Extending Enhanced Universal Dependencies – addressing subjects in pro-drop languages

Magali S. Duran[1], Elvis A. de Souza[1],
Maria das Graças V. Nunes[1], Adriana S. Pagano[2], Thiago A. S. Pardo[1]
[1]Núcleo Interinstitucional de Linguística Computacional, Universidade de São Paulo – Brazil
[2]Faculdade de Letras, Universidade Federal de Minas Gerais – Brazil
magali.duran@gmail.com, elvis.desouza99@gmail.com,
gracan@icmc.usp.br, apagano@ufmg.br, taspardo@icmc.usp.br

Abstract

Enhanced Universal Dependencies (EUD) serve as a crucial link between syntax and semantics. Beyond basic syntactic dependencies, EUD provide valuable refined logical connections for downstream tasks such as semantic role labeling, coreference resolution, information extraction, and question answering. While the original EUD framework defines six types of relations, this paper introduces an extension designed to address subject propagation in pro-drop languages. This "Extended EUD" proposal increases the number of dependency relations that may be annotated in sentences, improving linguistic representation. Additionally, we report our experiments on a corpus of Portuguese (a pro-drop language), which we make publicly available to the research community.

1 Introduction

Syntax-based approaches for multilingual Natural Language Processing (NLP) have evolved in the last decade mostly due to the growing number of languages that have gold standard treebanks annotated under the Universal Dependencies framework (De Marneffe et al., 2021; Nivre et al., 2016). Currently, there are 296 corpora available on the UD website, but only 42 of them[1] provide Enhanced Universal Dependencies (EUD) in the ninth column of the well-known CoNLL-U format, called *deps*. EUD constitutes a "bridge" between syntax and semantics. In column "deps", basic dependencies are transformed into a more semantic-like representation which associates each token to every head token they modify, regardless of whether this relation is explicit or implicit. For this reason, each token in EUD may have multiple heads. Besides that, EUD includes case markers (dependents of *cc*, *case* and *mark* relations) in the relation name of their heads (*nmod*, *obl*, *acl*, *advcl*), providing helpful clues to their respective semantic roles.

By making explicit logical relations that were previously implicit, EUD increases the number of responses that can be obtained from an annotated corpus and paves the way for downstream applications that need these responses, such as semantic role labeling, coreference resolution, information extraction and question answering.

EUD allows for six types of "enhancements", including the propagation of subjects shared by coordinate clauses (*conj* clauses) and the assignment of "external subjects" for clauses with null subject (*xcomp* clauses). Once a clause with implicit subject receives an enhanced subject, it can share it with clauses dependent on it, creating a propagation chain. This recursive nature of EUD shows that one EUD relation may rely on the product of another EUD relation, and not only on basic syntactic dependencies. As an illustration, Figure 1 shows a sentence with EUD annotation where a *nsubj* is first propagated to a *conj* dependent clause and then to a *xcomp* dependent clause (in red edges).

Considering this recursive property of subject propagation within a sentence, EUD could further leverage syntax, also including propagation of subject for *acl*, *acl:relcl*, *advcl* and *ccomp*. This would be particularly relevant for so-called pro-drop languages, which present high occurrence of subject ellipsis since a verb form encodes Person, Number and sometimes Gender features that indicate who the subject is. If there is at least one explicit subject within a series of clauses in a sentence, such subject can be propagated to the other clauses with implicit subject. As it is possible to have different sequences of clauses in a sentence, the subject propagation path must be clear so as not to interrupt propagation. If, for example, a *ccomp* has an *xcomp* or *conj* as dependents, these clauses will only receive subject propagation once *ccomp* has received subject propagation before them. This

[1]According to UD v2.15 homepage: https://universaldependencies.org

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 143–152

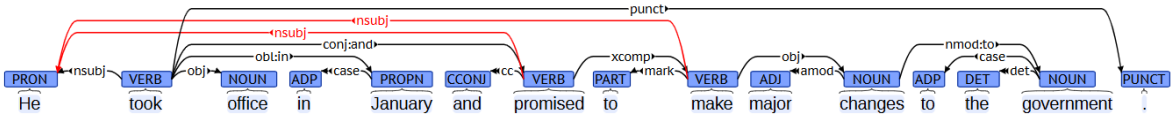

Figure 1: Example of EUD annotation with recursive propagation of subjects for the sentence "He took office in January and promised to make major changes to the government."

may be observed if we compare Figures 2 and 3. Whereas Figure 2 follows the current EUD guidelines and has only one subject, Figure 3, following the proposed extended EUD, has four subjects, as *ccomp* needs to receive subject propagation from its head before propagating it to the *xcomp* and *conj* dependent clauses.

