Increasing manually annotated resources for Galician: the Parallel Universal Dependencies Treebank

Xulia Sánchez-Rodríguez[*1,2] and **Albina Sarymsakova**[*1] and **Laura Castro**[1] and **Marcos Garcia**[1]
[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
[2]Language Variation and Textual Categorisation (LVTC)
Universidade de Vigo
xulia.sanchez.rodriguez@usc.gal, albina.sarymsakova@usc.gal
laura.sanchez@usc.gal, marcos.garcia.gonzalez@usc.gal

Abstract

This paper presents the development of the Parallel Universal Dependencies (PUD) treebank for Galician. PUD treebanks were originally created for the CoNLL 2017 Shared Task on Multilingual Parsing, and have subsequently been used both to develop NLP tools and to perform cross-linguistic analysis using parallel resources. The Galician PUD consists of 1000 sentences manually reviewed by professional translators and aligned with the other 23 available PUD treebanks. The linguistic annotation was first carried out using state-of-the-art NLP tools for Galician, and then reviewed by two experts, achieving a high inter-annotator agreement. We describe the process of translating, pre-processing, and reviewing the corpus, and discuss the annotation of some linguistic phenomena in comparison with other PUD treebanks. The release of Galician PUD will double the size of the available treebanks for this linguistic variety, as only 1000 reviewed sentences were available to date. It will also be useful for carrying out cross-linguistic analyses including Galician, and as an additional test corpus for machine translation systems.

Key words: Galician, Syntax, Universal Dependencies, PUD

1 Introduction

Universal Dependencies (UD) is a multilingual framework of natural language processing (NLP). It functions as a cross-linguistic, standardizing system for morphological and syntactic annotation, fostering a collaborative initiative to generate annotated corpora across numerous languages, forming an expanding repository of such resources that serve as fundamental data for various language-specific applications and linguistic studies (de Marneffe et al., 2021). At present, the UD project encompasses over 217 treebanks representing 122 languages from 24 distinct language families.[1] However, there is a considerable disparity regarding the volume of the treebanks available for each language. In fact, the scarcity of manually annotated data for low-resource varieties such as Galician poses a challenge for those interested in conducting both cross-linguistic and NLP studies.

A core component of UD are the Parallel Universal Dependencies (PUD) treebanks, which are a set of parallel corpora composed of the same 1000 sentences consistently ordered, with sentence alignment between languages, and sourced from news articles and Wikipedia. PUD treebanks are currently available for 23 languages and were established for the *CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (Zeman et al., 2017). The initial PUD treebanks have been translated to different languages such as Turkish (Türk et al., 2019), Icelandic (Jónsdóttir and Ingason, 2020), or Bengali (Majumdar et al., 2022), among others.

Besides providing annotated corpora in multiple languages, PUD treebanks have been used for different purposes, such as facilitating multilingual comparative analyses of automatic parsers (Alves et al., 2022), developing NLP tools such as sentiment analysis systems (Kanayama and Iwamoto, 2020), or examining syntactic differences among languages, shedding light on quantifying the prevalence of different syntactic divergences across language pairs (Nikolaev et al., 2020).

With the above in mind, this paper introduces the development of a new manually annotated treebank for Galician (Galician PUD), which will be incorporated into the official PUD repository. The Galician PUD treebank has been translated by professionals, pre-processed using state-of-the-art NLP tools, and finally annotated by two experts. The quality of the manual annotation was assessed using

[*]Equal contribution.

[1]https://universaldependencies.org/

both inter-annotator agreement and automatic parsing measures, and the results indicate that the new Galician PUD has a high-quality and consistent annotation.

Before the release of this new PUD, Galician possessed only one digital corpus with manual annotation of syntactic dependencies —TreeGal (Garcia et al., 2018)—, comprising a total of 1000 manually revised sentences. With the introduction of this new corpus, an additional 1000 sentences are incorporated, effectively doubling the size of the existing resource, which is a significant development for the linguistic resources available for Galician. The release of this new resource will contribute to the PUD cross-linguistic data repository serving as a valuable parallel corpus for improving and assessing the performance of natural language processing systems, facilitating comparisons between different language varieties, or evaluating machine translation systems.

