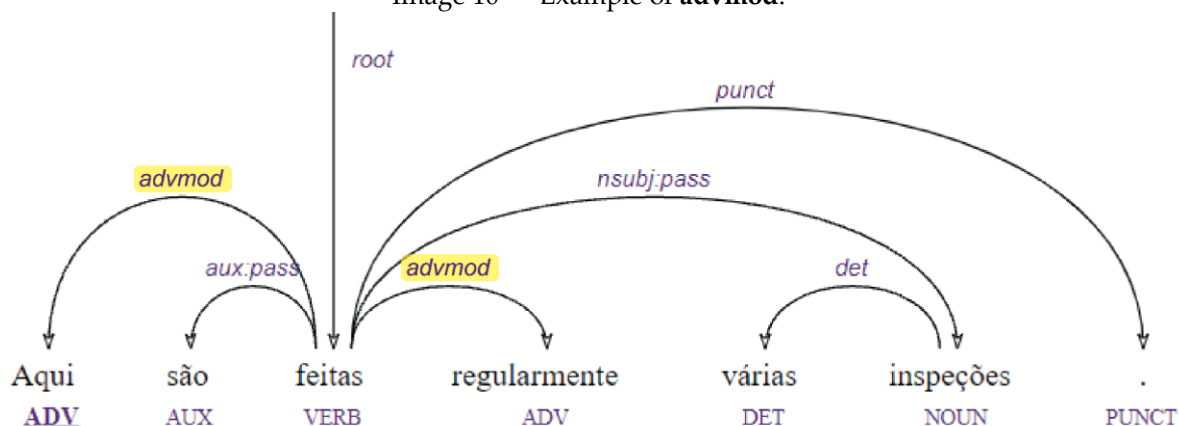
Image 8 — Indirect object noun clause dependent on **xcomp**.

Source: syntactic tree in UD format generated using the Arborator-Grew software.

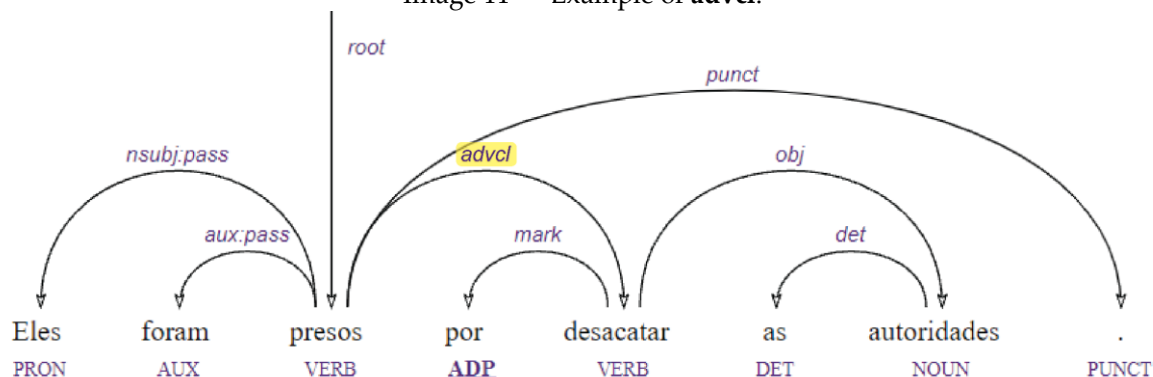
Image 9 — Indirect object noun clause dependent on **ccomp**.

Source: syntactic tree in UD format generated using the Arborator-Grew software.

The modifiers consisting of adverbs (ADV) are annotated with the **advmod** relation (Image 10). However, when addressing an adverbial clause, it will be denoted an **advcl** relation (Image 11).

Image 10 — Example of **advmod**.