It is important to note that ellipsis of the subject is not exclusive to pro-drop languages and other languages could take advantage of this extension in EUD. In English, for example, it is not uncommon to find an *advcl* with subject ellipsis (in bold in the following examples):

> Following his discharge, he taught English at a Buddhist girls' school **while also taking classes at Kyoto University**.[2]

> The meal was extremely overpriced and lacked flavor, **especially for being a special NYE menu**.[3]

However, when there is no explicit subject in a sentence, there is no possible propagation. In these cases, one might consider inserting an empty token to represent the elided subject, but this is a separate task, not related to subject propagation. Moreover, EUD emphasizes that empty tokens are only allowed for elided predicate insertion (predicates that are suppressed in the dependent clause but may be recovered from the matrix clause).

We took on the task of implementing EUD in Portuguese and exploring the full propagation of explicit subjects for dependent clauses that have elided subjects, writing rules based on features typical of verb forms in pro-drop languages. The strategy for doing this was to customize a multilingual rule-based EUD system, pre-annotate the corpus and manually revise the results, making changes to the rules, adding new rules, or fixing the UD gold standard annotation when necessary. This paper reports the finalized enhanced dependency annotation for the Porttinari-base corpus (Duran et al., 2023). A detailed account of the annotation process is not included here, as it has already been partially described by de Souza et al. (2024). More emphasis will be placed on describing the annotation process of the proposed EUD extension. The contributions of this paper are:

- to report the experience of extending subject propagation for other dependent clauses besides those defined in EUD, which we call "*Extended* Enhanced Universal Dependencies" (EEUD), and
- to release a Portuguese corpus annotated with EUD and EEUD (as far as we know, the first one for this language)[4].

The remainder of this paper is organized as follows: Section 2 briefly presents the related work and Section 3 reports the methodology we adopted; the results are reported in Section 4 while Section 5 explores the relation between basic dependencies and EUD; some limitations and final remarks are presented in Sections 6 and 7, respectively.

2 Related work

One of the precursors of UD, the Stanford Dependency model, already included an output containing inferred relations (De Marneffe et al., 2006; De Marneffe and Manning, 2008), intended to provide a semantically richer representation of syntactic relations, making them more readily employable by other applications. The idea was to replicate modifiers for each token they modify and to replicate case markers for each token they introduce. The term "enhanced" first appeared in de Marneffe et al. (2014) referring to a series of transformations: marking external subjects and the external role in relative clauses, renaming dependencies to include

[2]Example retrieved from UD English GUM 2.16.
[3]Example retrieved from UD English EWT 2.16.
[4]Both versions of the annotated corpus are available: the one with EUD in UD homepage, and the one with the full subject propagation (EEUD) in our website.

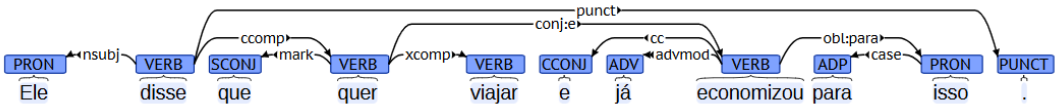

Figure 2: EUD annotation not extended to ccomp in "Ele disse que quer viajar e já economizou para isso." (lit.: He said that wants to travel and has already saved up for it.)

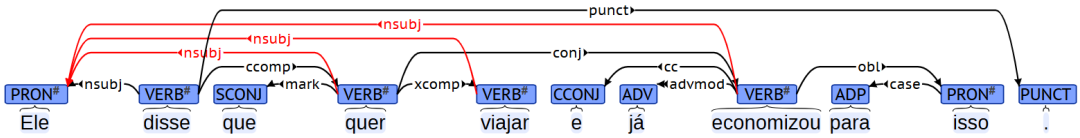

Figure 3: EUD annotation extended to *ccomp* in "Ele disse que quer viajar e já economizou para isso." (lit.: He said that wants to travel and has already saved up for it.)

case markers, and propagating relations over conjunctions.

Recognizing the relevance of such ideas, Schuster and Manning (2016) developed an English converter to produce an enhanced output according to UD restrictions, that is, expanding basic dependencies without changing them.