2 Galician PUD

This section describes the translation process of the Galician PUD followed by the annotation and revision steps and their results.

2.1 Translation

The source text for this study consisted of English sentences extracted from the English PUD (Zeman et al., 2017). The translations into Galician were made by three professional translators, all of them native speakers of Galician, and comprehensive translation guidelines were established to maintain consistency throughout the process. Alongside the original English sentences, two automatic translations (from the Spanish and Portuguese PUDs) were also presented as suggestions in order to promptly address any potential doubts that may arise during the translation process. Automatic translations were performed using a state-of-the-art neural machine translation system from Spanish (Gamallo et al., 2023b), and a rule-based transliteration system from Portuguese (Ortega et al., 2022).

Upon the completion of the translation phase, we compared the BLEU scores obtained with this new resource to those of the original translation models. The evaluation was carried out between the automatic translation of the sentences in Spanish, Portuguese and English (as a new English-Galician translation system was published in this period (Gamallo et al., 2023a)), and their Galician translation performed by specialists. The highest BLEU score was achieved with Spanish (56.4), followed by a high score for English (42.1), and a slightly lower BLEU on the Portuguese transliteration (36.8), although it still yielded a commendable result. The fact that the BLEU scores in Spanish are noticeably lower than those of the NMT system (74.3), while the English ones are similar (42.7), suggests that the Galician PUD sentences were not primarily based on any of the automatic translations.

2.2 Pre-processing

The annotation task involved a multi-step linguistic processing approach. After translating the 1000 sentences, they underwent tokenization and tagging using the linguistic toolkit Freeling (Padró, 2011). Subsequently, a specialized script was applied to resolve split contractions and to convert the FreeLing output into UD standard format CoNLL-U.[2] Following this, UDPipe (v1.2) (Straka and Straková, 2017) was used as a parsing tool, with the TreeGal-based model (Garcia et al., 2018) to provide the automatic annotation of syntactic dependencies. The python implementation of *udapi* was used throughout the annotation process to verify the treebank consistency.[3]

2.3 Annotation

The treebank has been annotated by two experts: a native speaker with a strong background in Linguistics and syntax, and a postdoctoral researcher in Linguistics with high competence in Galician. Both annotators initially annotated 30 sentences to familiarize themselves with the procedure and make sure that they were following the same parameters for annotation. These initial sentences used for training were not included in the final PUD.

For the annotation process, the 1000 sentences of the dataset were divided into different files, each containing 50 sentences. These files were then assigned to the annotators, who conducted individual labeling using the INCEpTION (Klie et al., 2018) platform. Regular follow-up meetings with additional language experts were conducted to address any uncertainties or questions that arose during the annotation process. It is worth noting that each file was reviewed only by one annotator, except for the last 50 sentences (951-1000), which were

[2] https://universaldependencies.org/format.html

[3] https://github.com/udapi/udapi-python

annotated again by both of them. This allowed us not only to compute inter-annotator agreement at a final stage, but also to compare it with the initial one obtained from the training sentences, and therefore to assess whether the agreement had improved as more sentences were annotated.

2.4 Results

Inter-annotator agreement: We calculated the annotators' agreement taking into account both the dependency head of each token and the specific syntactic relation of each dependency. To do so, we used Cohen's κ (Cohen, 1960) for the Head and Deprel columns, and the standard Labeled and Unlabeled Attachment Score (LAS and UAS, respectively), in both the training sentences and those annotated for the treebank (Table 1).

Dataset	Head	Deprel	LAS	UAS
Training	0.83	0.87	85.04	90.78
Treebank	0.96	0.96	93.79	96.48

Table 1: Inter-annotator agreement for the 30 training sentences and the final 50 sentences of the treebank. *Head* and *Deprel* are the Cohen's κ of both annotations, while LAS and UAS refer to the Labeled and Unlabeled Attachment Scores, respectively.