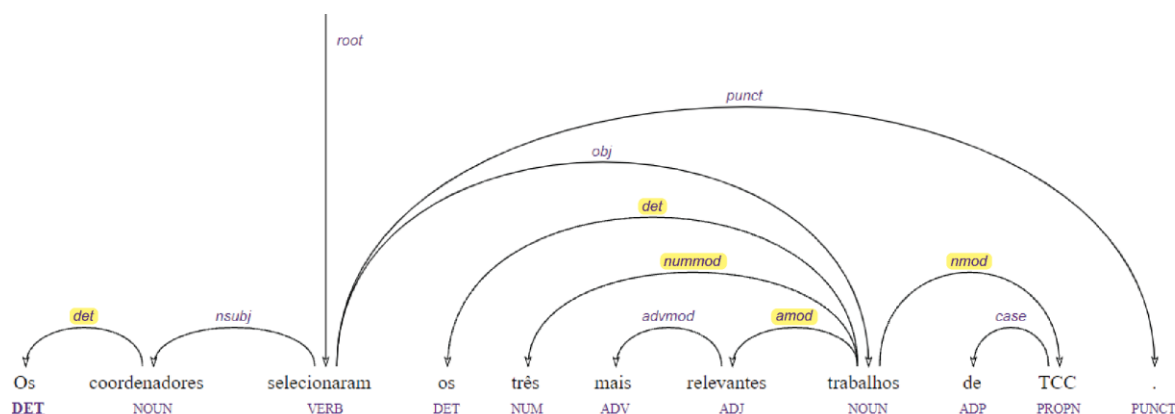
Source: syntactic tree in UD format generated using the Arborator-Grew software.

Image 11 — Example of **advcl**.

Source: syntactic tree in UD format generated using the Arborator-Grew software.

The UD scheme also allows for greater granularity in nominal modifiers than prescriptive grammars. Nominal modifiers can be: **amod** (for ADJ), **nmod** (for NOUN and PROPN) and **nummod** (for NUM). Prepositions (ADP) that introduce **nmods** are linked by the **case** relation to the *head* nominal of the phrase. Determiners (DET) are linked to nominals they modify through the **det** relation, as shown in Image 12.

Image 12 — Examples of nominal modifiers: **det**, **nummod**, **amod**, **nmod**.

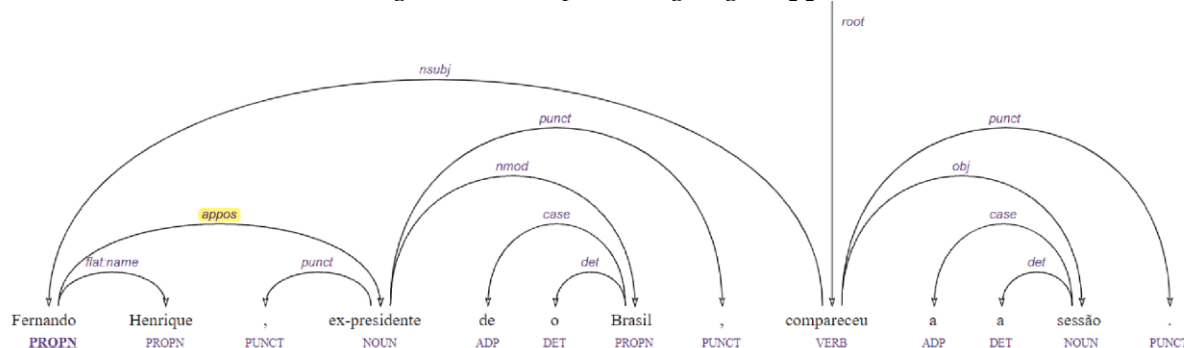


Source: syntactic tree in UD format generated using the Arborator-Grew software.

If, nevertheless, a nominal modifier takes the form of a clause, the dependency relation will always be **acl**, because the UD does not distinguish between nominal complement clauses and adjectival clauses.