Nivre et al. (2018) evaluated two cross-lingual techniques to automatically annotate EUD in UD treebanks: a rule-based English system and a data-driven Finnish system. They used Swedish and Italian corpora to test both systems. The rule-based system performed better at assigning subjects, and the data-driven system performed better at propagating coordinate dependents. Overall, both systems contributed to bootstrap the EUD annotation.

The proposal to develop non-language-dependent systems to annotate EUD inspired two shared tasks held at IWPT: one in 2020 (Bouma et al., 2020) and another in 2021 (Bouma et al., 2021). However, their multilingual dataset did not include Portuguese. EUDs were instantiated in Portuguese by Pagano et al. (2023) and, later, de Souza et al. (2024) developed a Portuguese customization: the rule-based system UDtoEUD[5] (Guillaume and Perrier, 2021), a Graph Rewriting System that uses GREW (Bonfante et al., 2018) to convert basic dependencies into EUD. Droganova and Zeman (2019) clarify that EUD annotation is optional in treebanks, and it is allowed to annotate only one, several or all of the six types of EUD. The UD framework, until now, does not allow the annotation of other types of EUD out of the six types already described in the guidelines[6].

[5] https://gitlab.inria.fr/grew/udtoeud

[6] https://universaldependencies.org/u/overview/enhanced-syntax.html

3 Methodology

Revising EUD annotation is an extremely heavy and time-consuming task, which requires even more qualified human resources than those needed to revise basic dependencies. To bootstrap EUD annotation, we customized the UDtoEUD converter to Portuguese (de Souza et al., 2024), after a in-depth linguistic study on enhanced dependencies for the Portuguese language (Duran, 2024).

We chose UDtoEUD because it has all the right qualities for our task: 1) it is based on rules, which enabled us to improve the system to increase the quality of the whole annotation without relying on previously annotated data; 2) it scored above 98% over gold basic UD data for Italian, which is the language closest to Portuguese among those that took part in the shared task; and 3) it allows entering language-specific lexicon: control and raise verbs (to determine the subject of *xcomp*), and adverbs (to determine which adverbs left to the head of *conj* should propagate to the *conj* dependent).

By customizing UDtoEUD, we automatically annotated the six EUD types in Porttinari-base, which has 8,418 sentences. As the converter reached 96.05% ELAS in a gold standard EUD composed of sentences assembled from Porttinari-base, it was not deemed necessary to perform human-revision on a one-by-one basis in this phase.

In the next phase, we improved the converter to include propagation of subjects to other clauses: *ccomp*, *advcl*, *acl* and *acl:relcl*. To find candidates for propagation, we used the following rules:

- the dependent clause should not be head of *nsubj* or *nsubj:pass* or *csubj*, that is, if the

clause has a subject, the place for the subject is already filled;

- the dependent clause should not be head of *expl:impers*, that is, if there is a mark of impersonalization, there is no place for a subject;
- the dependent clause should not be an impersonal verb ("haver", "chover", "anoitecer", etc.), that is, if a verb is impersonal, it does not have a place for a subject.

Besides that, for *ccomp* and *advcl*, their head should be head of *nsubj* or *nsubj:pass*, i.e., they should have a subject to propagate, regardless of whether it was an explicit subject or a subject resulting from a previous propagation. In other words, if the head does not have a subject to propagate, no propagation is possible. For *acl:relcl* and some types of *acl*, the head of the head should have a subject to propagate.

Another restriction to propagation was that the dependent clause (or its *aux* or *aux:pass* or *cop*) should have the same values for the features Person and Number of the clause (or its *aux* or *aux:pass* or *cop*) from which it would inherit the subject.

As this is a new task, all the cases of propagation (2,147 sentences) and non-propagation (759 sentences) were verified by linguists that are experts in UD, leading to the improvement of the system. In this phase, we detected some non-recurrent errors in basic dependencies, which were promptly corrected. Since we used a strategy of converting basic dependency trees into EUD, EUD automatic annotation only works if the basic dependencies are correctly annotated; hence they are a very good source for checking the basic dependencies logic.

4 Results

In this section, we discuss separately the implementation of EUD and EEUD, but present the increase in relations achieved after each of these phases.

