Parsing the Switch: LLM-Based UD Annotation for Complex Code-Switched and Low-Resource Languages

Olga Kellert[1*] **Nemika Tyagi**[1*] **Muhammad Imran**[2] **Nelvin Licona-Guevara**[1]
Carlos Gómez-Rodríguez[2]
[1]Arizona State University [2]Universidade da Coruña, CITIC
{olga.kellert, ntyagi8, nliconag}@asu.edu, {m.imran, carlos.gomez}@udc.es

Abstract

Code-switching presents a complex challenge for syntactic analysis, especially in low-resource language settings where annotated data is scarce. While recent work has explored the use of large language models (LLMs) for sequence-level tagging, few approaches systematically investigate how well these models capture syntactic structure in code-switched contexts. Moreover, existing parsers trained on monolingual treebanks often fail to generalize to multilingual and mixed-language input. To address this gap, we introduce the *BiLingua Pipeline*, an LLM-based annotation pipeline designed to produce Universal Dependencies (UD) annotations for code-switched text. First, we develop a prompt-based framework for Spanish-English and Spanish-Guaraní data, combining few-shot LLM prompting with expert review. Second, we release two annotated datasets, including the first Spanish-Guaraní UD-parsed corpus. Third, we conduct a detailed syntactic analysis of switch points across language pairs and communicative contexts. Experimental results show that BiLingua Pipeline achieves up to **95.29%** LAS after expert revision, significantly outperforming prior baselines and multilingual parsers. These results show that LLMs, when carefully guided, can serve as practical tools for bootstrapping syntactic resources in under-resourced, code-switched environments[1].

1 Introduction

Code-switching (CSW) is a widespread linguistic phenomenon observed in multilingual communities around the world. Despite its prevalence in spoken and informal digital communication, it remains a complex challenge for natural language processing (NLP), particularly for syntactic parsing. One of the central issues is that most state-of-the-art parsing models are trained on monolingual treebanks and thus lack robustness when applied to mixed-language data (Özateş et al., 2022).

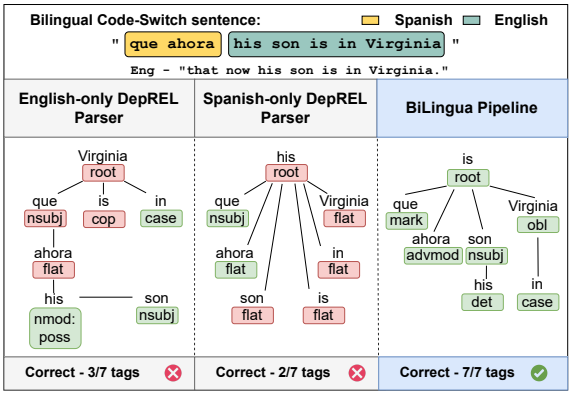


Figure 1: Comparison of dependency relation predictions (DepREL) for a Spanish-English CSW sentence across three parsing tools. The English-only and Spanish-only models misassign key relations due to monolingual bias. In contrast, the *Bilingua Pipeline* correctly analyzes the full structure across the language boundary.

Previous works of Özateş et al. (2022); Rijhwani et al. (2017); Bhat et al. (2018) took an important step toward addressing this gap by proposing, for instance, a semi-supervised dependency parsing framework that augments training with auxiliary sequence labeling tasks (Özateş et al., 2022). Their model improved parsing accuracy on Turkish-German spoken corpus by learning better representations of syntactic structure in a multilingual setting. However, even with such enhancements, existing models often rely on large amounts of annotated data, which is particularly limiting for under-resourced language pairs.

