The First Spoken Spanish Treebank in Universal Dependencies: Annotation, Distributional Analysis, and Parsing Evaluation

El primer treebank de español oral en Universal Dependencies: anotación, análisis distribucional y evaluación de parsing

Johnatan E. Bonilla[1]
[1]Humboldt-Universität zu Berlin - Ghent University - Instituto Caro y Cuervo
j.bonilla@hu-berlin.de

Abstract: This paper presents the first treebank for spoken Spanish within the Universal Dependencies (UD) framework, built from transcriptions of the Corpus Oral y Sonoro del Español Rural (COSER). The treebank comprises sentences from 17 Spanish regions, manually annotated using Arborator-Grew. We provide a distributional analysis of dependency relations revealing differences between spontaneous speech and written text, including higher frequencies of discourse adverbs, disfluencies, and coordination, alongside reduced nominal complexity. Evaluation using UDPipe 2 as a pre-trained baseline yields a UAS of 75.51% and LAS of 66.48% on the orthographic version of the treebank. Fine-tuning a spaCy transformer via 10-fold cross-validation achieves a UAS of 77.94% and LAS of 67.52%, surpassing the pre-trained spaCy transformer baseline by +8.81 pp in LAS—the dominant gain attributable to learning speech-exclusive dependency labels—and surpassing the stronger UDPipe 2 baseline by +1.04 pp in LAS. Error analysis reveals that word order variability (obj↔nsubj confusion) and spoken-specific labels (reparandum, discourse) constitute the primary challenges for written-trained parsers.
Keywords: Treebank, spoken Spanish, Universal Dependencies, parsing.

Resumen: Este artículo presenta el primer treebank de español oral en el marco de Universal Dependencies (UD), construido a partir de transcripciones del Corpus Oral y Sonoro del Español Rural (COSER). El treebank comprende oraciones de 17 regiones españolas, anotadas manualmente mediante Arborator-Grew. Presentamos un análisis distribucional de las relaciones de dependencia que revela diferencias entre el habla espontánea y el texto escrito, incluyendo mayor frecuencia de adverbios discursivos, disfluencias y coordinación, junto con una menor complejidad nominal. La evaluación de UDPipe 2 como baseline muestra un UAS de 75,51% y un LAS de 66,48% sobre la versión ortográfica del treebank. El ajuste fino de un transformer spaCy con validación cruzada alcanza un UAS de 77,94% y un LAS de 67,52%, con una mejora de +8,81 pp en LAS respecto al baseline spaCy sin ajustar—atribuible principalmente al aprendizaje de relaciones exclusivas del habla—y +1,04 pp sobre el baseline UDPipe 2. El análisis de errores revela que la variabilidad en el orden de palabras (confusión obj↔nsubj) y las etiquetas específicas del habla oral (reparandum, discourse) constituyen los principales desafíos para los analizadores entrenados en texto escrito.
Palabras clave: Treebank, español oral, Dependencias Universales, parseado.

1 Introduction

One of the notable gaps in the current landscape of NLP resources is the underrepresentation of spoken language, especially in dialects that do not have a substantial written footprint. The majority of parsers and language models, such as those in the widely used spaCy library (Honnibal et al., 2020), are predominantly trained on data derived from written sources. This inherently lim-

its their effectiveness in processing spoken language, which often diverges significantly from its written counterpart in syntax, semantics, and pragmatics. Features such as flexible word order, colloquial expressions, and the presence of disfluencies pose unique challenges not typically encountered in written text.

In the context of Spanish, the Universal Dependencies (UD) project (Nivre et al., 2016; De Marneffe et al., 2021) has played a vital role in providing annotated treebanks for various languages. The Spanish AnCora treebank (Martínez-Alonso and Zeman, 2016; Taulé, Martí, and Recasens, 2008) stands out as a significant contribution within the UD project. However, despite its comprehensiveness, AnCora primarily focuses on written Spanish, leaving a notable void in resources tailored to spoken language. This gap is particularly relevant given that Spanish is the world's second most-spoken native language, with rich dialectal variation across Europe and the Americas (Vítores, 2023).

Recognizing this lacuna, the present work develops a dedicated treebank for spoken Spanish with a focus on rural dialects. The *Corpus Oral y Sonoro del Español Rural* (COSER) (Fernández-Ordóñez, 2005-to date) forms the backbone of this initiative. The COSER-UD treebank,[1] derived from this corpus, provides a framework for annotating the morphosyntactic and syntactic structures of spoken Spanish, adapting the UD guidelines to the peculiarities of spoken language.

The PoS tagging layer of the treebank has already been evaluated in previous work (Bonilla, 2025). The results showed a drop in accuracy from 98–99% on written data to 94–95% on spoken data. After fine-tuning with the gold standard, accuracy recovered to 98% for UPOS and 97% for FEATS. The present study complements this previous research by addressing the syntactic layer. We provide: (i) a distributional analysis of dependency relations comparing spoken and written Spanish, with a principled distinction between genuine differences and artifacts of sentence length; (ii) a detailed per-relation error analysis using UDPipe 2 (Straka, 2018) as a pretrained baseline, identifying the three main sources of parsing errors on spoken data; and (iii) a fine-tuning evaluation via 10-fold cross-validation, demonstrating that even a small spoken treebank yields substantial improvements, especially in labeling accuracy.