4.1 Phase 1 - EUD identification

As already mentioned, since the UDtoEUD customization for Portuguese achieved an overall ELAS of 96.05% in a gold dataset developed for the task (de Souza et al., 2024), no human-revision was deemed necessary for most of the automatically annotated EUD in Porttinari-base. As expected, the biggest challenge in EUD was elided predicates. This EUD inserts an empty token whenever a token, dependent on a *conj* relation, is head of an *orphan* relation. The relation *orphan* is relatively rare in our corpus (67 occurrences). Inserting an empty token whenever an *orphan* occurs is a trivial task, but deciding where this empty token shall be inserted and naming the relations it establishes with the ex-head and the ex-dependent of the *orphan* relation is not easy for a rule-based system. So we decided to manually revise all the sentences containing *orphan* relations in the corpus, which required a lot of manual editing until the respective sentences reached their final version. For this reason, we excluded the *orphan* relation from the statistics reported in this paper. This issue will be addressed in depth in a forthcoming study. Table 1 shows the ELAS of the other five EUD types before and after customizing UDtoEUD for Portuguese.

Type of EUD	Original	Customized
1 - case assignment	95.17%	98.90%
2 - *xcomp* subjects	92.54%	97.06%
3 - prop. *conj* head	84.13%	96.43%
4 - prop. *conj* dependents	96.67%	96.67%
5 - relative pronoun *ref*	94.23%	94.23%

Table 1: ELAS of UDtoEUD per EUD type before and after customization. Source: de Souza et al. (2024)

Out of these five EUD, the only one that required an amendment in the customized UDtoEUD was the assignment of an external subject to *xcomp*. We found cases of *xcomp* subject corresponding to the *obj*, *obl* or *iobj* of the matrix clause, and sometimes such elements are elided in Portuguese, leading to an *xcomp* whose external subject cannot be determined. To avoid the wrong assignment of the subject of the matrix clause to the *xcomp*, we had to improve the rules, informing a list of verbs that allow elided complements, for example: "permitir" (*to allow*), "deixar" (in the sense of *to allow*) and "mandar" (*to order*). The following example shows an elided *obj*:

> Os leilões permitem [Ø] pagar o animal em até 50 parcelas. (lit.: "Auctions allow [Ø] to pay for the animal in up to 50 installments" meaning "allow anybody to pay")

4.2 Phase 2 - EEUD identification

Departing from the basic rules cited before, we generated CoNLL-U files containing, for each of the four relations (*acl*, *acl:relcl*, *advcl* and *ccomp*),

cases of subject propagation and cases of non-propagation. These files were revised by three expert annotators, who indicated the cases in which propagation succeeded, the cases in which propagation failed (the subject should not be propagated or the subject propagated was not the correct one), the cases in which non-propagation was correct and the cases in which non-propagation was incorrect (the subject should have been propagated). Table 2 shows the result of this revision: 87.98% of the propagated subjects were considered correct and 12.02% were considered incorrect; 73.88% of the non-propagated subjects were considered correct and 26.12% were considered incorrect. The high percentage of subjects which should have been but were not propagated (26.12%) is due to the fact that our rules were initially designed to deal more with verbal predicates than with nominal predicates.

Dealing with each of the four relations separately made it easier to analyze the cause of the errors and come up with solutions to improve the rules. In what follows, we present an example of subject propagation for each focused relation and comment on errors that we found.

acl: The *acl* dependent rarely contains a subject and rarely presents a finite form. Most of the time it is a passive voice construction without an *aux:pass* or an active construction without an *aux*. Its subject is the nominal that is its head in basic dependencies.[7] In the sentence below, the propagated subject is in bold and the *acl* dependent is underlined.

> O último **discurso** gravado de Al-Baghdadi é de novembro de 2016. (lit.: The last **speech** recorded of Al-Baghdadi is from November 2016.)

Two problems were found in the propagation of *acl*: one relating to the annotation of the feature Voice=Pass, and another relating to the propagation of coordinate nominals as subject.

Regarding the first problem, we found inconsistencies in the annotation of past participles, as some of them should have the feature Voice=Pass and did not, and others should not have it and did, thus affecting the type of subject propagated – *nsubj* or *nsubj:pass*. In Portuguese, sometimes a past participle form is not a reduced form of a passive construction. This is the case with intransitive verbs, as shown in the following example, which clearly does not constitute a passive voice sentence. These problems required a complex revision of the Voice=Pass feature in the corpus.

