I Speak for the Árboles: Developing a Dependency Treebank for Spanish L2 and Heritage Speakers

Anonymous ACL submission

Abstract

We introduce the first set of Universal Dependencies (UD) annotations for Spanish learner writing from the UC Davis COWSL2H corpus. Our annotations include lemmatization, POS tagging, and syntactic dependencies. We adapt the existing UD framework for Spanish L1 to account for learner-specific features such as code-switching and non-canonical syntax. A suite of parsing evaluation experiments shows that parsers trained on learner data together with moderate sizes of Spanish L1 data can yield reasonable performance. Our annotations and parsers will be openly accessible to motivate future development of learner-oriented language technologies.

1 Introduction

Morphosyntactic information for learner data has the potential to benefit a variety of research topics, ranging from characterizing morphological production, modeling the syntactic developmental trajectory of language learners, to advancing natural language processing (NLP) tools tailored specifically for learners and their education (Meurers and Dickinson, 2017). Datasets consisting of learner production manually annotated with morphosyntactic features, however, are relatively scarce (Kyle, 2021; Sung and Shin, 2024).

The current paper contributes to this research gap by developing a dependency treebank for Spanish second-language (L2) and heritage speakers. We choose Spanish given its status as an extremely important L2 for students with varied educational backgrounds (U.S. Census Bureau, 2013). Our annotations follow the framework of Universal Dependencies (UD) (Zeman et al., 2024), a substantially community-led project addressing the need for consistent and cross-linguistic annotation. Although numerous grammatical frameworks exist, we employ UD because of the continuous collaborative efforts devoted to its expansion, which ensures the sustainability of its annotation guidelines and developed resources. Additionally, there exists UD treebanks for Spanish first-language (L1) data (e.g. Ancora (Taulé et al., 2008)) along with treebanks for a few other L2s such as English (Kyle, 2021) and Korean (Sung and Shin, 2024). These resources help guide our own annotations.

Description	Count
Total number of annotated essays	23
Total number of tokens	6,604
Total number of sentences	383
Total number of topics	8
Total number of levels	20

Table 1: Descriptive statistics for our treebank.

To that end, we use the publicly accessible UC Davis Spanish learner corpus, COWSL2H[1], which has writing samples collected from college students enrolled in Spanish courses of varying proficiency levels. Our treebank consists of 23 essays across 8 topics and 20 distinct course levels randomly sampled from COWSL2H, totaling 383 sentences and 6,604 tokens (Table 1). We adapt the UD framework for Spanish L1 with morphosyntactic features such as code-switching and production errors commonly found in learner production. In particular, we provide manual annotations and develop models at three linguistic levels: lemmas, part-of-speech (POS) tags and syntactic dependencies.

2 Related Work

Standard NLP tools often yield worse performance on learner corpora, particularly when models trained on native-speaker data are applied to non-native input or other out-of-domain texts with differing linguistic characteristics (McClosky et al., 2006). This performance gap has motivated researchers and the community to build non-native corpora to support more generalizable models.

[1] https://github.com/ucdaviscl/cowsl2h

With dependency treebank specifically, one of the first scalable efforts to annotate bilingual learner (written) data was for English by Berzak et al. (2016a), who developed the Treebank of Learner English (TLE) (Berzak et al., 2016b) following UD. This treebank includes parallel annotations of both the original learner sentences and corrected versions which provides for a comparative framework. Follow-up study by Kyle et al. (2022) expanded dependency annotations to spoken discourse by L2 English speakers learner.

Subsequent work expanded to other L2s. The Korean L2 treebank by Sung and Shin (2024) includes over 7,500 annotated sentences from learner essays. Their work involved adapting UD guidelines to Korean's agglutinative structure and possible morphological errors. Li and Lee (2020) developed a parallel UD treebank for L2 Chinese, consisting of 600 learner sentences and 697 corrected targets from intermediate-level narrative writing. Each sentence pair was manually annotated with POS, heads, and dependency relations, enabling contrastive syntactic analysis of L2 productions. Lastly, Di Nuovo et al. (2019) introduced an UD-guideded Italian learner treebank with automated parsing and manual post-editing.

