A proposal for the accurate syntactic annotation of Galician

Xulia Sánchez-Rodríguez[1*], Albina Sarymsakova[2*] and Laura Castro[2]

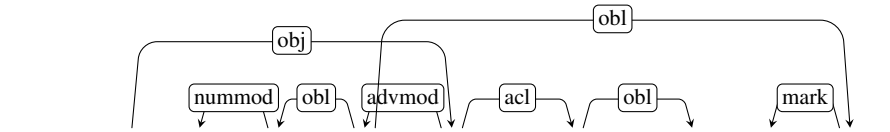
[1]Language Variation and Textual Categorisation (LVTC), Universidade de Vigo
[2]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela
[*]Equal contribution

Keywords: Galician, Syntax, Universal Dependencies, PUD

Parallel Universal Dependencies (PUD) treebanks are an essential component of the Universal Dependencies (UD) framework. They are a set of parallel corpora consisting of the same 1000 sentences aligned across languages and sourced from articles and Wikipedia. PUD treebanks exist in 23 languages and were developed for the *CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (Zeman et al., 2017).

In this project, we address different syntactic annotation challenges by proposing annotation guidelines tailored to the unique linguistic features of Galician—a language with limited resources and minimal representation within the UD framework. These ambiguities arose during the development of the Galician PUD (Sánchez-Rodríguez, Sarymsakova, Castro, & Garcia, 2024), a new manually annotated treebank for Galician which will be incorporated into the official PUD repository. Our proposals will focus on the following cases: comparatives, auxiliaries, and foreign terminology.

In the context of comparative sentences, given the absence at the time of a standardized model in the guidelines and a lack of consensus across languages within the PUD, we propose annotating the second part of comparative constructions with the 'obl' (oblique nominal) label, with dependence on the adverbial modifiers 'máis' ('more') or 'menos' ('less', 'fewer') (Figure 1). This approach is exemplified in various examples across languages from the first PUD edition (e.g., Portuguese and English PUD examples (v2.13) in sentence ids n01061016, n05002004, n01004017, and n04002020).



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

Figure 1: Example of a comparative sentence and its annotation proposal (id: n01015036).

Building upon the established definition of auxiliaries in UD and considering that the current Galician guidelines already encompass semi-copulative verbs like *semellar* ('to seem', 'to appear'), our proposal for the Galician PUD extends this scope to include additional verbs not accounted for as auxiliaries in TreeGal (Garcia, Gómez-Rodríguez, & Alonso, 2018), the only digital corpus with manual annotation of syntactic dependencies in Galician before the release of the new PUD. An illustrative example of this expansion can be observed in Figure 2, wherein the verb *parecer* (a synonym for *semellar*) is identified as an auxiliary.

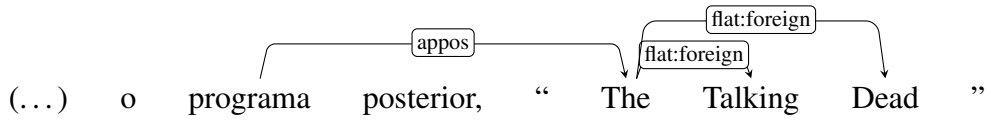
Regarding foreign terminology, it was noted that in some PUDs, annotators adhered to the syntactic conventions of their respective languages, while others simply applied the 'flat' label. To

(. . .) Hong Kong parece estarse preparando (. . .)

"Hong Kong (. . .) appears to be bracing [for a wave of protests] "

Figure 2: Example of our proposal to annotate the verb *parecer* ('to seem') as an auxiliary (id: n01101017).

address this issue, a decision was made to uniformly annotate all such instances from languages other than Galician, Portuguese, or Spanish with the 'flat:foreign' label (Figure 3), in line with the recommended practice outlined in the UD guidelines[1]. Notably, structured annotations were retained for Portuguese due to the mutual intelligibility between different varieties (i.e., Galician and Portuguese being generally regarded as belonging to the same language), and for Spanish, given that virtually all Galician speakers are proficient in Spanish as well.



"[New episodes are followed by] after show, "The Talking Dead""

Figure 3: Example of foreign terminology annotation; in this case, a show title (id: n01138026).

Participating in the first UniDive training programme would provide us with significant insights into automatic syntactic analysis and treebank creation for low-resource languages. Furthermore, the PARSEME corpus would provide an invaluable resource for parsing multiword expressions (MWEs), which are important but complex aspects in syntactic analysis. Access to such data would allow us to improve our parsing systems, potentially increasing their ability to accurately handle linguistic phenomena specific to low-resource languages such as Galician. Furthermore, the brainstorming hackathon and poster sessions are intended to promote new ideas and collaborations among participants. Engaging with fellow researchers would offer a unique opportunity to form networks and build connections that could potentially lead to future collaborations in the field of NLP, with a particular emphasis on addressing the issues experienced in low-resource language research.

Regarding the open questions for the brainstorming hackathon, we propose the following:

- How do the unique syntactic features of individual languages impact the transferability of parsing models trained on cross-linguistic data, and what strategies can be employed to mitigate potential challenges in model adaptation?
- How can the annotation of specific syntactic phenomena in low-resource languages contribute to the improvement of parsing models trained on data from high-resource languages?

At present, our project is in a development phase. Despite the fact that the Galician PUD already has its first version completed and will be released in the v2.14 of Universal Dependencies on 15 May, our objective is to continue enhancing its annotation through the development of more accurate and effective natural language processing applications for this language, as well as the revision of automatic annotation for future versions of the corpus.

[1] `https://universaldependencies.org/`

References

Garcia, M., Gómez-Rodríguez, C., & Alonso, M. A. (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.

Sánchez-Rodríguez, X., Sarymsakova, A., Castro, L., & Garcia, M. (2024, March). Increasing manually annotated resources for Galician: the parallel Universal Dependencies treebank. In P. Gamallo et al. (Eds.), *Proceedings of the 16th international conference on computational processing of portuguese* (pp. 587–592). Santiago de Compostela, Galicia/Spain: Association for Computational Lingustics. Retrieved from https://aclanthology.org/2024.propor-1.65

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., ... Li, J. (2017, August). 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* (pp. 1–19). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/K17-3001 doi: 10.18653/v1/K17-3001