> Outra coisa, no entanto, é a avalanche de 1,5 milhão de **refugiados** chegados em 2015 e 2016. (lit.: Another thing, however, is the avalanche of 1.5 million **refugees** arrived in 2015 and 2016)

The second problem is related to the heuristics used to propagate subjects: when a dependent clause has a subject, it is not a candidate for propagation. This was a problem when two or more coordinate tokens are the subject, causing the system to propagate only the last one. It was solved for all propagation rules, not only for *acl*. A limitation of coordinate tokens as subject, as Schuster and Manning (2016) points out, is that it is not always clear whether the intention is a distributive or collective interpretation. Although we are aware of this, we have chosen to ignore it for the time being, and adopted a distributive interpretation.

acl:relcl: In Portuguese, the *acl:relcl* dependent almost always has an internal subject. When it does not, and the *acl:relcl* clause does not contain a relative pronoun, its subject is in most cases a nominative relative pronoun that is, at the same time, the head of the *acl:relcl* itself, as seen in the next example.

> Quem ouviu julgou que o treinador estava perdido. (lit.: **Who** heard it thought the coach was lost)

In other cases, the subject of *acl:relcl* is the head of the head of *acl:relcl*:

> **Ele** e **Saud** pediram desculpas publicamente pelo que disseram. (lit.: **He** and **Saud** have publicly apologized for what said)

When Person and Number features of the copula verb and the *acl:relcl* dependent present different values, propagation is not performed (Figure 4).

advcl: The *advcl* dependent presents several patterns, depending on its semantic function. We observed good results with temporal and causal *advcl*, but comparative and conformative *advcl* rarely received a propagation. Finally, *advcl* behaves as *xcomp* if its head has an *obj* dependent and such

[7] This circularity is a situation that also occurs when the head of an *acl:relcl* substitutes a relative pronoun as subject: "the man who died" turns into "the man died" when the *ref* relation is annotated in EUD.

	PROPAGATED		NOT PROPAGATED			
	correct (TP)	incorret (FP)	correct (TN)	incorret (FN)	Total	Accuracy
acl	896	57	29	7	989	93.53%
acl:relcl	221	9	134	59	423	83.92%
advcl	621	151	228	119	1,119	75.87%
ccomp	172	44	186	19	421	85.04%
Total	1,910	261	577	204	2,952	84.25%
(%)	87.98%	12.02%	73.88%	26.12%		

Table 2: Results of EEUD revision.

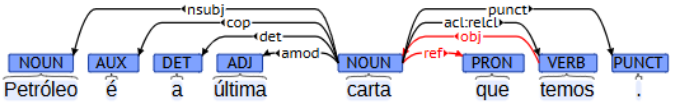

Figure 4: An example in which the subject did not propagate because Person and Number did not match (the AUX "é" is third person singular and the VERB "temos" is first person plural): "Petróleo é a última carta que temos" (lit.: Oil is the last card that have)

obj is the subject to be propagated to the *advcl*. Although we do not have any subrelation for *advcl* at the moment, this could be very productive for EUD purposes. A particularly noteworthy case is that of an *advcl* that has its own subject, while its head does not. Therefore, the subject may be propagated from the dependent to the head, as Figure 5 shows.

ccomp: The *ccomp* dependent is almost always a finite construction (the verb or its auxiliary is a finite form). We analyzed those *ccomp* related to reported speech (which we annotated as *ccomp:speech*) and concluded that subject propagation was not feasible or suitable: almost all present a different Person and Number in the reported speech and the head of *ccomp:speech*.

Some infinitive constructions dependent on *ccomp* seem to be impersonal, and the propagation rules lead to errors, as in the next sentence, where "revitalização" is not the subject of "devolver".

> A revitalização *implica* em devolver a esses centros a vitalidade. (lit.: Revitalization **implies** to restore vitality to these centers.

Table 3 shows that 146 (31.4%) of the 465 errors found were solved by fine-tuning the rules and 73 (15.7%) were solved by corrections in the basic dependencies annotation. The remaining 246 errors (52.9% of all errors) were solved by manually editing EUD. This does not necessarily mean that there are no patterns in the remaining errors; rather, we were unable to identify them.

	rules	basic	other
acl	11	32	21
acl:relcl	3	5	60
advcl	112	21	137
ccomp	20	15	28
Total	146	73	246
(%)	31.4%	15.7%	52.9%

Table 3: Sources of errors.

We used the resulting gold standard to compute accuracy and precision of the rules before and after fine-tuning (Table 4).