Motivated by this lack of resources, we introduce Bilingua Pipeline, a bilingual syntactic parser pipeline based on large language models (LLMs), specifically the GPT-4.1 model, to generate syntactically annotated CSW datasets. Figure 1 illustrates

[*]Equal contribution.
[1]Data and source code are available at https://github.com/N3mika/ParsingProject.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15934–15949
November 4-9, 2025

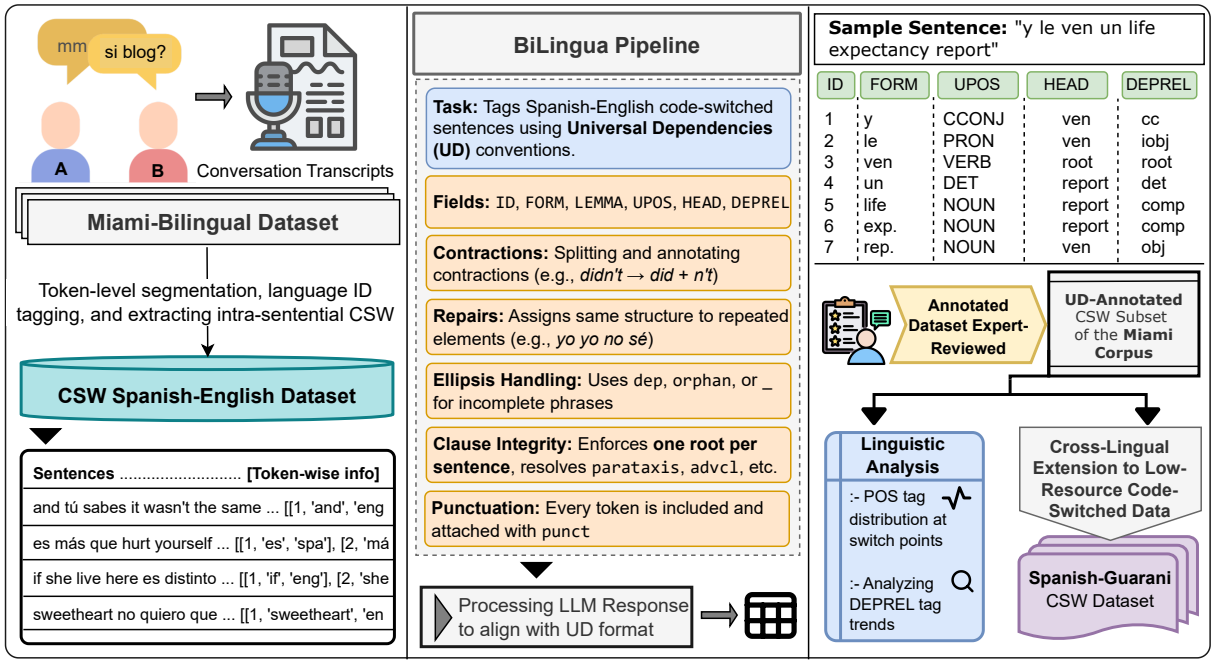


Figure 2: Overview of *Bilingua Pipeline* for Spanish-English code-switching. **Left:** Conversation transcripts from the Miami-Bilingual Corpus are processed through token-level segmentation, language ID tagging, and filtering of intra-sentential code-switches. **Center:** The Pipeline assigns UD tags to CSW sentences, handling contractions, repetitions, ellipsis, and clausal structure. **Right:** The resulting annotated dataset is reviewed by linguistic experts and enables downstream tasks such as POS/DEPREL analysis and extension to low-resource settings, including Spanish-Guaraní.

how current monolingual parsers perform significantly worse than Bilingua Pipeline on a widely studied CSW language pair. Next, we tackle this issue for two language pairs in particular: Spanish-English (a relatively well-resourced CSW language pair) and Spanish-Guarani (a low-resource language pair for which most linguistic tools are largely unavailable) (Chiruzzo et al., 2023). To our knowledge, these are the first datasets for Spanish-English and Spanish–Guaraní code-switching with UD-based syntactic annotations reviewed by native speakers. The entire workflow for creating and using Bilingua Pipeline is shown in Figure 2.

In addition to developing Bilingua Pipeline, we also examine the limitations of current syntactic parsing evaluation metrics. To this end, we introduce additional methods for assessing the performance of our parser with the help of linguistic experts. Moreover, the annotated datasets we create also support a wide range of linguistic analyses involving bilingual speakers, including understanding of fine-grained switch-point behavior. While most previous structural studies in NLP on CSW have focused on part-of-speech (POS) tags (Martínez, 2020; Rijhwani and Solorio, 2016; Solorio and Liu, 2008), our analysis using dependency parsing shows that syntactic subjects (nsubj) are among the most frequent switch points in both language pairs and that Spanish-Guarani CSW exhibits higher variation in switch points. This finding underscores the value of dependency parsing for analyzing switch points across languages. The main contributions of our work are as follows:

- We introduce a method for generating UD-style syntactic annotations using LLM-based prompting, and compare it against baselines from previous work on dependency parsing of CSW texts, as well as a parser trained on a synthetic combination of monolingual treebanks.
- We release two CSW datasets, with POS and dependency annotations, reviewed by native speakers, including a new resource for Spanish-Guaraní.
- We conduct a linguistic case study of common syntactic structures at CSW boundaries, revealing cross-linguistic switching patterns.