During the training phase, the values ranged from 0.83 (κ for the syntactic head) to 0.91 (90.78 UAS), which are reasonably high scores considering it was the initial phase of annotation for both annotators. These values significantly improved with the final 50 sentences of the treebank, increasing to 93.79 LAS and 96.48 UAS, and with $\kappa = 0.96$ for both the Head and Deprel columns. This represents a very high level of agreement, demonstrating (i) the similarity between the two annotators in their annotations, and (ii) the usefulness of the training process as well as the follow-up meetings during the annotation. This improvement in annotation quality as the process advances is evident, with increased agreement achieved at the end of the PUD. Consequently, a discussion of the disagreements and a review of the initial annotations allowed the annotators to identify some discrepancies in the labeling of some syntactic phenomena, whose final annotation was revised in the treebank as a whole.

Automatic parsing: To assess the quality of the annotation indirectly, we evaluated the performance of different models in the final version of the treebank. We used both UDPipe v1.2 (the one used for the initial annotation) and UDPipe v2 (Straka, 2018) with the two available models for Galician: TreeGal and CTG.[4] The first one was trained with Galician-TreeGal, a 1000 sentences treebank with manual annotation following to the latest UD guidelines. CTG models are based on the Galician-CTG treebank, a larger corpus (3993 sentences) with automatic syntactic annotation provided by FreeLing and automatically converted to UD (Gómez Guinovart, 2017).

The results in Table 2 show that the annotation of the Galician PUD is consistent with that of TreeGal, as both models (TreeGal-based UDPipe v1.2 and v2) obtain very similar results on the two manually annotated treebanks. This finding is reinforced by the performance of the CTG-based models, which achieve high results on the same data but much lower values on both PUD and TreeGal. There may be a bias in the annotation as we used UDPipe v1.2 for pre-processing the data, but in general, the results of the two versions of UDPipe and the inter-annotator agreement values suggest that the manual review of the Galician PUD is of good quality.

3 Discussion

Following the existing definition of auxiliaries in UD[5] and the fact that the current Galician guidelines already include semi-copulative verbs like *semellar* ('to seem', 'to appear'), our proposal for the Galician PUD incorporates other verbs not included in Treegal as auxiliaries. An example of this can be seen in Figure 1 with the verb *parecer* (a synonym for *semellar*).

Parecía desexar que (...) actuasen sen el

"He seems to have wished [for the Senate and the state] to (...) act without him"

Figure 1: Example of our proposal to annotate the verb *parecer* ('to seem') as an auxiliary (sentence id: w01062063).

Regarding comparative sentences, we encountered challenges in determining the dependency relationships between various elements within the comparison structure. Due to the absence of a

[4] UDPipe v1.2 models were trained with the 2.5 version of the treebanks, while UDPipe v2 used the 2.12 release. However, both treebanks are essentially identical in these two versions.

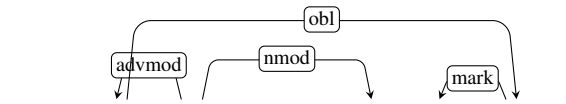
[5] https://universaldependencies.org/u/dep/aux_.html

Model	Galician PUD		TreeGal		CTG	
	LAS	UAS	LAS	UAS	LAS	UAS
UDPipe_v1.2 TreeGal	78.56	84.25	77.50	81.70	52.58	63.96
UDPipe_v1.2 CTG	59.46	71.98	55.80	68.68	81.20	85.50
UDPipe_v2 TreeGal	79.78	85.71	82.78	86.99	65.82	78.26
UDPipe_v2 CTG	64.37	78.14	59.32	74.78	84.31	86.86

Table 2: LAS and UAS of UDPipe models for Galician on the new Galician PUD and in other UD treebanks.

standardized model in the guidelines and a lack of consensus across languages within the PUD, our proposal for such sentences is to annotate the second part of comparative constructions with the 'obl' (oblique nominal) label, dependent on the adverbial modifiers, i.e., *máis* ('more') or *menos* ('less', 'fewer'), as can be seen in various examples and languages from the first PUD edition (Zeman et al., 2017)[6,7].

We have provided examples for these proposed dependencies, which are illustrated in Figures 2 (comparison of inferiority) and 3 (comparison of superiority).



"(...) fewer km of underground (...) than Madrid"

Figure 2: Example of a comparative sentence and its annotation proposal (sentence id: n04002020).