Table 5 specifies the number of subjects propagated by dependency relation (deprel), both in EUD and in EEUD. It is important to note that the application order of the subsets of rules concerning subject propagation is relatively free; however, they should be applied recursively until no additional subject propagation is obtained. In some cases, this means repeating these rules up to four or five times, and there is room to propose ways to optimize this repetition.

Table 6 shows the number of each type of relation in the corpus, in 3 columns: the first one shows basic dependencies, the second one shows EUD and the third one includes EEUD. Sentences with an *orphan* relation were not taken into account for this table due to the difficulty to align the null nodes inserted throughout the three versions

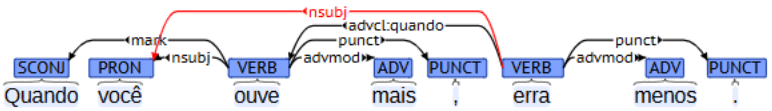

Figure 5: An example in which the subject propagated from the dependent to the head: "Quando você ouve mais, erra menos" (lit.: When you listen more, [you] make fewer mistakes)

	Total	Accuracy (Before)	F1-Score (Before)	Accuracy (After)	F1-Score (After)
acl	858	0.937063	0.966871	0.967366	0.983071
acl:relcl	349	0.787966	0.787356	0.828080	0.831461
advcl	1,877	0.798082	0.797435	0.928077	0.935002
ccomp	441	0.888889	0.887872	0.929705	0.931567

Table 4: Performance of rules before and after fine-tuning, calculated over the gold standard.

	Basic subjects	EUD		EEUD	
acl	135	139	(+2.96%)	1,101	(+692.09%)
acl:relcl	1,570	1,621	(+3.25%)	1,792	(+10.55%)
advcl	514	522	(+1.56%)	1,556	(+198.08%)
ccomp	690	706	(+2.32%)	919	(+30.17%)
conj	435	1,278	(+193.79%)	1,385	(+8.37%)
xcomp	-	1,712	-	1,869	(+9.17%)

Table 5: Subjects propagated by deprel in EUD and EEUD.

of the annotation. As may be observed, EEUD contributed to increasing 15.39% the number of *nsubj* and 81.61% of *nsubj:pass*

5 EUD as feedback on basic dependencies annotation

When using a conversion strategy, EUD and EEUD strongly rely on the annotation in other CoNLL-U columns, mainly on Head (dependency head) and Deprel (dependency relation), but also on Upos (universal part-of-speech) and Feat (column that encodes morphological features).

In Portuguese, several verbs have ambiguous forms in the first and third person, mainly in the past imperfect tense. As this feature is used to propagate subjects, annotation errors can lead to EUD errors. When we noticed this type of error, we looked for all the verb forms that allowed more than one value for the Person feature and revised their annotation.

As already mentioned, we also revised the Voice feature of past participles without a dependent *aux:pass*, as they may require Voice=Pass or not. When two predicates have different Voice values, the respective relation of the propagated subject needs to be adjusted from *nsubj* to *aux:pass* (or the inverse), as shown in Figure 6.

6 Limitations

The rules we implement rely heavily on comparing the Person and Number features of the head and the dependent: if they are the same, there is propagation; otherwise, there is not. However, even when these features have the same values, propagation may not work. This problem occurs mainly with Person=3, as Person=1 and Person=2 are normally not ambiguous within the discourse situation.

Two examples illustrate the above problem. In the first example (Figure 7), "o veículo" (the vehicle) is clearly not the subject of "acordou" (woke up), even though the propagation is licensed because the Person and Number rules have been met. In the second example (Figure 8), only world knowledge allows us to detect the error, as the subject of "preferia" (would prefer) is the obl "ex-bailarina" and not "chefe" (boss), which is *nsubj* of the head.