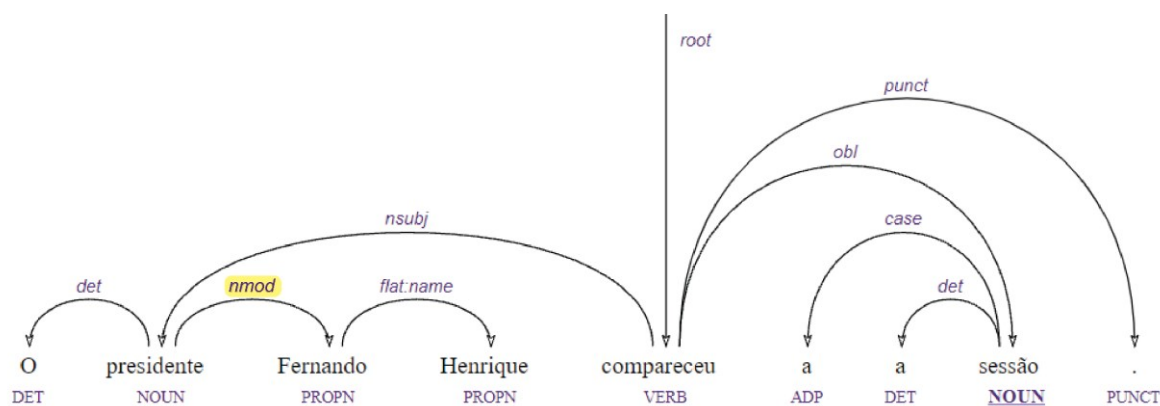
Another difference worth noting is that of the concept of appositions in the UD. According to the UD instructions, only the relation between two interchangeable elements that are independent and whose order can be swapped without affecting the grammaticality of the sentence should be annotated as **appos**, as shown in Image 13. Cases such as the one shown in Image 14, however, traditionally treated as appositions, are annotated as an attribute in the UD, **nmod**.

Image 13 — Exame of assigning of **appos**.



Source: syntactic tree in UD format generated using the Arborator-Grew software.

Image 14 — Attribute **nmod** which is commonly confused with **appos**.

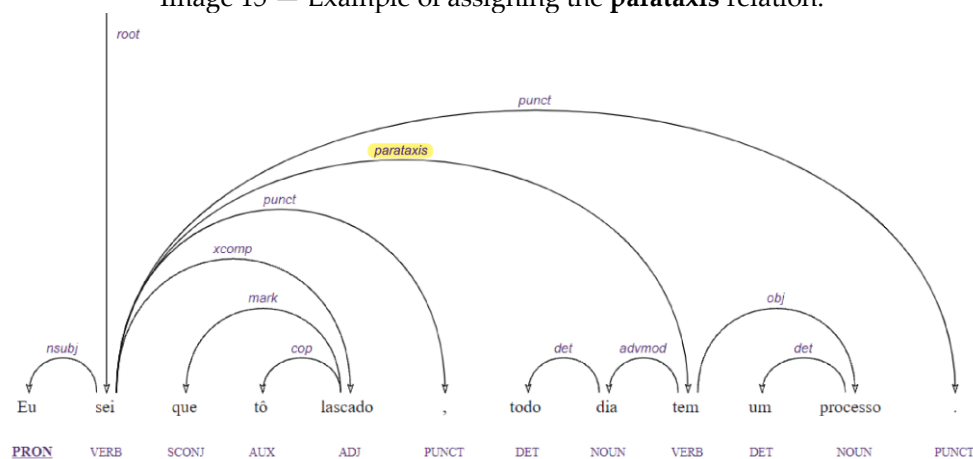


Source: syntactic tree in UD format generated using the Arborator-Grew software.

Along with dependency relations associated with syntactic functions of descriptive grammar, the UD provides a series of relations to link sentence elements that have no syntactic relation with other elements, such as **parataxis**, **discourse**, **reparandum**, **dislocated** and **intj**.

The **parataxis** relation (shown in Image 15), for instance, is used to link two clauses without explicit logical cohesion within the same sentence.

Image 15 — Example of assigning the **parataxis** relation.



Source: syntactic tree in UD format generated using the Arborator-Grew software.