[1] https://github.com/UniversalDependencies/UD_Spanish-COSER

2 Background and Related Work

2.1 Spanish Treebanks in UD

The Cast3LB treebank, initiated by Civit and Martí (2004), was the first systematic effort to build a Spanish treebank as part of the 3LB project (Civit, Martí, and Bufí, 2006). Cast3LB's subsequent transformation into dependency format by Gelbukh, Torres Ramos, and Calvo Castro (2005) employed heuristic rules achieving a 99.9% match between manually marked and generated heads, although it did not conform to the UD framework. The AnCora project (Taulé, Martí, and Recasens, 2008) expanded on Cast3LB by adding richer linguistic annotations using journalistic corpora from sources including the EFE news agency and the newspaper El Periódico. AnCora adopted a surface-order, theory-neutral annotation approach.

The CoNLL-X Shared Task evaluated Spanish dependency parsers on AnCora, with top LAS scores of 82.3% from McDonald, Lerman, and Pereira (2006) and 81.3% from Nivre, Hall, and Nilsson (2006). The UD project further contributed the GSD treebank (McDonald et al., 2013) and the PUD treebank, aligned with UD v2 guidelines. Together, AnCora and GSD have been essential in improving parser accuracy for Spanish. Table 1 shows reported UAS and LAS across current NLP libraries, with UDPipe 2 achieving the highest accuracy (UAS 93.87, LAS 92.10).

2.2 Spoken Language Treebanks in UD

Treebanks for spoken language in UD address unique challenges such as disfluencies and non-standard grammar. The work of Dobrovoljc and Nivre (2016) on a spoken Slovenian treebank was foundational, introducing labels like *parataxis:restart* for disfluencies. Subsequent research by Braggaar and van der Goot (2021) handled spoken Dutch-Frisian code-switching, developing specialized annotation guidelines for non-standard sentence structures.

Kahane et al. (2021) introduced practical and theoretical guidelines for the devel-

NLP library	UAS	LAS	UD ver.
spaCy sm	90.38	86.85	2.8
spaCy md	91.26	88.00	2.8
spaCy lg	91.40	88.19	2.8
spaCy trf	94.85	93.10	2.8
Stanza NLP (Qi et al., 2020)	93.09	91.30	2.12
UDPipe 1.2 (Straka, Hajic, and Straková, 2016)	90.20	87.60	2.5
UDPipe 2 (Straka, 2018)	**93.87**	**92.10**	2.12

Table 1: UD_Spanish-AnCora UAS and LAS reported accuracy across NLP libraries. Abbreviations: sm = small, md = medium, lg = large, trf = transformer.

opment of spoken language treebanks in UD and SUD, addressing text-sound alignment, segmentation into sentences, use of punctuation in transcriptions, disfluencies, and paratactic constructions. Other significant contributions include learner English (Kyle et al., 2022) and the comparative overview by Dobrovoljc (2022), which highlighted transcription inconsistencies across treebanks such as Turkish-German (Çetinoğlu and Çöltekin, 2023), Naija (Caron et al., 2019), and French (Lacheret et al., 2014).

A key issue, as noted by Dobrovoljc (2022), is the lack of standardized transcription for spoken language, resulting in variations in orthography and the treatment of disfluencies, complicating morphosyntactic analysis across treebanks. Harmonization of these standards remains crucial for advancing cross-linguistic research on spoken syntax.

3 COSER Corpus and Data Preparation

The COSER-UD treebank is based on the COSER corpus (Fernández-Ordóñez, 2005-to date), which focuses on rural European Spanish dialects. As of December 2025, COSER includes data from 3,079 informants (average age: 74.3 years), primarily reflecting speech patterns that can be traced back to the early 20th century. The recordings, collected from 1,467 rural locations across Spain since 1990, span 2,002 hours. The corpus now has 272 transcriptions covering 371 hours and 58 minutes, totaling over 4 million words.

In COSER transcriptions, special attention is given to the representation of phonetic, phonological, and morphophonological variants. Segment suppression is transcribed faithfully, as in *dejao* for *dejado* ("left") or *pa* for *para* ("for"). Additions or changes in segment order use conventional orthography, e.g., *toballa* for *toalla* ("towel"). Stress changes are indicated with diacritical accents, e.g., *áhi* instead of standard *ahí* ("there"). Other conventions include writing out numerals (except years) and transcribing foreign words without translation. Personal names are anonymized with the marker *NP*.

To accommodate both dialectological and NLP research, the treebank maintains the original phonological transcription as the primary FORM field while providing standardized orthographic forms in the MISC column via the `Ortho=` feature. In the current dataset, 154 tokens (2.17%) carry such variants. For example, the entry for the phonological form *tá* includes `Ortho=está` in the MISC field, allowing automatic generation of an orthographic version of the treebank. This dual representation enables evaluation of parsers on both transcription variants (Section 6.3).