deprel	Basic	EUD		EEUD	
acl	1,635	1,748	(+6.91%)	1,748	-
acl:relcl	1,899	2,039	(+7.37%)	2,039	-
advcl	2,311	2,522	(+9.13%)	2,522	-
advmod	6,107	6,051	(-0.92%)	6,051	-
amod	6,595	6,833	(+3.61%)	6,833	-
appos	1,036	1,153	(+11.29%)	1,153	-
aux	806	831	(+3.10%)	831	-
aux:pass	983	1,008	(+2.54%)	1,008	-
case	22,359	23,196	(+3.74%)	23,196	-
cc	4,182	4,183	(+0.02%)	4,183	-
ccomp	1,102	1,216	(+10.34%)	1,216	-
ccomp:speech	669	811	(+21.23%)	811	-
conj	4,518	4,518	-	4,518	-
cop	2,871	3,043	(+5.99%)	3,043	-
csubj	356	411	(+15.45%)	411	-
csubj:outer	4	4	-	4	-
csubj:pass	1	1	-	1	-
det	23,897	23,881	(-0.07%)	23,881	-
discourse	254	277	(+9.06%)	277	-
dislocated	77	85	(+10.39%)	85	-
expl	562	562	-	562	-
expl:impers	154	154	-	154	-
fixed	1,339	1,339	-	1,339	-
flat	87	87	-	87	-
flat:foreign	68	68	-	68	-
flat:name	3,331	3,332	(+0.03%)	3,332	-
iobj	95	95	-	95	-
list	27	27	-	27	-
mark	4,301	4,520	(+5.09%)	4520	-
nmod	12,534	13,549	(+8.10%)	13,549	-
nsubj	9,480	12,117	(+27.82%)	13,982	(+15.39%)
nsubj:outer	97	97	-	97	-
nsubj:pass	709	968	(+36.53%)	1,758	(+81.61%)
nummod	1,759	1,784	(+1.42%)	1,784	-
obj	7,200	7,650	(+6.25%)	7,650	-
obl	9,058	9,743	(+7.56%)	9,743	-
obl:agent	498	547	(+9.84%)	547	-
parataxis	877	966	(+10.15%)	966	-
punct	22,094	22,094	-	22,094	-
ref	0	1,814	-	1,814	-
reparandum	5	5	-	5	-
root	8,354	8,354	-	8,354	-
vocative	26	29	(+11.54%)	29	-
xcomp	2,330	2,479	(+6.39%)	2,479	-

Table 6: Relations increase, by deprel, due to EUD and EEUD. There are two relations that saw a decrease from Basic to EUD: *advmod* and *det*. The former is due to the relative adverb "onde", which is labeled as *advmod* dependent on the verb of the relative clause in the basic tree, but is labeled as *ref* dependent on the nominal that is modified by the relative clause in the EUD graph. The latter is due to the relative pronoun "cujo", which is labeled as *det* dependent on a nominal to its right in the basic tree, but is labeled as *ref* dependent on the nominal to its left in the EUD graph.

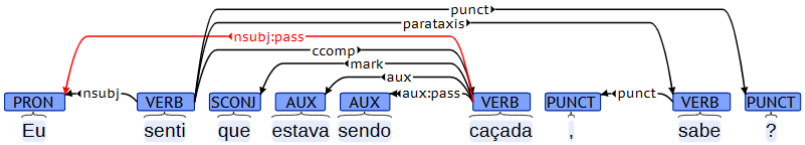

Figure 6: An example of *nsubj* propagated as *nsubj:pass*: "Eu senti que estava sendo caçada, sabe?" (lit.: I felt that was being hunted, you know?

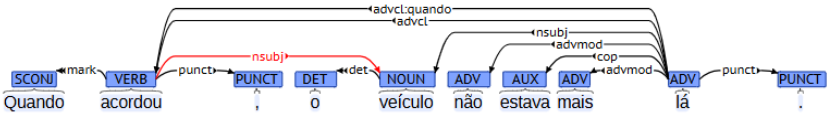

Figure 7: Error analysis – the subject of the head does not fill the *advcl* subject: "Quando acordou, o veículo não estava mais lá" (lit.: When woke up, the vehicle was not there anymore)

7 Final remarks

We produced two versions of the corpus: one with EUD as described in the UD guidelines (corpus available on the UD website) and another with all enhanced relations, including those extensions we implemented, i.e., EEUD[8].

Working with a rule-based system to annotate enhanced dependencies has given us a deeper understanding of our language. At the same time, when we checked the automatically assigned enhancements, we noticed some inconsistencies in the annotation of the basic dependencies, which were corrected, improving the corpus quality as a whole.

EEUDs are dependency relations more difficult to annotate automatically, as the number of errors produced by the rules was greater than the number of errors in the original EUD, and required many manual corrections. Even so, the task proved to be productive, as we managed to increase the number of propagated subjects by 56%.

Drawing on our two versions of the golden standard corpus (one annotated with EUD and another with EUD + EEUD), as future work, we intend to train automatic EUD and EEUD parsers for Portuguese. We also plan to take a closer look at the enhancement of the *orphan* dependency relation. For this, we intend to produce data augmentation of *orphan* cases, since the sparsity of the data, added to the diversity of patterns, will probably prevent automatic approaches from achieving good performance in EUD annotation.