The UD guidelines also define three dependency relations to annotate multi-word expressions that have no syntactic relation with one another (**compound**, **flat** and **fixed**). These labels should be used sparingly, as they merely indicate that words

linked by them have no syntactic relation to one another, which makes them opaque for the parser to use in learning syntax with them. Example:

João Paulo - PROPN PROPN linked by **flat** relation
in addition - ADV ADP linked by **fixed** relation
going away - VERB ADV linked by **compound** relation

It cannot be said, therefore, that the UD guidelines determine the annotation of all multi-word expressions, since those that have syntactic relations between their constituent tokens are annotated with common relations. Example:

greenhouse effect is a NOUN NOUN phrase linked by the **nmod** relation, in which the head is “*effect*” and the dependent (modifier) is “*greenhouse*”
sun lotion is a NOUN ADJ linked by the **amod** relation, where the head is “*lotion*” and the dependent (modifier) is “*sun*”
bureau de change is a NOUN ADP NOUN linked by the relations **nmod** (between *bureau* and *change*) and **case** (between *change* and *de*)

Something worth noting in the examples presented in this section is the fact that some syntactic relations only accept constituents with a specific morphosyntactic label: **det** only accepts DET, **case** only accepts ADP, **advmod** only accepts ADV, **nmod** only accepts NOUN or PROPN, **amod** only accepts ADJ, **nummod** only accepts NUM, **aux** only accepts AUX, **cop** only accepts AUX, **punct** only accepts PUNCT, **intj** only accepts INTJ and **mark** only accepts SCONJ and ADP. This illustrates how the UD automatically restricted the annotation, preventing most labeling errors from occurring.

6 The challenges of the instantiation of the UD guidelines and the construction of the annotation manual

The task of producing two annotation manuals (one for morphosyntactic labels and one for dependency relation labels) started with a thorough reading of the UD guidelines, supplemented by reading other reference sources about the assigning of UD labels, such as articles and discussion forums available on the UD github. We carried out the work in two stages, since the annotation of morphosyntactic labels would be performed prior to the dependency relation annotation (which is why we organized the guidelines in two manuals).

Before starting the annotation of each stage, we already had a preliminary version of the respective manual, which were used to train the annotators. These preliminary versions of the manuals were prepared by a linguist and reviewed by a NLP researcher (who is not a linguist) who is involved in the annotation process. The team of annotators was subsequently expanded, and the people responsible for the preliminary version of the manual remained on the team and participated in the annotation process.

During each stage of the annotation process, it became clear that the definitions needed to be improved and the instructions needed to be more detailed. We use the discrepancies among the annotators to help improve the manuals, and we also use examples from the *corpus* to illustrate issues that proved challenging for annotation.

Another source of reference was the *corpora* already annotated in Portuguese and other languages using the UD model. To do this, we used the Grew-Match search engine¹⁰, which allows searches by words, headwords, morphosyntactic tags, and dependency relationships. This practice was important both for identifying solutions used in other *corpora* that could be applicable to our *corpus* and for realizing that 1)

¹⁰ <http://match.grew.fr/>

different *corpora* of the same language sometimes use different ways of assigning the same label or relation; 2) *corpora* of different languages may use different tags and relations to annotate similar phenomena; and 3) often a pattern is not recognized by annotators, and it ends up being annotated in different ways throughout the same *corpus*, leading to inconsistency.

While compiling the manuals, we encountered three types of situations: 1) those in which the UD guidelines were clear and easily applicable in Portuguese; 2) those in which there were gaps in the UD guidelines, but which, by following linguistic principles, we were able to resolve with clear decisions; and 3) those in which the UD guidelines were unclear or had gaps, and the linguistic justification was insufficient to guarantee a decision that could be considered “correct,” situations in which we had to make arbitrary decisions in order to ensure consistency in the annotation.

For example, Portuguese UD corpora list different numbers of verbs with auxiliary verbs: the PUD lists 13, the GSD lists 59, and the Bosque corpus lists 6 verbs¹¹. In other languages, there is also considerable variation in the use of this label. In English, for example, modals such as *would*, *should*, *can*, *could*, *may*, *might*, and *must* are marked, while in French, the causative verb *faire* is also marked as AUX.