Overall, we believe our findings and released resources will support future work in both compu-

Study	Language Pair	POS Accuracy	LAS (Parsing)
Solorio et al. (2014)	Spanish–English	85.1%	–
Solorio et al. (2014)	Hindi–English	83.3%	–
Rijhwani et al. (2017)	Spanish–English	80.0%	–
Özateş et al. (2022)	Hindi–English	–	71.93%
Özateş et al. (2022)	Turkish–German	–	66.30%
Bhat et al. (2018)	Hindi–English	–	71.03%

Table 1: POS tagging and dependency parsing performance on CSW datasets in prior work.[2]

Datasets	Statistics	Spanish-English	Spanish-Guaraní
Original	Sentences	≈ 56,000	1,140
	Tokens	242,475	≈ 17,100
Code-Switched	Sentences	2,837	1,140
	Tokens	30,811	≈ 17,100

Table 2: Sentence and token counts in the original and code-switched subsets of the Spanish-English and Spanish-Guaraní datasets.

tational modeling and linguistic analysis of code-switching, especially for under-resourced and typologically diverse language pairs.

2 Related Work

Parsing CSW text is considerably more challenging than monolingual parsing due to structural variability, mixed grammar rules, and limited annotated corpora. Prior work (Rijhwani et al., 2017; Solorio and Liu, 2008; Solorio et al., 2014; Özateş et al., 2022) has demonstrated that models trained exclusively on monolingual data perform poorly on CSW without specific adaptation. Özateş et al. (2022) addressed this gap by proposing a semi-supervised parsing framework that incorporates auxiliary sequence labeling tasks. Their approach improved parsing performance on Turkish-German CSW with labeled attachment scores (LAS (Buchholz and Marsi, 2006)) reaching up to 73%. Similarly, Bhat et al. (2018) and Rijhwani et al. (2017) reported LAS scores of about 70-72% for Hindi-English data using adapted models (Table 1).

Most prior research focuses on part-of-speech tagging rather than full syntactic parsing, and available resources remain limited to a few language pairs or L2 languages (Sung and Shin, 2025). Aguilar et al. (2020) introduced LinCE, a centralized benchmark for linguistic code-switching covering tasks such as POS-tagging for various language pairs, including Spanish-English. A draft UD treebank exists for Spanish-English code-switching, but it is not publicly released, further illustrating the scarcity of syntactically annotated CSW data. Efforts to increase parsing speed, such as recasting dependency parsing as a sequence labeling task (Strzyz et al., 2019; Roca et al., 2023), have improved runtime, but still depend heavily on monolingual training data.

Our work builds on these foundations but shifts toward LLM-based annotation. LLMs like OpenAI GPT can be prompted with examples and linguistic rules to produce annotations without requiring extensive supervised training. We apply this technique to generate and evaluate new CSW datasets, including one for Spanish-Guaraní, thus expanding the reach of syntactic tools to underserved language communities.

[2]LAS comparisons across prior works involve different language pairs and data sources, and such comparisons are meant to be indicative, not directly comparable.

3 Dataset and Experiments

3.1 Datasets

We use two existing datasets for implementing our structure for Bilingua Pipeline and creating linguistically annotated CSW datasets. The first dataset we use is the Miami Corpus (Deuchar et al., 2014), a well-known Spanish-English dataset widely used in bilingualism and NLP research (Fricke and Kootstra, 2016; Chi and Bell, 2024; Martínez, 2020). The second is the Spanish-Guaraní dataset from the shared task GUA-SPA: Guarani-Spanish Code-Switching Analysis (Chiruzzo et al., 2023), including social media and news content with spontaneous multilingual usage in a low-resource setting.