4 Annotation Process

The development of the COSER-UD treebank followed several stages. First, the COSER corpus was preprocessed to create a geographically balanced sample of conversation turns, excluding conversation markers for easier segmentation into sentences (Bonilla, Bouzouita, and Segundo-Díaz, 2022). Seventeen Spanish regions were represented.

SpaCy's sentence segmenter and the `spacy_conll` module (Honnibal et al., 2020) were used to segment conversation turns into sentences and generate CoNLL-U files with tokenization and metadata (including the original turn ID, location, and time range). Crucially, these CoNLL-U files were generated *without dependency annotation*—the HEAD and DEPREL columns were left empty. This is an important methodological point: no automatic parser was used to pre-annotate dependency structures, ensuring that the resulting gold standard is not

biased toward any particular parsing model.

The empty-dependency CoNLL-U files were imported into Arborator-Grew (Guibon et al., 2020), where all dependency structures were annotated manually from scratch. Arborator-Grew combines Arborator's collaborative graphical annotation interface with Grew's graph querying and rewriting capabilities, offering functionalities such as tree comparison, error detection via pattern matching, and controlled annotation access through role-based permissions.

The annotation protocol involved one primary annotator and two validators. Arborator-Grew's validation system allows validators to review all trees independently and select the correct version when disagreements arise. This consensus-based peer adjudication approach is standard in UD treebank development and has been employed in projects such as the Naija and Beja treebanks (Kahane et al., 2021). All annotation decisions were guided by the UD v2 guidelines, with adaptations for spoken language phenomena informed by Kahane et al. (2021) and the UD documentation for Spanish.

A gamification approach was previously introduced to expedite manual review of PoS tags (Bonilla, Bouzouita, and Segundo-Díaz, 2022; Bonilla, Díaz-Segundo, and Bouzouita, 2023), and a high-accuracy gold standard for PoS tagging—achieving 98% for UPOS and 97% for FEATS—was developed and evaluated in Bonilla (2025). The current treebank release comprises 474 sentences and 7,083 tokens (Table 2), with a mean sentence length of 14.9 tokens (median: 10.0, SD: 15.3).

Region	Sents	Tokens
Andalusia	30	383
Aragon	30	318
Asturias	31	518
Balearic Islands	30	503
Basque Country	20	384
Canary Islands	30	371
Cantabria	30	367
Castile-León	30	459
Castilla-La Mancha	30	406
Catalonia	30	325
Extremadura	31	445
Galicia	30	456
La Rioja	20	562
Madrid	30	309
Murcia	30	507
Navarre	22	321
Valencian Community	20	449
Total	**474**	**7,083**

Table 2: Distribution of sentences and tokens by region in the treebank.

5 Specificities of Treebank Annotations

5.1 Parts of Speech

The morphosyntactic tagging of spoken Spanish, aligned with UD guidelines, was refined by comparing tagging practices with the AnCora and GSD treebanks (Bonilla, 2025). Key revisions included reclassifying possessive adjectives as determiners (DET), as in *la casa mía* ("my house"), and tagging ordinal numbers as adverbs (ADV) in verbal contexts. Features specific to spoken language were added, including the `Degree=Dim` tag for diminutives (*-ito*, *-ín*) and the `Foreign=Yes` tag for non-Spanish words, capturing regional and colloquial variation. Pronouns were further refined to distinguish between Interrogative and Exclamative forms, while spoken verb conjugations such as *vosotros* forms and subjunctive uses were addressed to reflect regional nuances.

5.2 Dependency Relation Distribution

Table 3 compares the distribution of dependency relations in the spoken treebank with the written AnCora treebank (train subset; 453,039 tokens). Relations are sorted by frequency in the spoken data and only those with notable differences or theoretical interest are shown.

A critical methodological observation is that some distributional differences are artifacts of sentence length rather than genuine linguistic properties of spoken language. The spoken corpus has a mean sentence length of 14.9 tokens, versus approximately 27 in AnCora. Since each sentence has exactly one root, the proportion of *root* is the inverse of the mean sentence length: $1/14.9 = 6.71\%$, which matches the observed 6.69% almost exactly. Similarly, the higher proportion of *punct* (20.57% vs. 11.69%) reflects the proportionally greater weight of sentence-boundary punctuation in shorter utterances.

Beyond these artifacts, three categories of genuine differences emerge. First, relations exclusive to spoken language: *reparandum* (2.09%) marks self-corrections and hesitations, *discourse:filler* (0.30%) captures filled pauses such as *eh*, and *discourse* (1.99%) an-

notates discourse markers and interjections. These relations are absent from AnCora.

Second, several relations are substantially more frequent in speech. The increase in *advmod* (+4.82 pp) reflects the pervasive use of adverbs as discourse connectors (*pues*, *ya*, *luego*, *entonces*). The higher frequency of *cc* (+2.51 pp) and *advcl* (+2.04 pp) indicates a preference in spoken language for coordination and adverbial clauses over the complex subordination typical of written text.