The issue of auxiliary and copular verbs was one that required us to make an arbitrary decision (and a provisional one until a better solution is found), since the UD guidelines leave it up to each language to make that decision. The only recommendation of the UD is that only highly grammaticalized auxiliary verbs and copular verbs that have lost their semantic meaning should be classified as AUX. Currently, we are classifying as AUX only the auxiliary verbs for tense and the passive voice, as well as the copular verbs “*ser*” and “*estar*”.

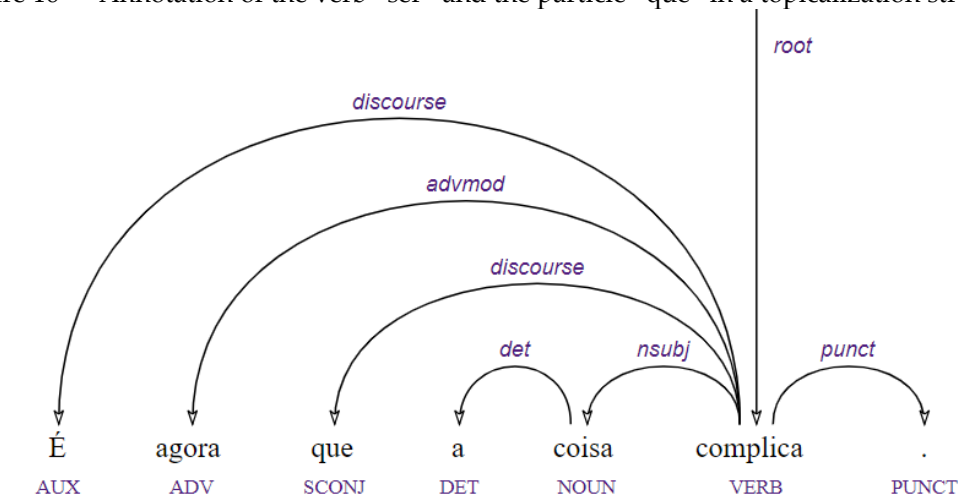
One issue that required a decision based on linguistic criteria was the annotation of prepositions that introduce reduced infinitive clauses as ADP rather

¹¹ For a detailed comparison: <https://universaldependencies.org/treebanks/pt-comparison.html>

than SCONJ. The UD simply states that SCONJ is used for *complementizers* (in Portuguese, “*que*” and “*se*”) and words that introduce adverbial clauses (such as “*quando*” in temporal clauses, and “*como*” in causal and adverbial clauses). According to Azeredo (2013, item 291), subordinating conjunctions are used to introduce full clauses rather than reduced clauses; this has led us to conclude, therefore, that prepositions introducing reduced infinitive clauses should continue to be marked as ADP, even though the corresponding syntactic relationship is **mark**. This annotation may prove useful in future tasks involving the annotation of semantic roles, as prepositions that introduce noun phrases or infinitive clauses provide important clues for identifying semantic roles within a verb’s argument structure.

The annotation of the verb “*ser*” and the conjunction “*que*” in a topicalizing function was a case in which we replicated the arbitrary decision made in the Bosque-UD corpus: we annotated the verb as AUX, the conjunction as SCONJ, and the relationship between the two and the predicate as **discourse**.

Figure 16 — Annotation of the verb “*ser*” and the particle “*que*” in a topicalization structure.



Source: syntactic tree in UD format generated using the Arborator-Grew software.

One instance where we did not blindly follow the UD guidelines was in the spelling of proper names. The UD stipulates that proper nouns that function as common nouns and adjectives in the language should be tagged with the labels

corresponding to their original categories, and that their syntactic relationships should be tagged using the syntactic relations applied to common words. Thus, “*Maria das Dores*” would consist of a proper noun, an adverbial phrase, a determiner, and a noun, while “*Companhia de Água e Esgoto*” would consist of a noun, an adverbial phrase, a noun, a compound noun, and a noun. However, we took a conservative approach and treated all words written in uppercase as PROPNS, linking them using the **flat:name** relationship, as we did not want to restrict NLP tasks that depend on the *parser’s* output, such as named entity recognition.

