Integrating Geographically Diverse Low-Resourced Spanish Varieties into Universal Dependencies

Johnatan E. Bonilla
Language and Translation Technology Team (Lt3) - Ghent University
Institute für Romanistik - Humboldt-Universität zu Berlin
johnatan.bonillahuerfano@ugent.be

Lucia Amorós
Centre for Studies on Educational Memory (CEME) - Murcia University
International University of La Rioja
lamoros@um.es

Date: 8-12 July 2024

UD Spanish

AnCora (567K)

GSD (431K)

PUD (23K)

→ Advanced PoS taggers and parsers for Spanish are only trained on written texts from normative/standard varieties (e.g. Madrid).

→ There are no open-source treebanks available for spoken Spanish varieties.

Corpus Oral y Sonoro del Español Rural (COSER)
(Fernández-Ordóñez 2005-)
'Audible Corpus of Spoken Rural Spanish'

3009 interviews

254 locations

3.3M transcribed words

Manual validation
PoS tags (2020-2022)

Crowdsourcing validation GWAPS
PoS tags (2020-2022)

Tesoros

Anotatlón

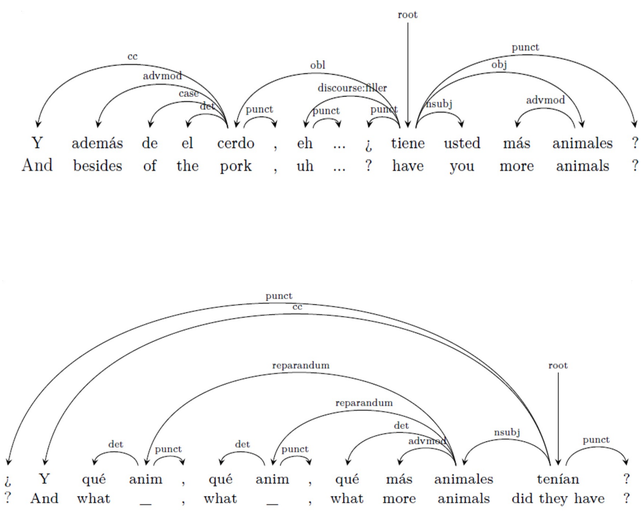
Agentes

www.juegosdelespanol.com

	COSER-UPOS Benchmark		COSER-UD Treebank	
Region	**Sentences**	**Tokens**	**Sentences**	**Tokens**
Andalusia	982	14217	30	383
Aragon	635	8264	30	322
Asturias	787	12454	31	532
Balearic Islands	617	11215	30	505
Canary Islands	1170	17272	30	376
Cantabria	664	8746	30	377
Castile	722	10542	30	461
Catalonia	892	12122	30	328
Extremadura	1118	16912	31	451
Galicia	942	16315	30	461
Madrid	596	9161	30	312
Castilla-La Mancha	731	10326	30	411
Murcia	612	7954	30	513
Navarre	839	10650	22	322
La Rioja	600	9841	20	568
Valencian Community	692	11241	20	451
Basque Country	620	9150	20	388
Total	**13219**	**196372**	**539**	**8158**

https://github.com/UniversalDependencies/UD_Spanish-COSER

https://github.com/johnatanebonilla/COSER-PoS.v2



→ **Expand the COSER-UD treebank** with the COSER-PoS GS.

→ **Expand Spanish intralinguistic diversity in UD:** Spain represents only 8% of Spanish speakers worldwide.

New project - Spanish varieties Treebank (SvarT)

- Involves participation from Instituto Caro y Cuervo and Cartagena and Antioquia Universities from Colombia, Murcia University from Spain, Humboldt-Universität zu Berlin.
- Integrates data from various open source corpus of rural and urban spoken Spanish (AMERESCO, PRESEEA, ALEC).
- Sentences are selected using semi-automatic methods prioritizing significant morphosyntactic variation from low-resource varieties (e.g., voseo, periphrastic future).
- PoS tag and parse simultaneously