Third, relations associated with nominal complexity are markedly reduced in speech: *case* (−7.81 pp), *det* (−6.05 pp), *nmod* (−5.51 pp), *amod* (−4.33 pp), and *flat* (−2.47 pp). This pattern is consistent with the well-documented tendency in spontaneous colloquial Spanish to favor pronominal reference over full, complex noun phrases (Vigara Tauste, 1994).

5.3 Indirect Objects and Obliques

In Spanish traditional grammar, dative pronouns like *le* are classified as indirect objects, associated with recipients or beneficiaries. However, in the UD framework for Spanish, these constructs are treated as oblique dependents and labeled *obl:arg* to differentiate them from temporal or locational adjuncts. The treebank consistently applies this UD convention: dative clitics are annotated as *obl:arg* (82 instances), with zero instances of *iobj*, following the analysis established for Spanish in UD.

5.4 Discourse Elements

The *discourse* label is applied to interjections, discourse particles, and non-adverbial markers. In the treebank, the most frequent discourse markers by lemma are: *pues* (42 occurrences), *sí* (32), *bueno* (10), and *claro* (8). These elements are essential in conveying speaker intentions, guiding conversation flow, and maintaining coherence. Figure 1 illustrates the annotation for the interjection *Claro* in a sentence from the Catalonia subset.

Following Dobrovoljc and Nivre (2016), the subtype *discourse:filler* annotates sounds that lack syntactic or semantic function, corresponding to a pause to think rather than an actual word. In the treebank, this phenomenon commonly occurs with the particle *eh* (21 instances).

5.5 Conversational Fragments and Ellipsis

Ellipsis in Spanish is a phenomenon well documented in analyses of colloquial language (Vigara Tauste, 1994). Spoken data contains numerous utterances without a finite verb, such as *¿Y la leche de las ovejas?* ("And the milk from the sheep?") or *¿Y de qué color?* ("And what color?"). The treebank contains 139 sentences with a non-verbal root, representing 29.3% of all sentences. A quantitative analysis of these cases reveals three distinct categories:

Copular constructions (40 cases): A content word is promoted to root in the presence of a *cop* dependent, following standard UD promotion rules. Example: *Y bueno, esta pregunta es un poco impertinente* ("Well, this question is a bit impertinent"), where the adjective *impertinente* serves as root with *es* as copula.

Conversational fragments (87 cases): Pragmatically complete turns that do not contain a recoverable elided verb. These include follow-up questions (e.g., *¿Y abuela?*, "And grandma?"), backchannel responses (*Sí, hombre*, "Yes, man"), and topic-establishing phrases (*Con la mano*, "By hand"). For these, the *orphan* relation is not employed, following the principle that they are not instances of grammatical ellipsis in the transformational sense, but rather autonomous pragmatic units characteristic of dialogue (Kahane et al., 2021). The most salient content word is promoted to root.

Potential verbal ellipsis (12 cases): Utterances with verbal dependents (e.g., *obl*, *nsubj*) but no copula. These are the only cases where the *orphan* relation could be considered; however, given their low frequency and the pragmatic completeness of the utterances in context, we adopt the same promotion strategy. Figure 2 illustrates a typical conversational fragment.

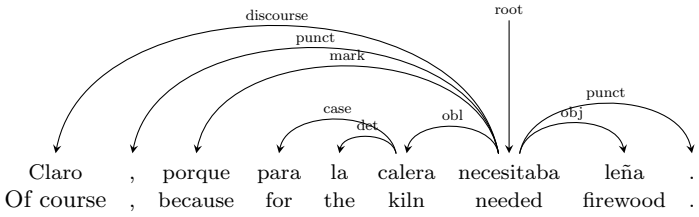
5.6 Disfluencies

The annotation of disfluencies is integral to capturing the essence of spoken language. The relation *reparandum* denotes words or phrases that are replaced or corrected by the speaker during the flow of speech. The treebank contains 148 instances of *reparandum* across 101 sentences. These predominantly involve function words (pronouns: 32, X/unintelligible: 20, determiners: 19,

Relation	Spoken %	Written %	Diff
punct*	20.57	11.69	+8.88
det	9.06	15.11	−6.05
advmod	8.27	3.45	+4.82
root*	6.69	3.15	+3.54
obj	6.58	4.61	+1.97
case	5.87	13.68	−7.81
cc	5.24	2.73	+2.51
mark	4.38	3.47	+0.91
nsubj	4.33	5.29	−0.96
conj	4.15	2.91	+1.24
advcl	3.66	1.62	+2.04
obl	3.64	4.22	−0.58
reparandum	2.09	—	spoken
discourse	1.99	0.00	+1.99
nmod	1.37	6.88	−5.51
amod	0.92	5.25	−4.33
flat	0.11	2.58	−2.47
discourse:filler	0.30	—	spoken

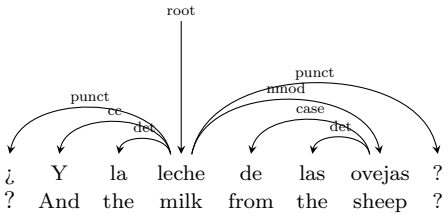
*Artifact of sentence length (see text). "spoken" = exclusive to spoken data. "—" = not attested in corpus.
N-spoken = 7,083; N-written = 453,039.