7 Concluding Remarks and Future Research

As noted above, the work of developing the annotation manual for the UD model in a Portuguese *corpus* required considerable effort and constant refinement. For practical and instructional reasons, we have divided the guidelines into two manuals: one for annotating morphosyntactic tags (*PoS Tag Annotation Manual*) and another for dependency relations (*Dependency Relation Annotation Manual*). These two complementary manuals are available on the POeTiSA¹² project website and can help streamline the effort required for new projects involving syntactic annotation of corpora using the UD model. New versions of these manuals will be published on the same website whenever updates to the content are necessary.

Since our project involves annotating other *corpora* across different domains, it is likely that new guidelines will be added to the existing material, for example, in the form of appendices dedicated to specific domains.

¹² <https://sites.google.com/icmc.usp.br/poetisa>

Our thank you

The authors would like to thank the Center for Artificial Intelligence at the University of São Paulo (C4AI-<http://c4ai.inova.usp.br/>), funded by IBM and FAPESP (grant #2019/07665-4), for its support.

References

AFONSO, S.; BICK, E.; HABER, R.; SANTOS, D. Floresta sintá(c)tica: a treebank for Portuguese. *In*: RODRÍGUEZ, M. G.; ARAUJO, C. P. S. (ed.), **Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)**. European Language Resources Association, 2002. p. 1698-1703.

AZEREDO, J. C. S. **Fundamentos de Gramática do Português**. Rio de Janeiro: Jorge Zahar Editores, 2013. E-book.

BAKER, C. F.; FILLMORE, C. J.; LOWE, J. B. The Berkeley Framenet Project. *In*: **36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics**, Vol. 1. Quebec: Association for Computational Linguistics, 1998. p. 86-90.

BICK, E. **The Parsing System PALAVRAS**: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press, 2000.

BICK, E. Constraint grammar-based conversion of dependency treebanks. *In*: **Proceedings of the 13th International Conference on Natural Language Processing (ICON)**. Varanasi: NLP Association of India, 2016, p. 109–114.

BRESNAN, J. Control and Complementation. **Linguistic Inquiry**, Vol. 13, n. 3, p. 343-434, 1982.

BUCHHOLZ, S.; MARSI, E. CoNLL-X Shared Task on Multilingual Dependency Parsing. **Proceedings of the Tenth Conference on Computational Natural Language Learning**. New York: Association for Computational Linguistics, 2006. p. 149–164.

FILLMORE, C. J. The Case for Case. *In*: BACH, E.; HARMS R. T. (ed.) **Universals in Linguistic Theory**. London: Holt, Rinehart and Winston. p. 1-88, 1968.

FREITAS, C.; ROCHA, P.; BICK, E. "Floresta Sintá(c)tica: Bigger, Thicker and Easier". *In*: TEIXEIRA, A.; LIMA, V. L. St. de; OLIVEIRA, L. C. de; QUARESMA, P. (ed.),

Proceedings of Computational Processing of the Portuguese Language, 8th International Conference, (PROPOR 2008), vol. 5190. Springer Verlag, 2008. p. 216-219.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.. **Deep Learning**. Cambridge MA: MIT Press, 2016.

HENRIQUES, C. C. **Nomenclatura Gramatical Brasileira: cinquenta anos depois**. São Paulo: Parábola, 2009.

HONNIBAL, M.; JOHNSON, M. An Improved Non-monotonic Transition System for Dependency Parsing. *In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Lisboa: Association for Computational Linguistics, 2015. p. 1373-1378.

IDE, N. Introduction: The Handbook of Linguistic Annotation. *In: Ide, Nancy; Pustejovsky, James (ed). Handbook of Linguistic Annotation*. Springer, 2017.

KARLSSON, F. Constraint Grammar as a Framework for Parsing Unrestricted Text. *In: KARLGREN, H. (ed.), Proceedings of the 13th International Conference of Computational Linguistics*, Vol. 3. ACM Digital Library, 1990. p. 168-173.