The rules of the customized version of UDtoEUD are available at Github[9]. We believe the rules we designed to deal with elided subjects may be useful for other languages, especially for the so-called pro-drop languages. Extending EUD to allow any kind of subject propagation would partially solve the inequality in the number of subjects between pro-drop and non-pro-drop languages, remaining cases where there is no subject to propagate. The gold standard corpus may also be useful for those wishing to experiment with a non-rule-based approach.

Acknowledgments

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 from São Paulo Research Foundation (FAPESP grant #2019/07665-4) and IBM Corporation. The project was also supported by 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. Adriana Pagano thanks the National Council for Scientific and Technological Development (CNPq 404722/2024-5; 313103/2021-6) and Minas Gerais State Agency for Research and Development (FAPEMIG).

[8] https://sites.google.com/icmc.usp.br/poetisa/resources-and-tools

[9] https://github.com/alvelvis/eud-portugues/blob/main/flask/conjunto_regras_porttinari.grs. The EUD rules can be found as the "eud_portuguese" strategy, while the EEUD rules can be found as the "eud_portuguese_extended" strategy.

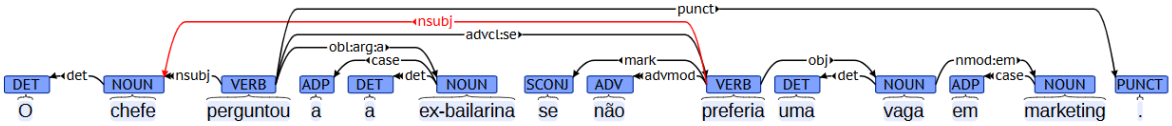

Figure 8: Error analysis – the logical subject of the *ccomp* should be the *obl* of its head (ex-bailarina): "O chefe perguntou à ex-bailarina se não preferia uma vaga em marketing" (lit.: The boss asked the former dancer if wouldn't prefer a job in marketing)

References

Guillaume Bonfante, Bruno Guillaume, and Guy Perrier. 2018. *Application of Graph Rewriting to Natural Language Processing*, volume 1 of *Logic, Linguistics and Computer Science Set*. ISTE Wiley.

Gosse Bouma, Djamé Seddah, and Daniel Zeman. 2020. Overview of the IWPT 2020 shared task on parsing into enhanced Universal Dependencies. In *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, pages 151–161, Online. Association for Computational Linguistics.

Gosse Bouma, Djamé Seddah, and Daniel Zeman. 2021. From raw text to enhanced Universal Dependencies: The parsing shared task at IWPT 2021. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 146–157, Online. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Iceland. European Language Resources Association (ELRA).

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, and 1 others. 2006. Generating typed dependency parses from phrase structure parses. In *Lrec*, volume 6, pages 449–454.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Elvis A de Souza, Magali S Duran, V Nunes Maria das Graças, Gustavo Sampaio, Giovanna Belasco, and Thiago Pardo. 2024. Automatic annotation of enhanced universal dependencies for brazilian portuguese. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 217–226. SBC.

Kira Droganova and Daniel Zeman. 2019. Towards deep universal dependencies. In *Proceedings of the fifth international conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 144–152.

Magali S Duran. 2024. Guidelines for annotating enhanced syntactic dependency relations in portuguese, following the guidelines of the Universal Dependencies (UD) approach (in Portuguese). Technical Report 448, ICMC-USP.

Magali S Duran, Lucelene Lopes, Maria das Graças Nunes, and Thiago Pardo. 2023. The dawn of the Porttinari multigenre treebank: Introducing its journalistic portion. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 115–124, Porto Alegre, RS, Brasil. SBC.

Bruno Guillaume and Guy Perrier. 2021. Graph rewriting for enhanced universal dependencies. In *IWPT 2021-17th International Conference on Parsing Technologies*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, and 1 others. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Joakim Nivre, Paola Marongiu, Filip Ginter, Jenna Kanerva, Simonetta Montemagni, Sebastian Schuster, and Maria Simi. 2018. Enhancing universal dependency treebanks: A case study. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 102–107.

Adriana Pagano, Magali S Duran, and Thiago Pardo. 2023. Enhanced dependencies para o português brasileiro. In *Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival*, pages 461–470.

Sebastian Schuster and Christopher D Manning. 2016. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378.