Table 3: Distribution of selected dependency relations in the spoken treebank vs. AnCora (written), sorted by spoken frequency.



(*Of course, because she needed firewood for the kiln.*)

Figure 1: Annotation of the interjection *Claro* with the *discourse* relation (sent_id: cata-488).

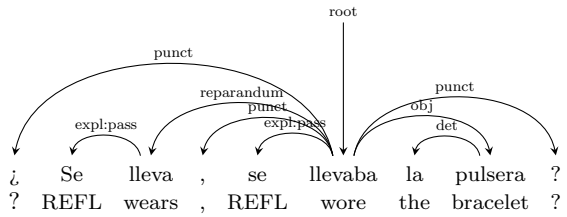


(*And the milk from the sheep?*)

Figure 2: Conversational fragment with a nominal root (sent_id: bale-77).

prepositions: 16) and verbs (15), reflecting that speakers most commonly self-correct on grammatical elements and tense/mood choices.

Figure 3 shows a tense self-correction from the Extremadura subset, where the speaker begins with the present tense *lleva* before correcting to the imperfect *llevaba*. The repaired element attaches to the correcting element via *reparandum*, preserving the relationship between the two attempts.



(*Did they wear—did they used to wear the bracelet?*)

Figure 3: *Reparandum* annotating a tense self-correction (sent_id: extr-1121).

The relation *parataxis* (29 instances) is

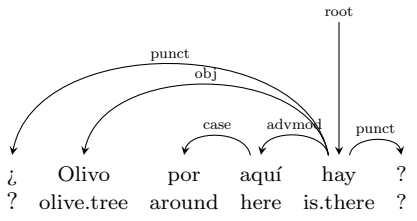
employed to describe the juxtaposition of clauses that could stand alone but are presented together without a connecting conjunction. This relation is particularly relevant in spoken data where clauses are linked by punctuation or intonation rather than explicit syntactic connectors.

5.7 Word Order Variability

In Spanish, the canonical SVO order is most commonly observed in formal written language (Villalba, 2012). However, the flexibility of Spanish syntax allows for variations such as SOV, VSO, and OVS in spoken language, particularly when a speaker wishes to topicalize a specific element or conform to a conversational style reflecting immediacy (Morales, 2006).

For parsing algorithms, this flexibility presents a considerable challenge. As our error analysis confirms (Section 6.2), the confusion between *nsubj* and *obj* annotations—especially when the subject follows the verb—is the single most frequent parsing error, with 44 *obj* tokens misclassified as *nsubj* and 30 *nsubj* tokens misclassified as *obj*. This confusion arises from the parsers' tendency to adhere to the expected SVO order, leading to systematic misclassification when speakers deviate from it.

Figure 4 illustrates a representative OVS construction from the treebank: *¿Olivo por aquí hay?*, where the fronted nominal *Olivo* occupies the pre-verbal position typically associated with subjects in written Spanish, yet is correctly annotated as *obj* of the existential verb *hay*, which takes no overt subject.



(*Are there any olive trees around here?*)

Figure 4: OVS construction: fronted object *Olivo* precedes the existential verb *hay* with no overt subject. A parser biased towards SVO order tends to misclassify *Olivo* as *nsubj* (sent_id: anda-230).

6 Parsing Evaluation

6.1 Pre-trained Model Evaluation

This section compares the performance of models trained on the AnCora treebank when applied to the spoken treebank. A critical methodological point is that the treebank was annotated entirely from scratch in Arborator-Grew, without any automatic pre-annotation of dependency structures. This eliminates the possibility of annotation bias toward any particular parser—a concern raised in previous evaluations.

Table 4 reports UAS and LAS on the spoken treebank compared to each model's reported accuracy on AnCora. All models exhibit a substantial performance drop on spoken data. UDPipe 2, the best performer on AnCora (UAS 93.87, LAS 92.10), drops to UAS 75.52 and LAS 66.36—a decrease of 18.35 and 25.74 points, respectively. SpaCy's non-transformer models (sm, md, lg) performed worst, with UAS ranging from 69 to 71 and LAS from 54 to 56. The spaCy transformer model fared better (UAS 74.15, LAS 58.44), while Stanza NLP and UDPipe 2 achieved the closest results to written-data performance.

The difference between the phonological and orthographic versions of the treebank is consistently small across all models, ranging from 0.04 to 1.32 pp in UAS and 0.04 to 1.26 pp in LAS, consistent with the observation that only 2.17% of tokens carry phonological variants.

6.2 Error Analysis

To understand *where* parsers fail on spoken data, we conducted a per-relation error analysis using UDPipe 2 predictions—the best-performing pre-trained baseline. Table 5 presents accuracy for the relations with the most errors, and Table 6 shows the top confusion pairs.

