The Spanish Learner and Heritage Speaker Dependency Treebank

Valeria Pagliai **Sergio José Salazar Rodó**

Emiliana Pulido **Andres Gutierrez-Quintero**

Zoey Liu
University of Florida
liu.ying@ufl.edu

Introduction Data-driven analysis of syntactic development in second language (L2) acquisition requires learner corpora with (morpho)syntactic annotations (Wulff, 2017). Such datasets, however, remain scarce. While a few L2 corpora, focusing on English (e.g., the Learner English Treebank (TLE) (Berzak et al., 2016)), Italian (the learner Italian treebank (Di Nuovo et al., 2019)), or Korean (L2 Korean treebank (Sung and Shin, 2024)) provide syntactic annotations, they lack systematic documentation of learner deviations, i.e., linguistic patterns that reflect the learning and adoption of a language during the acquisition process.

To address these gaps, we present a manually curated L2 Spanish dataset (N = 49,247) based on the COWS-L2H corpus (Davidson et al., 2020). We adopted the Universal Dependencies (UD) (de Marneffe et al., 2021) framework, which offers clear annotation guidelines that are cross-linguistically adaptable and have been applied in previous L2 corpora (see above). Our annotations cover lemmatizations, part-of-speech (POS) tags, and syntactic dependencies. Compared to existing L2 corpora, our treebank: (1) captures learner deviation by considering instances of pro-drop, ungrammatical structures, and several dependency subrelations; (2) includes learners of different proficiency levels; (3) offers an additional dedicated annotation manual that can be straightforwardly employed and expanded by future research. Although there exists one Spanish L2 treebank (Pulido et al., 2025), it was annotated by a single annotator with 6,604 tokens only and contains no information of learner deviations. In contrast, our dataset constitutes the largest syntactic treebank for L2 Spanish with detailed annotations, which will be freely accessible.

Annotation process Four graduate students specializing in linguistics worked on the annotations. We first conducted a pilot phase in which we annotated 30 sentences randomly selected from the Spanish AnCora treebank (AnCora) (Taulé et al., 2008) from scratch. Inter-annotator agreement across the features of interest, including lemmas, POS tags, and syntactic dependencies, reached at least 90% on average. During this stage and throughout the initial phase of learner data annotations, one annotator constructed a detailed annotation manual providing instructions and examples on how to consistently address learner data.

We used the COWS-L2H corpus, which consists of 5,383 essays written by Spanish learners across different classes and instructional levels. The corpus includes lower, intermediate, advanced, and heritage speakers courses. We include at least one essay (N=168) from each of the 17 different courses, covering 165 unique student authors out of a total of 167, as some students contributed to multiple essays. After automatically parsing the learner essays with `spaCy`, annotators corrected parsing errors following the project's annotation guidelines. We added miscellaneous information about ungrammatical constructions, typos, and language identification. Finally, annotators also explicitly marked missing tokens in all sentences, including omissions of non-overt subjects (i.e., pro-drop).

Experiments Due to space limitation, we focus on the analysis of missing tokens and dependency parsing. Overall, the resulting treebank includes 2,415 missing tokens, the distribution of which appears to be influenced by both the POS tag of the omitted token and the proficiency level (Table 1): learners seem to drop pronouns (77.97%) and prepositions (9.65%) much more often than other POS tags. The frequency of missing tokens gradually decreases from lower to upper level classes, indicating potential improvement through the stages of acquisition.

With dependency parsing, we examined two data partitioning strategies coupled with different data representations. One split followed UD's guideline of including at least 10k tokens in the test set, and the other was based on the commonly used 8:1:1 ra-

POS tag	Freq.
PRON	1,883
ADP	233
DET	99
PUNCT	50
Other	152

Level	Freq
Lower	1,075
Intermediate	905
Heritage	266
Upper	169

Table 1: Missing tokens and pro-drop across POS tags and class levels.

tio. For data representation, we experimented with prioritizing essays from upper-proficiency learners and from a combination of upper and intermediate learners in the training set. Additionally, we also explored excluding missing tokens from a given data split, as well as including our explicit annotations of such tokens. We used Label Attachment Scores (LAS) as the evaluation metric.

We carried out three training configurations, specifically, training parsers: (1) exclusively on AnCora, (2) solely on our manually annotated L2 data, and (3) on AnCora and subsequently fine-tuning on the L2 data. All models were trained with the MaChAmp toolkit (van der Goot et al., 2021) using the multilingual BERT-cased model (Devlin et al., 2019) and the default training parameters.

	AnCora	L2 treebank	Finetune AnCora
UD partition			
no empty tokens	0.858	0.866	0.875
with empty tokens	0.808	0.860	0.870
Regular partition			
no empty tokens	0.853	0.858	0.876
with empty tokens	0.814	0.865	0.877

Table 2: LAS scores from different partitioning strategies and training configurations, when training data contains both upper and intermediate learner data.

The different partitioning strategies and data representation yield reasonable LAS scores, ranging from 0.790 to 0.877. Table 2 presents results for when the training set consists of mostly upper and intermediate-level learner essays. With the UD partition, it seems that when excluding missing tokens from our annotated dataset, the parser derived from AnCora yields relatively good result (LAS=0.858); that said, we observe comparable performance when the parser was trained on the L2 data alone, which is approximately 13 time smaller compared to the AnCora training set. This demonstrates the effectiveness of our in-domain annotations. Finetuning Ancora with our L2 training data did not notably improve model performance, which potentially points to pronounced structural and distributional differences between Spanish L1 and L2 writing. Similar trends are present when missing tokens are explicitly labeled, across data splits and training configurations.

References

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal dependencies for learner English. In *Proceedings of the 54th Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 737–746.

Sam Davidson, Aaron Yamada, Paloma Fernandez Mira, Agustina Carando, Claudia H. Sanchez Gutierrez, and Kenji Sagae. 2020. Developing NLP tools with a new corpus of learner Spanish. In *Proceedings of the Twelfth LREC*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the NAACL:HLT, Volume 1 (Long and Short Papers)*.

Elisa Di Nuovo, Cristina Bosco, Alessandro Mazzei, Manuela Sanguinetti, and 1 others. 2019. Towards an Italian learner treebank in universal dependencies. In *CEUR Workshop Proceedings*, volume 2481, pages 1–6.

Emiliana Pulido, Robert Pugh, and Zoey Liu. 2025. I speak for the árboles: Developing a dependency treebank for Spanish L2 and heritage speakers. In *Proceedings of the 63rd Annual Meeting of the ACL (Volume 4: Student Research Workshop)*.

Hakyung Sung and Gyu-Ho Shin. 2024. Constructing a dependency treebank for second language learners of Korean. In *Proceedings of LREC-COLING 2024*.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on LREC*.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the EACL: System Demonstrations*.

Stefanie Wulff. 2017. What learner corpus research can contribute to multilingualism research. *International Journal of Bilingualism*, 21(6):734–753.