Although there are a number of Spanish L2 datasets (e.g., CAES (Miaschi et al., 2020), CEDEL2 (Lozano, 2021)), none (including COWSL2H) provides UD-style morphosyntactic annotations. Aside from COWSL2H, other aforementioned datasets do not include heritage speaker data. We hope that contingent on gradual expansion of data availability and our annotation framework, future work will be able to computational assess the structural differences in the production between L2 and heritage speakers (Montrul, 2010).

3 Annotation guidelines and process

While annotations for lemmas and POS tags were relatively more straightforward, challenges arose when annotating syntactic dependencies. Our annotation guidelines mainly followed the UD framework (Nivre et al., 2020), especially the annotation schemes of Ancora (Taulé et al., 2008). For instance, we adopted AnCora's guidelines regarding the removal of the *iobj* dependency relation with regards to prepositional indirect objects. Albeit with these references, we had to use our best judgment when encountering learner constructions that were not clearly addressed in existing guidelines. For sentences that were long and continuous that lacked punctuation and conjunctions, we used *parataxis* to connect the heads of the subclauses. We also adopted obl:tmod (Zeldes and Schneider, 2023) to distinguish temporal modifiers from their parent *obl*. Additionally, we purposefully tried to avoid assigning *dep* (unspecified dependency), despite that phrases containing errors can obscure syntactic or semantic interpretation of the sentence; and instead, we manually reassigned a more specific label based on syntactic context.

Since spelling errors are common in learner writing, we kept the original misspellings in the FORM column (Table 2) to reflect what the student actually wrote. When the intended word was clear, we corrected it in the LEMMA column to keep things consistent for downstream tools like lemmatizers and parsers. For instance, in the sentence "*El pasisaje es fenomenal* (The scenery is phenomenal)", we kept *pasisaje* as the FORM but used *paisaje* as the LEMMA. This way, we balance staying true to learner output with keeping the data clean and usable.

ID	FORM	LEMMA	UPOS	DEPREL
1	El	el	DET	det
2	pasisaje	paisaje	NOUN	nsubj
3	es	ser	AUX	cop
4	fenomenal	fenomenal	ADJ	root
5	.	.	PUNCT	punct

Table 2: Example of a learner spelling error preserved in the FORM column but corrected in the LEMMA column.

Most likely due to Spanish being the heritage or second language of the university students, there were code-switched sentences with certain words or phrases being in English. We followed the guidelines of the UD English Web Treebank (EWT) for those specific tokens (Silveira et al., 2014).

The specific guidelines were developed in a continuous manner mostly by Annotator A, an undergraduate double majoring in Linguistics and Psychology who is a heritage speaker of Spanish. Idiosyncratic cases in early annotation stages were discussed among all authors to refine the guidelines. Annotator A continued to annotate the full treebank. 48 sentences (805 tokens) were cross-annotated by Annotator A and Annotator B, who is a doctoral candidate in computational linguistics. Disagreements were resolved through discussion. Table 3 shows the inter-annotator agreement; only one lemma disagreement was recorded[2].

[2]For *"A el crecer, me sentí tan mal por mi misma y seria*

Annotation	Agreement Score
POS tag	0.98
Syntactic head	0.93
Syntactic deprel	0.91
Syntactic head+deprel	0.88

Table 3: Annotator agreement scores for POS tagging and syntactic annotations.

Metric	learner_only	ancora_only	ancora+learner
LAS	0.792	0.816	**0.824**
UAS	0.854	**0.890**	**0.890**
Lemma Acc.	0.938	0.971	**0.983**
UPOS Acc.	**0.976**	0.972	0.973

Table 4: Parser performance across training schemes.