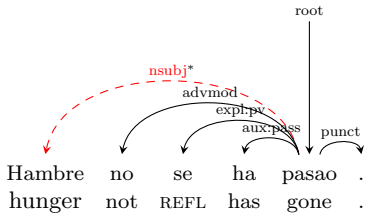
The errors fall into three categories. First, *spoken-exclusive relations* that the parser never encountered during training account for 18.0% of all LAS errors (429 out of 2,383): *reparandum* (148 errors, LAS 0%), *discourse* (141 errors, LAS 0%), *aux:pass* (42 errors, LAS 0%), *parataxis* (29 errors, LAS 0%), *acl:relcl* (27 errors, LAS 0%), and *discourse:filler* (21 errors, LAS 0%). While the parser can sometimes attach these tokens to the correct head (e.g., UAS for *discourse*:

Pre-trained Model	AnCora		Spoken (Phono)		Spoken (Ortho)	
	UAS	LAS	UAS	LAS	UAS	LAS
spaCy sm	88.77	82.44	69.34	54.03	69.85	54.63
spaCy md	89.73	83.54	70.00	55.38	70.94	56.35
spaCy lg	89.78	83.74	70.37	55.48	71.10	56.21
spaCy trf	93.85	88.69	74.15	58.44	74.87	58.71
UDPipe 1.2	88.13	80.83	72.25	57.05	73.57	58.31
UDPipe 2	**93.87**	**92.10**	**75.52**	**66.36**	**75.51**	**66.48**
Stanza NLP	93.25	89.94	75.21	65.46	75.80	66.20

Table 4: UAS and LAS for pre-trained models on AnCora vs. the spoken treebank. UDPipe 2 results obtained via its REST API with gold tokenization and PoS tags.

61%), it *never* assigns the correct label, instead defaulting to labels from its training vocabulary such as *mark*, *advcl*, or *dep*.

Second, *structurally ambiguous relations* where spoken word order confuses the parser: *obj* has 147 LAS errors, with 44 tokens misclassified as *nsubj* and 41 as *obl:arg*; *nsubj* has 84 LAS errors, with 30 tokens misclassified as *obj*. This bidirectional obj↔nsubj confusion (74 errors total) is the single largest error cluster and directly reflects the parser's expectation of SVO order.



(*Hunger has not been suffered.*)
Gold: `obj` UDPipe 2: `nsubj`*

Figure 5: UDPipe 2 error on a fronted object: *Hambre* is correctly annotated *obj* (gold), but predicted *nsubj* (dashed, marked with *), reflecting the parser's SVO prior (sent_id: rioj-500).

Figure 5 illustrates the most frequent error type (obj→nsubj, 44 cases): in *Hambre no se ha pasao* ("Hunger has not been suffered"), the fronted noun *Hambre* is the object of an impersonal passive construction, but UDPipe 2 assigns it *nsubj*—filling the canonical pre-verbal subject slot. The reverse confusion (nsubj→obj, 30 cases) also occurs: in *subastan, los palos* ("They auction off, the poles"), the post-verbal subject *palos* is predicted as *obj*, as the parser defaults to treating post-verbal nominals as objects in the absence of a pre-verbal subject.

Third, *clausal relations*: *conj* (LAS 47.96%, frequently confused with *advcl*) and *advcl* (LAS 46.72%, confused with *root* and *acl*), reflecting the difficulty of identifying clause boundaries in spoken language without explicit subordinating cues.

Relation	N	UAS%	LAS%	Err
punct	1457	58.75	58.75	601
advmod	586	80.89	77.65	131
obj	466	92.06	68.45	147
conj	294	57.82	47.96	153
advcl	259	55.98	46.72	138
reparandum	148	8.78	0.00	148
discourse	141	60.99	0.00	141
nsubj	307	86.64	72.64	84
obl	258	72.09	55.04	116
discourse:filler	21	38.10	0.00	21

Table 5: Per-relation UDPipe 2 accuracy on relations with highest error counts. Err = total LAS errors.

Gold	Predicted	N
obj	nsubj	44
aux:pass	aux	42
obj	obl:arg	41
conj	advcl	36
discourse	mark	34
discourse	advcl	32
nsubj	obj	30
root	advcl	25
acl:relcl	acl	25

Table 6: Top 9 confusion pairs from UDPipe 2 evaluation (gold → predicted).

6.3 Fine-tuning with Cross-Validation

Given the small treebank size, 10-fold cross-validation was used to evaluate fine-tuning performance (Kohavi, 1995). This method divides the data into ten segments, using nine for training and one for testing in each itera-

tion, maximizing data use while providing a robust estimate of variance.

Although UDPipe 2 constitutes the strongest pre-trained baseline, it relies on a deprecated software stack (Python 3.7 / TensorFlow 1.15) that is incompatible with current transformer libraries (`transformers`, `protobuf`), making reproducible custom fine-tuning infeasible with modern tooling. SpaCy, by contrast, provides a production-ready and actively maintained ecosystem with native transformer support, and was therefore selected as the fine-tuning framework.

The model was trained using spaCy's neural network framework with the pre-trained BETO model (`dccuchile/bert-base-spanish-wwm-cased`) (Cañete et al., 2020). The architecture used spaCy's TransitionBasedParser.v2 with a hidden layer width of 128, 3 maxout pieces, and `use_upper=false`. The transformer component was configured with a window size of 128 and stride of 96, without mixed precision. Training used gradient accumulation over 3 steps, a warmup linear learning rate (initial: 5e-5, 250 warmup steps, 20,000 total steps), the `batch_by_padded` strategy with batch sizes of 2,000 and a buffer of 256, early stopping with patience of 300 evaluations at a frequency of 50 steps, and Adam optimization with L2 weight decay of 0.01 and gradient clipping at 1.0.