KIPPER-SCHULER, K. **VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon**. Tese de Doutorado (Ciência da Computação). University of Pennsylvania, 2005.

KONDRATYUK, D.; STRAKA, M. 75 Languages, 1 Model: Parsing Universal Dependencies Universally. *In: INUI, K.; JIANG, J.; NG, V.; WAN, X. (ed.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP)*, Hong Kong, China. Association for Computational Linguistics, 2019. p. 2779–2795.

LEVIN, B. **English Verb Classes and Alternations: A Preliminary Investigation**. University of Chicago Press, 1993.

MANNING, C.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge, MA: MIT Press, 1999.

MARNEFFE, M.-C. de; MANNING, C. D.; NIVRE, J.; ZEMAN, D.. Universal Dependencies. **Computational Linguistics** 47 (2), p. 255–308, 2021.

MITCHELL, T. Machine Learning. New York: McGraw Hill, 1997.

NIVRE, J. Towards a Universal Grammar for Natural Language Processing. *In: Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, 2015. p. 3-16.

NIVRE, J.; HALL, J.; NILSSON, J.; ERYİĞİT, G.; MARINOV, S. Labeled Pseudo-Projective Dependency Parsing with Support Vector Machines. *In: Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*. New York: Association for Computational Linguistics, 2006. p. 221-225.

NIVRE, J.; MARNEFFE, M.-C. de; GINTER, F.; HAJIČ, J.; MANNING, C. D.; PYYSALO, S.; SCHUSTER, S.; TYERS, F.; ZEMAN, D. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. *In: Proceedings of the 12nd International Conference on Language Resources and Evaluation (LREC 2020)*. Marseille: European Language Resources Association, 2020. p. 4034-4043.

PALMER, M.; GILDEA, D.; KINGSBURY, P. The Proposition Bank: An Annotated *Corpus* of Semantic Roles. *Computational Linguistics*, 31:1., p. 71-105, March, 2005.

PUSTEJOVSKY, J.; BUNT, H.; ZAENE, A. Designing Annotation Schemes: From Theory to Model. *In: IDE, N.; PUSTEJOVSKY, J. (ed.). Handbook of Linguistic Annotation*. Springer, 2017.

QI, P.; ZHANG, Y.; ZHANG, Y.; BOLTON, J.; MANNING, C. D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *In: JURAFSKY, D.; CHAI, J.; SCHLUTER, N.; TETRAULT, J. R. (ed.). Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. p. 101-108.

RADEMAKER, A.; CHALUB, F.; REAL, L.; FREITAS, C.; BICK, E.; PAIVA, V. de. Universal Dependencies for Portuguese. *In: Proceedings of the Fourth International Conference on Dependency Linguistics*. Linköping University Electronic Press, 2017. p. 197-206.

SARDINHA, T. B. Linguística de *corpus*: Histórico e Problemática. *Documentação e Estudos em Linguística Teórica e Aplicada (DELTA)*, 16:2, 2000. p. 323-367.

SILVA, J.; BRANCO, A.; CASTRO, S.; REIS, R.. Out-of-the-Box Robust Parsing of Portuguese. *In: Proceedings of the 9th International Conference on the Computational Processing of Portuguese*. Springer, 2010. p. 75-85.

SOUZA, E. de; CAVALCANTI, T.; SILVEIRA, A.; EVELYN, W.; FREITAS, C. **Diretivas e documentação de anotação UD em português (e para língua portuguesa)**. Rio de Janeiro: PUC-RIO, 2020. Disponível em: <http://comcorhd.letas.puc-rio.br/recursos/>.

STRAKA, M.; STRAKOVA, 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*. Association for Computational Linguistics, 2017. p. 88-99.

TESNIÈRE, L. **Éléments de Syntaxe Structurale**. Paris: Librairie C. Klincksieck, 1959.

TESNIÈRE, L. **Elements of Structural Syntax**. Tradução de OSBORNE, Timothy; KAHANE, Sylvain. Amsterdam: John Benjamins, 2015.