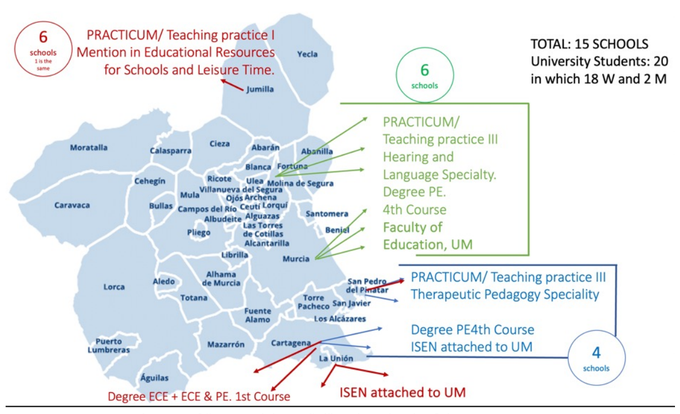
```
# sent_id = ALEC-ALEC_C3_A30_4-148
# time = 00:19:56,920--> 00:20:45,300
# location = Colombia-Antioquia-Ituango
# turn_text = I16: El conejo le dijo: "Quedáte aquí y cuando yo te pegue el grito te abrís de patas y manos, abrís la boca y cerrás los ojos".
# text = I16: El conejo le dijo: "Quedáte aquí y cuando yo te pegue el grito te abrís de patas y manos, abrís la boca y cerrás los ojos".
# text_orth = I16: El conejo le dijo: "Quedáte aquí y cuando yo te pegue el grito te abrís de patas y manos, abrís la boca y cerrás los ojos".
# text_en = I16: The rabbit told him: "Stay here and when I scream at you, open your legs and hands, open your mouth and close your eyes."
# lang = es

# sent_id = AMERESCO-BUE_001_03_19-537
# time = 00:24:23.11-00:24:26.50
# location = Argentina-Buenos Aires-Buenos Aires
# turn_text = C: ¿vos seguís a la<alargamiento/>/ la Bonaerense?/ creo que ya [hablamos]
# text = C: ¿vos seguís a la<alargamiento/>/ la Bonaerense?/ creo que ya [hablamos]
# text_orth = C: ¿vos seguís a la<alargamiento/>/ la Bonaerense?/ creo que ya [hablamos]
# text_en = C: Do you follow the <lengthening/>/ the Buenos Aires?/ I think we already [talked]
# lang = es

# sent_id = AMERESCO-VVU_001_02_17-12
# time = 00:00:47.02-00:00:48.77
# location = Bolivia-Santa Cruz-Santa Cruz
# turn_text = B: iba a salir más tarde/ [vos siempre] salís tarde
# text = B: iba a salir más tarde/ [vos siempre] salís tarde
# text_orth = B: iba a salir más tarde/ [vos siempre] salís tarde
# text_en = B: I was going to leave later/ [you always] leave late
# lang = es

# sent_id = PRESEEA-ALCA_H11_037-155
# time = 26:30
# location = España-Madrid-Alcalá de Henares
# turn_text = E:</> y<alargamiento/> ¿qué <vacilación/> y qué salís / casi siempre<alargamiento/>
# text = E:</> y<alargamiento/> ¿qué <vacilación/> y qué salís / casi siempre<alargamiento/>
# text_orth = E:</> y<alargamiento/> ¿qué <vacilación/> y qué salís / casi siempre<alargamiento/>
# text_en = E: </> and <lengthening/> what <hesitation/> and what do you come out / almost always <lengthening/>
# lang = es
```

- Continue exploring the pedagogical potential of treebanking for teaching and reflecting on linguistic diversity in Spanish language varieties
- Expert evaluation of gamification approaches in non-expert scenarios for diverse educational purposes.



Expert evaluation of "Juegos del Español" in the schools of the Region of Murcia (Spain).

Funded by the European Union

REFERENCES

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

Bonilla, J., Segundo, L. y Bouzouita, M. (2023). Using GWAPs for Verifying PoS Tagging of Spoken Dialectal Spanish. *10th International Conference on Behavioural and Social Computing* (BESC). Larnaca: IEEE. Doi: 10.1109/BESC59560.2023.10386542

Carcelén, A. & Uclés, G. (2019). Designing and developing a multidialectal oral corpus. The AMERESCO Corpus. *Normas*, 9, 17-36. doi: 10.7203/Normas.v9i1.16007

Fernández-Ordóñez, I. (2015a). COSER, Corpus Oral y Sonoro del Español Rural. Historia. [página web]. http://www.corpusrural.es/historia.php

Fernández-Ordóñez, I. (2015b). Dialectos del español peninsular. En J. Gutiérrez Rexach (Coord.), *Enciclopedia de Lingüística Hispánica*. Vol. 2 (pp. 387-404). Londres: Routledge Reino Unido. https://lc.cx/vnga9G

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

Martínez Sánchez, F., Prendes, M. P., Alfageme, M. B., Amorós, L., Rodríguez Cifuentes, T. y Solano, I. M. (2002). Herramienta de evaluación de multimedia didáctico. *Revista Pixel-Bit*, 18, 71-88. https://recyt.fecyt.es/index.php/pixel/article/view/61188

Ministry of Territorial Policy and Democratic Memory. (s.f.). Lenguas cooficiales en España. Ministerio de Política Territorial y Memoria Democrática. Retriever 23 of febrauery, 2024, https://lc.cx/9Gx8WM.

Moseley, Ch. (Ed.). (2010). *Atlas de las Lenguas del Mundo en Peligro*. (2ª ed.). París: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000189453

UD. Universal Dependencias Project. (2024). [web page]. Available in https://universaldependencies.org

Taulé, M., Borrega, O. y Martí, M. A. (2011). AnCora-Net: integración multilingüe de recursos lingüísticos semánticos. *Procesamiento del Lenguaje Natural*, 47, 153-160. http://hdl.handle.net/10045/18523