Table 7 presents the results alongside both pre-trained baselines evaluated on both treebank versions. UDPipe 2 constitutes the strongest baseline, since it is trained on the most recent AnCora release (UD 2.12) with up-to-date annotation guidelines, whereas the spaCy models were trained on the older UD 2.8 version. This difference in training data recency is reflected in the large gap between UDPipe 2's LAS (66.36/66.48) and spaCy trf's LAS (58.44/58.71) on the spoken data, despite comparable performance on AnCora itself.

The fine-tuned model surpasses UDPipe 2—the best pre-trained baseline—by +2.43 pp in UAS and +1.04 pp in LAS on the orthographic version. While this improvement is modest, it is notable because UDPipe 2 benefits from both a more recent training set and a strong underlying architecture (Straka, 2018), yet it lacks exposure to spoken-specific labels. The fine-tuned model overcomes this limitation by learning relations such as *reparandum*, *discourse*, and *discourse:filler* from the spoken training data.

The comparison against the spaCy trf baseline—i.e., the same architecture without fine-tuning—reveals the full effect of domain adaptation. UAS improves by 3–4 pp, while LAS improves by nearly 9 pp. This asymmetry is explained directly by the error analysis: the pre-trained spaCy model can often identify correct head attachment (explaining its reasonable UAS) but cannot assign spoken-specific labels that are absent from its AnCora 2.8 training data. Fine-tuning teaches the model these labels, accounting for the disproportionate LAS gain. The fact that the fine-tuned spaCy trf model, originally 8 pp behind UDPipe 2 in LAS, catches up and surpasses it after fine-tuning demonstrates that domain-specific data can compensate for both architectural and training data disadvantages.

The difference between phonological and orthographic versions remains minimal in the fine-tuned setting (UAS: 0.11 pp, LAS: 0.32 pp), confirming that the primary challenge for parsing spoken Spanish lies in syntactic structure and spoken-specific phenomena rather than surface-level transcription differences.

7 Conclusions and Future Work

This paper has presented the first spoken Spanish treebank within the Universal Dependencies framework, covering 17 Spanish regions, annotated entirely de novo using Arborator-Grew. The main contributions are threefold.

First, the distributional analysis reveals that differences between spoken and written dependency distributions fall into three principled categories: artifacts of sentence length (root and punct frequencies), genuine spoken-language features (discourse markers, disfluencies, coordination preference), and reduced nominal complexity (lower frequencies of det, case, nmod, amod). This three-way distinction provides a clearer framework for interpreting distributional differences than previous analyses.

Second, the per-relation error analysis identifies three key challenges for parsers trained on written data: spoken-exclusive relations that account for 18% of all LAS errors, word order variability causing sys-

Model	Phono		Ortho	
	UAS	**LAS**	**UAS**	**LAS**
Pre-trained baselines (no fine-tuning):				
UDPipe 2	75.52	66.36	75.51	66.48
spaCy trf	74.15	58.44	74.87	58.71
Fine-tuned spaCy trf (10-fold CV):				
	$77.83_{\pm 1.99}$	$67.20_{\pm 2.22}$	$77.94_{\pm 1.91}$	$67.52_{\pm 1.74}$
Improvement over UDPipe 2:				
Δ	+2.31	+0.84	+2.43	+1.04
Improvement over spaCy trf:				
Δ	+3.68	+8.76	+3.07	+8.81

Table 7: Fine-tuning results (10-fold CV, mean ± SD) compared to both pre-trained baselines on both treebank versions. Δ rows show the improvement of the fine-tuned model over each baseline.

tematic subject-object confusion, and clause boundary identification in paratactic speech. These findings provide actionable insights for improving spoken language parsing.

Third, fine-tuning with the spoken treebank yields a substantial LAS improvement of nearly 9 pp over the non-fine-tuned spaCy baseline, primarily attributable to learning spoken-specific dependency labels. This demonstrates that even small amounts of domain-specific annotated data can meaningfully improve parser performance on underrepresented varieties.

The treebank is an ongoing project with plans for expansion. The gold standard for PoS tagging already comprises over 200,000 tokens (Bonilla, 2025), and manual parsing annotation continues incrementally. Future work will focus on increasing the treebank size, potentially incorporating additional varieties of Spanish, and investigating cross-lingual transfer from other spoken UD treebanks. The treebank is publicly available as part of the Universal Dependencies project.

Acknowledgements

This work was carried out as part of a medium-scale research infrastructure project funded by the Flemish Research Fund (Fonds voor Wetenschappelijk Onderzoek, FWO) entitled "A Respeaking and Collaborative Game-Based Approach to Building a Parsed Corpus of European Spanish Dialects" (project reference: I000418N). The author wishes to thank Daniel Zeman, Véronique Hoste, and Miriam Bouzouita for their valuable contributions to this work.