4 Parsing Experiments

We randomly split our treebank into training and test set at a 4:1 ratio. We then developed three different parser models using different training data representation: (1) `learner_only`, trained exclusively on our small set of hand-annotated learner data (5k tokens)[3]; (2) `ancora_only`: trained on the entire AnCora training set. (3) `ancora+learner`, trained on the combination of the learner data and the full AnCora training set (453k tokens).

Each model jointly performed lemmatization, POS tagging, and dependency parsing. Each model was built using the default parameters of the MaChAmp toolkit (van der Goot et al., 2021), which fine-tunes contextual subword embeddings from a pretrained model (we used multilingual BERT (Devlin et al., 2019) on multiple tasks simultaneously). All tasks shared encoder parameters, but each had its own unique decoder: a transformation-rule classifier (Straka, 2018) for lemmatization, a softmax layer on the contextual embeddings for POS tagging, and a deep biaffine parser for dependency parsing (Gardner et al., 2018). We used accuracy as the evaluation metric for lemmatization and POS tagging, and both labeled and unlabeled attachment score (UAS/LAS) for dependency parsing.

5 Results and Discussion

As shown in Table 4, `learner_only` model achieved reasonable performance across the three tasks, and only lagged mildly behind `ancora_only` in some cases. This is particularly encouraging given that the training data for `learner_only` is almost 90 times smaller.

While POS accuracy is comparable between `learner_only` and `ancora_only`, lemma accuracy was notably weaker for `learner_only` (0.938 vs. 0.971). Manual inspection of parser predictions revealed the performance discrepancies largely resulted from `learner_only` mishandling lemmas for irregular verbs, which occur much less frequently in the learner training data due to size limitation. For example, the parser failed to learn root alternations, such as with *hizo* (past tense of "did") in Figure 1, where the correct lemma is *hacer* ("do"), but the `learner_only` model incorrectly predicted *hier*. This pattern somewhat mimics human learner behavior, overgeneralizing inflectional rules without lexical anchoring, a characteristic of early interlanguage development (Andringa and Rebuschat, 2015).

Aside from excessive productive suffixing (e.g., *-ar* inflections on verb classes), the `learner_only` model produced non-standard lemmas that are not attested in Spanish (e.g., *pudieer* and *sintiar*). These errors show that the model failed to restrict inference to grammatically well-formed lexical stems, a common issue in low-resource lemmatization (Kanerva et al., 2018; Mielke et al., 2021). However, this model also overapplies morphological rules in ways even human learners tend to avoid. For example, *sintió* (for *sentir*; "to feel") was lemmatized into *sintiar*, an imaginary form ending in *-iar*. The present participle verb *comiendo*, was mislemmatized as comier (should be *comer* which means "to eat"), likely due to confusion with the subjunctive form *comiera* or stem truncation, when the output is an incomplete root, omitting part of the predicted verb stem. These errors reflect the difficulty in predicting irregular morphology and tense variation.

For dependency parsing, `ancora_only` achieves moderately better performance compared to `learner_only`. The `learner_only` parser struggled more with dependency relations involving structural ambiguity or deeply embedded clauses, which are common in L2 writing. These sentences often lack clear punctuation or use repetitive structures, making it harder to identify clause boundaries and syntactic roles. Dependency relations like

tan insegura también.", Annotator B initially labeled *seria* as the verb *ser* (conditional), while Annotator A took it as the adjective *serio* ("serious"). We ultimately interpreted it as a misspelling of *era*, aligning better with the sentence's tense and meaning, and selected *ser* as the final lemma.

[3]To avoid unnecessary unseen tokens, we replaced the named entity placeholders (e.g., "`*FIRST_NAME*`") with standardized names.

