Integrating Geographically Diverse Low-Resourced Spanish Varieties into Universal Dependencies

Johnatan E. Bonilla, Humboldt Universität zu Berlin, Germany; Instituto Caro y Cuervo, Colombia

Keywords: Spoken Spanish, Universal Dependencies, Dialectology, Low-resourced dialects

Description of the Resource

Globally, over 599 million people speak Spanish. However, Spanish Universal Dependencies (UD) treebanks such as AnCora[1] (Taulé, 2008; Martínez & Zeman, 2016) and COSER-UD[2] (Bonilla, 2024) have predominantly documented the varieties from Spain. Despite Spain's significant cultural impact, its approximately 47.7 million inhabitants represent only about 8% of Spanish speakers worldwide. This project seeks to improve the representation of Spanish in the UD project by focusing on both rural and urban Spanish varieties from Latin America. The initiative integrates data from various sources, including the *América y España Español Coloquial corpus*[3] (AMERESCO, 'Coloquial American and Spanish Corpus', Albelda & Estellés, 2020), the *Proyecto para el Estudio Sociolingüístico del Español de España y América*[4] (PRESEEA, 'Project for the Sociolinguistic Study of Spanish from Spain and America', Moreno-Fernández, 2005), and the oral corpus from the *Atlas Lingüístico Etnográfico de Colombia*[5] (ALEC, 'Colombian Linguistic and Ethnographic Atlas', Bonilla, 2020). This expansion aims to provide a more comprehensive and representative dataset of the Spanish language as it is used globally.

The project, preliminarily named SVarT (Spanish Varieties Treebank), aims to compile an initial treebank containing 10 to 20 sentences from interviews conducted in a range of countries including Argentina, Bolivia, Chile, Colombia, Cuba, Ecuador, Honduras, Panama, Mexico, the United States, Peru, Venezuela, Uruguay, Puerto Rico, Guatemala, and Paraguay. Sentence selection will be made using semi-automatic methods that prioritize morphosyntactic variation identified as significant within dialectological literature. For example, Spanish from Argentina, Chile, Bolivia, and certain areas of Colombia will include examples of voseo, where the pronoun "vos" replaces "tú" (‘you’) for the informal second person singular, with corresponding verb forms such as "tenés" instead of "tienes" (‘have’). This linguistic phenomenon, alongside other morphosyntactic phenomena such as variations between compound and simple past tenses, distinctions in the use of "leísmo" and "loísmo", the adverbialization of adjectives, postposition of possessive adjectives, and the use of the periphrastic future tense employing constructions like "ir a" followed by an

[1] https://universaldependencies.org/treebanks/es_ancora/index.html
[2] https://github.com/UniversalDependencies/UD_Spanish-COSER
[3] https://esvaratenuacion.es/ameresco
[4] https://preseea.uah.es/
[5] https://clicc.caroycuervo.gov.co/corpus/ALEC

infinitive, are currently under-documented in existing resources like AnCora or COSER-UD. SVarT aims to address these gaps by focusing on these complex linguistic features, thereby enriching the representation of Spanish language diversity in linguistic datasets.

How the participation in the training school will be useful for the project

Participating in the 1st UNIDIVE Training Summer School will greatly benefit the development of the SVarT project. Despite my experience with the COSER-UD, I still encounter challenges with the annotation frameworks, especially in handling code-switching annotation in regions where Spanish intersects with indigenous languages or specific UD guidelines like the treatment of Spanish indirect object as oblique. This interaction will help clear up specific annotation challenges and reveal how similar syntactic phenomena are managed in treebanks of other languages.

Additionally, my goal to add a layer of manually searched idioms and colloquial multiword expressions to the SVarT and COSER-UD projects is an area I have not yet explored. Moreover, I'm interested in understanding how to use advanced collaborative tools for annotation infrastructure to make the annotation process more efficient and facilitate better data management and researcher collaboration. Sharing experiences and challenges with other treebank creators at the summer school will inspire innovative approaches and foster future collaborations, enhancing the SVarT project and ensuring its success as a comprehensive resource for the linguistic community.

Open Questions for Brainstorming Hackathon

- Dialectological Linguistic Approach for Sample Selection: Would using a dialectological linguistic approach help us better identify and select sentences that show a wide range of morphosyntactic diversity from the treebank?
- Transfer Learning and Its Limits: What are the challenges of using existing resources like AnCora and COSER-UD, or multilingual parsers to tag and parse sentences from newly integrated resources? Where do these methods fall short?
- Discourse Marker Consistency Across Resources: Given the varied use of discourse markers in COSER-UD, ALEC, AMERESCO, and PRESEEA, how can we standardize these markers to ensure consistency and usefulness across different datasets? This question also applies to treebanks created with transcription from spoken sources.

Current Project Phase: Data Gathering and Preprocessing

To date, the ALEC dataset has been fully compiled, with 40 sentences parsed, tagged, and included in version 2.14 of the COSER-UD treebank, although sentences were not selected based on specific morphosyntactic features. The AMERESCO dataset is systematically

arranged on the Hugging Face platform[6], and preparations are underway to similarly organize the PRESEEA dataset. This phase also includes a thorough examination and refinement of transcription protocols to ensure uniformity and accuracy across all datasets.

References

Albelda, M. y Estellés, M. (2020): Corpus Ameresco, Universitat de València, ISSN: 2659-8337, www.corpusameresco.com.

Bonilla, J. E. (2024). Universal Dependencies for Spoken Spanish. [Doctoral dissertation, Ghent University].

Bonilla, J. E., Rubio López, R. Y., Llanos Chávez, A. L., Bejarano Bejarano, D. E., & Bernal Chávez, J. A. (2020). Proyecto de digitalización y nuevas perspectivas del Atlas Lingüístico-Etnográfico de Colombia. In *Dialectología digital del español* (Vol. 80, pp. 13-28). Universidade de Santiago de Compostela.

Moreno-Fernández, F. (2005). Corpus para el estudio del español en su variación geográfica y social. El corpus PRESEEA. *Oralia: Análisis del discurso oral*, *8*, 123-139.

Martínez-Alonso, H., & Zeman, D. (2016). Universal Dependencies for the AnCora treebanks. *Procesamiento del Lenguaje Natural*, (57).

Taulé, M., Martí, M. A., & Recasens, M. (2008, May). Ancora: Multilevel annotated corpora for Catalan and Spanish. In *Lrec* (Vol. 2008, pp. 96-101).

[6] https://huggingface.co/datasets/johnatanebonilla/ameresco-asr