In addition to this, several ambiguous cases that required consensus between the two annotators arose during the annotation process. Firstly, there was uncertainty regarding how to annotate specific terminology in other languages, particularly titles (e.g., the song title "Her Father Didn't Like Me Anyway", sentence id: w01130102), as it was observed that, in some PUDs, the annotators followed the syntactic rules of their own language, while others only used the 'flat' label. In this case, the decision was to annotate all of these instances from languages other than Galician, Portuguese, or Spanish with the 'flat' label (Figure 4), as usually recommended in the UD guidelines.[8] Apart from Galician, we decided to keep the structured annotation in Portuguese because there is mutual intercomprehension between the different varieties (i.e., Galician and Portuguese are generally considered belonging to the same language), and in Spanish, as practically all Galician speakers can also speak Spanish.

A similar case occurred with certain expressions or idioms, as some languages analyzed them as regular phrases while others used the 'fixed' label. In alignment with the previous case, the decision was to annotate these expressions with the 'fixed' label (Figure 5).

In view of this, and as previously stated, a final review is being conducted in order to verify the consistency of the annotations, drawing special attention to these ambiguous cases, prior to the submission of the PUD to the UD initiative.

4 Conclusions and further work

In this paper, we presented the development of the PUD treebank for Galician, aimed at being incorporated to the Universal Dependencies repository. The sentences have been translated by professionals, automatically annotated in a first stage, and manually reviewed by two linguists. This new resource will contribute to the NLP community by doubling the size of manually annotated treebanks for Galician.

Our study revealed that the agreement between annotators consistently improved as the annotation progressed, demonstrating a high level of agreement in the later stages of the corpus. Additionally, the Galician PUD annotation closely matches the previously available treebank with manual annotation for Galician.

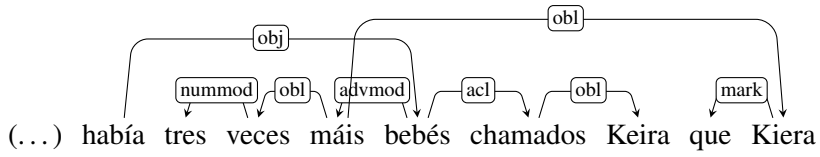
We also provide a brief discussion of various ambiguous cases during annotation, such as the annotation of comparative clauses or complex proper nouns in other languages, and present different solutions for them.

At the moment, we are carrying out a final review of the corpus, especially of those initial sen-

[6]Portuguese PUD examples (v2.13): sentence ids n01061016 and n05002004.

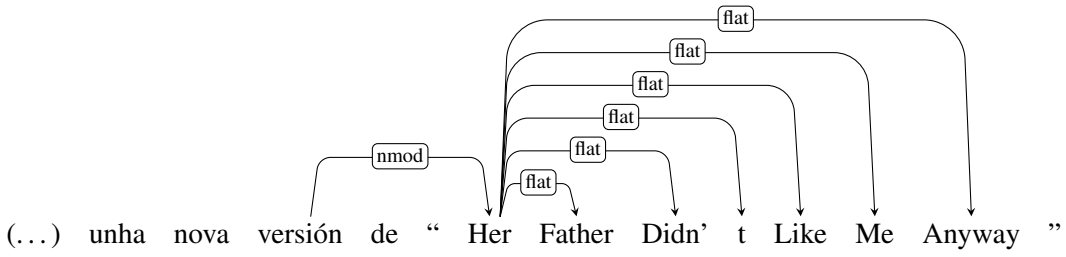
[7]English PUD examples (v2.13): sentence ids n01004017 and n04002020.

[8]https://universaldependencies.org/u/dep/flat.html



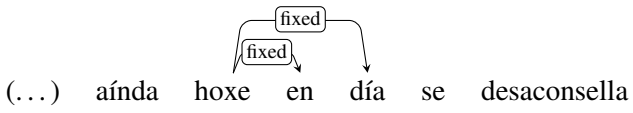
"(...) with three times more babies named Keira than Kiera"

Figure 3: Example of a comparative sentence and its annotation proposal (sentence id: n01015036).