Miami Spanish-English Corpus. This dataset comprises transcribed spoken interactions between bilingual speakers. Each sentence is tokenized and annotated with a language tag, POS tag, and morphological features (e.g., `be.V.3S.PRES`). Metadata includes token index, sentence and utterance IDs, speaker identity, and filename. An example utterance with a switch into English is shown below:

> **Speaker A:** *la composición es increíblemente asociada a Joachim because la tocó ahí primero.*
> [Eng – "The piece is strongly associated with Joachim because he played it there first."]

Spanish-Guaraní Dataset. This dataset contains Spanish-Guaraní utterances in social media and news contexts. Each example is a tokenized sentence, where every token is annotated with a language tag or named entity label (e.g., gn for

Guaraní, `es-b-ul` for Spanish beginning token, `ne-b-org` for the beginning of an organization entity). An illustrative example from the dataset is shown below:

> **@USER:** Movilización kakuaa opu'ãva tiranía venezolana rehe.
> [Eng – "A large mobilization rising up against the Venezuelan tyranny."]

In this example, `@USER` is labeled as a named entity (`ne-b-per`), while tokens such as *kakuaa opu'ãva* and *rehe* are labeled as Guaraní and the rest as Spanish. The mixture of Guaraní and Spanish illustrates natural code-switching behavior.

Code-Switch Subset. To analyze syntactic behavior in mixed-language contexts, we automatically filtered for CSW sentences in both of these datasets. A sentence was classified as CSW if it contained at least two tokens from different language tags (e.g., one in English and one in Spanish). Table 2 summarizes the number of sentences and tokens in both the full and code-switched subsets for each dataset.

3.2 Experimental Setup

To generate syntactic annotations for these datasets, we developed a lightweight pipeline powered by GPT-4.1 (version `gpt-4.1-2025-04-14`). We use the OpenAI API with a deterministic configuration: `temperature=0`, `top_p=1`, and `max_tokens=3000`. Each prompt consists of a system instruction followed by a user message including the CSW sentence and a request for token-level annotation in UD format. This pipeline is detailed further in Section 4. The Spanish-English dataset also includes conversational features typical of spontaneous speech, such as ellipsis, interjections, repetitions, and hesitations, which are known indicators of informal or spoken registers (Georgi et al., 2021). We flag such examples using a binary column `SPEC` to facilitate future syntactic and discourse-level studies that may benefit from separate treatment of these constructions.

Our resulting pipeline processes only the CSW subset of each dataset and outputs a CoNLL-like table with eight columns: token index (ID), token form (FORM), language tag (LANG), lemma (LEMMA), Universal POS tag (UPOS), syntactic head index (HEAD ID), syntactic head token (HEAD), and dependency relation (DEPREL). Native speakers of the respective language pairs reviewed and corrected the model outputs to ensure annotation accuracy. An example of this format, based on a CSW sentence from the Spanish-English dataset, is shown in Table 3. The resulting datasets are released under a permissive open-source license to encourage further research in low-resource and multilingual parsing.

ID	Token Form	LANG	LEMMA	UPOS	HEAD ID	HEAD	DEPREL
1	and	en	and	CCONJ	7	same	cc
2	tú	es	tú	PRON	3	sabes	nsubj
3	sabes	es	saber	VERB	7	same	conj
4	it	en	it	PRON	6	was	nsubj
5	was	en	be	AUX	7	same	cop
6	not	en	not	PART	5	was	advmod
7	same	en	same	ADJ	0	root	root
8	.	other	.	PUNCT	7	same	punct

Table 3: UD-style annotation of the CSW sentence "and tú sabes it wasn't the same," (Eng. "and you know it wasn't the same").

4 Methodology

Our methodology integrates four components to build and analyze syntactically annotated CSW data: (1) developing Bilingua Pipeline for generating UD annotations; (2) validating the annotations through expert review and evaluating accuracy; (3) conducting structural analysis on intra-sentential switch points; and (4) extending this framework to low-resource languages. Figure 2 provides an overview of the full pipeline.