References

Bonilla, J. E. 2025. Spoken spanish pos tagging: gold standard dataset. *Language Resources and Evaluation*, 59(2):983–1012.

Bonilla, J. E., M. Bouzouita, and R. L. Segundo-Díaz. 2022. La construcción del corpus oral y sonoro del español rural-anotado y parseado (coser-ap): avances en el etiquetado de partes del discurso. *Revista Internacional de Lingüística Iberoamericana*.

Bonilla, J. E., R. L. Díaz-Segundo, and M. Bouzouita. 2023. Using gwaps for verifying pos tagging of spoken dialectal spanish. In *2023 10th International Conference on Behavioural and Social Computing (BESC)*, pages 1–7. IEEE.

Braggaar, A. and R. van der Goot. 2021. Creating a universal dependencies treebank of spoken frisian-dutch code-switched data. *arXiv preprint arXiv:2102.11152*.

Caron, B., M. Courtin, K. Gerdes, and S. Kahane. 2019. A surface-syntactic ud treebank for naija. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 13–24. Association for Computational Linguistics.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *Practical Machine Learning for Developing Countries at the Tenth International Conference on Learning Representations*, pages 1–10.

Çetinoğlu, Ö. and Ç. Çöltekin. 2023. Two languages, one treebank: building a turkish–german code-switching treebank and its challenges. *Language Resources and Evaluation*, 57(2):545–579.

Civit, M. and M. A. Martí. 2004. Building cast3lb: A spanish treebank. *Research on Language and Computation*, 2(4):549–574.

Civit, M., M. A. Martí, and N. Bufí. 2006. Cat3lb and cast3lb: From constituents to dependencies. In T. Salakoski, F. Ginter, S. Pyysalo, and T. Pahikkala, editors, *Advances in Natural Language Processing*, pages 141–152, Berlin, Heidelberg. Springer Berlin Heidelberg.

De Marneffe, M.-C., C. D. Manning, J. Nivre, and D. Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Dobrovoljc, K. 2022. Spoken language treebanks in universal dependencies: an overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806.

Dobrovoljc, K. and J. Nivre. 2016. The universal dependencies treebank of spoken slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1566–1573.

Fernández-Ordóñez, I. 2005-to date. Corpus oral y sonoro del español rural (audible corpus of spoken rural spanish). Accessed: 2022-04-15.

Gelbukh, A., S. Torres Ramos, and F. H. Calvo Castro. 2005. Transforming a constituency treebank into a dependency treebank. *Procesamiento del Lenguaje Natural*, 35:145–152.

Guibon, G., M. Courtin, K. Gerdes, and B. Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *LREC 2020-12th Language Resources and Evaluation Conference*.

Honnibal, M., I. Montani, S. Van Landeghem, and A. Boyd. 2020. spacy: Industrial-strength natural language processing in python. `https://github.com/explosion/spaCy`.

Kahane, S., B. Caron, E. Strickland, and K. Gerdes. 2021. Annotation guidelines of ud and sud treebanks for spoken corpora. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages pp–35. Association for Computational Linguistics.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 2:1137–1145.

Kyle, K., M. Eguchi, A. Miller, and T. 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.

Lacheret, A., S. Kahane, J. Beliao, A. Dister, K. Gerdes, J.-P. Goldman, N. Obin, P. Pietrandrea, and A. Tchobanov. 2014. Rhapsodie: a prosodic-syntactic treebank for spoken french. In *Language Resources and Evaluation Conference*.

Martínez-Alonso, H. and D. Zeman. 2016. Universal dependencies for the ancora treebanks. *Procesamiento del Lenguaje Natural*, 57:91–98.

McDonald, R., K. Lerman, and F. Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 216–220.

McDonald, R., J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, O. Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Morales, A. 2006. Los sujetos 'ligeros' del español y su posición en la oración. *Haciendo lingüística. Homenaje a Paola Bentivoglio*, pages 487–501.

Nivre, J., M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, and others. 2016. Universal dependencies v1: A multilingual treebank

collection. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*.

Nivre, J., J. Hall, and J. Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *LREC*, volume 6, pages 2216–2219.

Qi, P., Y. Zhang, Y. Zhang, J. Bolton, and C. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. `https://arxiv.org/abs/2003.07082`.

Straka, M. 2018. Udpipe 2.0 prototype at conll 2018 ud shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207.

Straka, M., J. Hajic, and J. Straková. 2016. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In N. Calzolari et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož. European Language Resources Association (ELRA).

Taulé, M., M. A. Martí, and M. Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In N. Calzolari et al., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 96–101, Marrakech. European Language Resources Association (ELRA).

Vigara Tauste, A. M. 1994. Economía y elipsis en el registro coloquial español. *Tabanque: Revista pedagógica*, (9):9–20.

Villalba, X. 2012. *El orden de las palabras en español*. Castalia.

Vítores, D. F. 2023. El español: una lengua viva. informe 2023. In C. P. Villalba, editor, *El español en el mundo 2023. Anuario del Instituto Cervantes*. Instituto Cervantes, Alcalá de Henares, Madrid, pages 23–142.