"(...) a new version of (...) "Her Father Didn't Like Me Anyway""

Figure 4: Example of foreign terminology annotation; in this case, a song title (sentence id: w01130102). In this instance, 'flat' corresponds to 'flat:foreign' in the treebank, here simplified to facilitate visualization.



"(...) now it is still unadvisable"

Figure 5: Example of the multiword expression *hoxe en día* ('nowadays') labelled as 'fixed' (sentence id: w01095089).

tences with potentially less inter-annotator agreement. In future work, we plan to use the Galician PUD together with other parallel treebanks to explore cross-lingual analysis and to develop state-of-the-art parsers for this linguistic variety.

Acknowledgements

This research was funded by the Galician Government (ERDF 2014-2020: Call ED431G 2019/04, and ED431F 2021/01), by MCIN/AEI/10.13039/501100011033 (grants with references PID2021-128811OA-I00 and TED2021-130295B-C33, the latter also funded by "European Union Next Generation EU/PRTR"), and by a *Ramón y Cajal* grant (RYC2019-028473-I).

We would also like to thank Pablo Gamallo and Iria de-Dios-Flores for helpful discussions and feedback, and Sandra Rodríguez Rey and Helena Pérez Puente for their assistance with the translations.

References

Diego Alves, Marko Tadić, and Božo Bekavac. 2022. Multilingual comparative analysis of deep-learning dependency parsing results using parallel corpora. In *Proceedings of the BUCC Workshop within LREC 2022*, pages 33–42, Marseille, France. European Language Resources Association.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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

Pablo Gamallo, Daniel Bardanca, José Ramom Pichel, Marcos Garcia, Sandra Rodríguez-Rey, and Iria de Dios-Flores. 2023a. Nos_mt-opennmt-en-gl. https://huggingface.co/proxectonos/NOS-MT-OpenNMT-en-gl.

Pablo Gamallo, Daniel Bardanca, José Ramom Pichel, Marcos Garcia, Sandra Rodríguez-Rey, and Iria de Dios-Flores. 2023b. Nos_mt-opennmt-es-gl. https://huggingface.co/proxectonos/NOS-MT-OpenNMT-es-gl.

Marcos Garcia, Carlos Gómez-Rodríguez, and Miguel A Alonso. 2018. New treebank or repurposed? On the feasibility of cross-lingual parsing of romance languages with universal dependencies. *Natural Language Engineering*, 24(1):91–122.

Xavier Gómez Guinovart. 2017. Recursos integrados da lingua galega para a investigación lingüística. In Marta Negro Romero, Rosario Álvarez, and Eduardo Moscoso Mato, editors, *Gallæcia. Estudos de lingüística portuguesa e galega*, pages 1045–1056. Universidade de Santiago de Compostela.

Hildur Jónsdóttir and Anton Karl Ingason. 2020. Creating a parallel Icelandic dependency treebank from raw text to Universal Dependencies. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2924–2931, Marseille, France. European Language Resources Association.

Hiroshi Kanayama and Ran Iwamoto. 2020. How universal are Universal Dependencies? exploiting syntax for multilingual clause-level sentiment detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4063–4073, Marseille, France. European Language Resources Association.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Pritha Majumdar, Deepak Alok, Akanksha Bansal, Atul Kr. Ojha, and John P. McCrae. 2022. Bengali and Magahi PUD treebank and parser. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 60–67, Marseille, France. European Language Resources Association.

Dmitry Nikolaev, Ofir Arviv, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Maria Saeboe, and Omri Abend. 2020. Fine-grained analysis of cross-linguistic syntactic divergences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1159–1176, Online. Association for Computational Linguistics.

John E. Ortega, Iria de Dios-Flores, Pablo Gamallo, and José Ramom Pichel. 2022. A Neural Machine Translation System for Galician from Transliterated Portuguese Text. In *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations (SEPLN-PD 2022)*, volume 3224, pages 92–95. CEUR.

Lluís Padró. 2011. Analizadores multilingües en freeling. *Linguamática*, 3(1):13–20.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. 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*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Abdullatif Köksal, Balkiz Ozturk Basaran, Tunga Gungor, and Arzucan Özgür. 2019. Turkish treebanking: Unifying and constructing efforts. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 166–177, Florence, Italy. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.