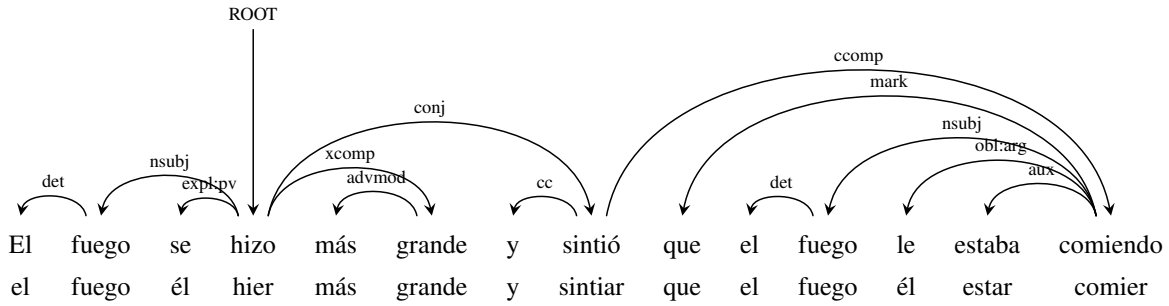

Figure 1: model-predicted dependency tree with predicted lemmas for the above sentence. Translation: "The fire grew larger, and they felt like the fire was consuming them." Punctuation not included due to spacing.

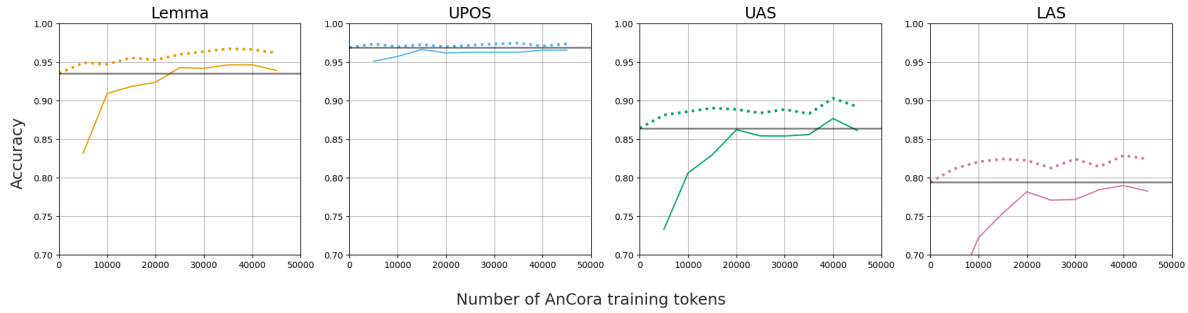

Figure 2: Learning curves of model performance across the three tasks with different training data representations; in each subfigure, the solid curve represents the performance from training data of different sizes subsampled from Ancora; the dash curve corresponds to the performance from the combination of the aforementioned Ancora subsamples with our learner training set; the solid horizontal line is the performance of the `learner_only` model, which remains constant given that the size of the learning training data is fixed.

`advcl`, `obl:arg`, and `xcomp` were particularly susceptible. For instance, in *"...a mi padre le dieron un premio"* ("...my father was given an award,") the gold label correctly assigns `obl:arg` to *padre*, reflecting its role as the receiver of the action. However, `learner_only` incorrectly labeled it as `nsubj`, failing to account for the fact that the subject of the verb *dieron* is implicit and not overtly expressed. This misclassification illustrates how the model overgeneralized subject role in the absence of explicit syntactic cues.

Across the three tasks, we have the best performance with `ancora+learner`. That said, its performance is mostly comparable to that of `ancora_only`. The lack of notable improvement between `ancora+learner` and `ancora_only`, raises the question of whether the predominant representation of Spanish L1 in the training data for `ancora+learner` hinders the model from learning observations in L2 production. To address this, we experimented with subsampling from Ancora datasets of different sizes ({5k, 10k, 15k, ..., 45k} tokens) then combining them individually with the learner training data to build parsers. The learning curve in Figure 2 shows that model performance does not improve consistently with more training data, but rather shows early increases up until 30-40k tokens followed by plateauing trends. Both UAS and LAS saw improvement up to 15k tokens, from 0.86 to 0.89 and 0.79 to 0.82, respectively. After this point, improvements were reduced, with UAS reaching a high of 0.90 at 40k tokens before plateauing. Lemma accuracy saw an early increase (from 0.94 to 0.96 by 15k tokens) to finish at 0.97 near 35k. UPOS tagging starts high at 0.969 and remains relatively stable with slight fluctuations.

Collectively, our study shows that even a modest amount of in-domain learner data can obtain reasonable performance, especially when combined with additional out-of-domain data. The observations here also suggest that training size does not always need to be bigger – instead, data representation that is possibly less affected by size can have a meaningful impact on model performance. We leave further investigation for future work.

6 Limitations

One limitation is the absence of manual morphological annotations, which we plan to add in future work. Including tags like `Typo=Yes` and `CorrectForm`, as in standard UD treebanks, would improve interpretability. Another limitation is the small corpus size, which led to many unseen forms, especially irregular or learner-specific ones, reducing lemmatization and parsing stability, a common challenge in low-resource NLP settings.

References

Sible Andringa and Patrick Rebuschat. 2015. New perspectives on the role of practice in second language learning. In Patrick Rebuschat, editor, *Implicit and Explicit Learning of Languages*, volume 48 of *Studies in Bilingualism*, pages 91–114. John Benjamins, Amsterdam.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016a. Treebank of learner english (tle).

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

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 North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Elisa Di Nuovo, Cristina Bosco, Alessandro Mazzei, and Manuela Sanguinetti. 2019. Towards an Italian learner treebank in universal dependencies. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR-WS.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2018. Universal lemmatizer: A sequence to sequence model for lemmatizing universal dependencies treebanks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 139–150.

Kristopher Kyle. 2021. Natural language processing for learner corpus research. *International Journal of Learner Corpus Research*, 7(1):1–16.

Kristopher Kyle, Masaki Eguchi, Aaron Miller, and Theodore Sither. 2022. A dependency treebank of spoken second language English. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 39–45, Seattle, Washington. Association for Computational Linguistics.

Yuxin Li and John Lee. 2020. L1-l2 parallel dependency treebank for learners of Chinese as a foreign language. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 901–909, Marseille, France. European Language Resources Association.

Cristóbal Lozano. 2021. CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*, 0(0):02676583211050522.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Reranking and self-training for parser adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 337–344, Sydney, Australia. Association for Computational Linguistics.

Detmar Meurers and Markus Dickinson. 2017. Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning*, 67(S1):66–95.

Alessio Miaschi, Sam Davidson, Dominique Brunato, Felice Dell'Orletta, Kenji Sagae, Claudia Helena Sanchez-Gutierrez, and Giulia Venturi. 2020. Tracking the evolution of written language competence in L2 Spanish learners. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–101, Seattle, WA, USA' Online. Association for Computational Linguistics.

Sabrina J. Mielke, Tal Linzen, and Jason Eisner. 2021. What kind of knowledge is captured by contextualized word representations? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 1250–1265.

Silvina Montrul. 2010. How similar are adult second language learners and Spanish heritage speakers? Spanish clitics and word order. *Applied psycholinguistics*, 31(1):167–207.

J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4027–4036.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for english. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Milan Straka. 2018. Udpipe 2.0 prototype at CoNLL 2018 ud shared task. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL) 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207.

Hakyung Sung and Gyu-Ho Shin. 2024. Constructing a dependency treebank for second language learners of Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3747–3758, Torino, Italia. ELRA and ICCL.

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 Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

U.S. Census Bureau. 2013. Spanish, chinese top non-english languages spoken; most of population is english proficient. Accessed: 2025-04-29.

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 European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Amir Zeldes and Nathan Schneider. 2023. Are UD treebanks getting more consistent? a report card for English UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 58–64, Washington, D.C. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Arofat Akhundjanova, Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Matthew Andrews, and 633 others. 2024. Universal dependencies 2.15. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.