4.1 Development of Bilingua Pipeline

To create the Bilingua Pipeline, we used GPT-4.1 via the OpenAI API to generate UD annotations for CSW sentences. The process of generating accurate UD annotations is already a complex and time-consuming task for monolingual data, the challenge becomes even greater in the context of bilingual or CSW input. Therefore, the prompts for Bilingua Pipeline were carefully crafted using few-shot examples and refined through iterative testing, incorporating feedback from linguistic experts familiar with the targeted language pairs. Our prompts were specifically designed to handle the non-canonical structures typical of spoken and informal language, such as contractions, repetitions, incomplete sentences, and elliptical coordination. The model was instructed to produce token-level annotations based on the traditional CoNLL-U format that include ID, FORM, LANG, LEMMA, UPOS, HEAD ID, HEAD, and DEPREL. Full details of the prompt structures are provided in Appendix B and C.

4.2 Handling of Informal Syntactic Structures

In conversational and code-switched speech, non-canonical structures such as dropped words, hesitations, and merged tokens frequently occur (Georgi et al., 2021). These phenomena pose challenges for automatic dependency parsing, as many UD parsers assume well-formed, complete sentences. Here we describe how Bilingua Pipeline's prompts account for these informal constructions so that resulting annotations remain linguistically coherent.

Incomplete or Elliptical Sentences. Conversational speech between multiple speakers often consists of interruptions between dialogues, leading to incomplete sentences or ellipses. We distinguish between truly incomplete sentences and elliptical ones that omit syntactic elements but remain interpretable. Table 4 and 5 show how we assign dependencies using `dep`, `orphan`, or _ in such cases.

FORM	LEMMA	UPOS	HEAD ID	HEAD	DEPEND
It	it	PRON	2	s'	nsubj
's	be	AUX	0	root	root
the	the	DET	4	end	det
end	end	NOUN	2	s'	attr
of	of	ADP	_	_	case
the	the	DET	_	_	det
.	.	PUNCT	2	s'	punct

Table 4: UD tagging of an incomplete sentence ['It's the end of the...'] with missing final noun phrase.

FORM	LEMMA	UPOS	HEAD ID	HEAD	DEPEND
Me	yo	PRON	2	gusta	iobj
gusta	gustar	VERB	0	root	root
comer	comer	VERB	2	gusta	xcomp
y	y	CCONJ	2	gusta	cc
a	a	ADP	6	ella	case
ella	ella	PRON	2	gusta	conj
bailar	bailar	VERB	6	ella	orphan
.	.	PUNCT	2	gusta	punct

Table 5: UD tagging of an elliptical sentence ['Me gusta comer y a ella bailar' (Eng- 'I like eating and she dancing.')] with gapping.

Repetitions. In spoken interaction, repetitions often arise due to hesitation or self-correction. When repetitions occur, both instances are assigned the same syntactic role and head to preserve structural alignment. We use a similar approach in our prompting to handle repetitions in the dataset. See Table 6 for an example.

Contractions and Punctuation. In traditional linguistic parsers, contractions (e.g., *don't, they're*) are tagged by splitting them into their components.

FORM	LEMMA	UPOS	HEAD ID	HEAD	DEPEND
Yo	yo	PRON	4	sé	nsubj
yo	yo	PRON	4	sé	nsubj
no	no	PART	4	sé	advmod
sé	saber	VERB	0	root	root
.	.	PUNCT	4	sé	punct

Table 6: UD tagging of a sentence ['Yo yo no sé.' (Eng - 'I I don't know.')] with hesitation and subject repetition.

The prompt for Bilingua Pipeline instructed the LLM to follow the same approach for English tokens and assign proper dependency roles to each part. Additionally, punctuation was consistently attached to the root or main clause verb using the `punct` label. See Table 7 for a typical output.

FORM	LEMMA	UPOS	HEAD ID	HEAD	DEPEND
She	she	PRON	3	go	nsubj
did	do	AUX	3	go	aux
n't	not	PART	2	did	advmod
go	go	VERB	0	root	root
.	.	PUNCT	3	go	punct

Table 7: UD tagging of a sentence ['She didn't go.'] illustrating contraction splitting.

4.3 Annotation Validation and Evaluation

It is important to note that evaluating the Bilingua Pipeline-generated UD annotations on CSW Spanish-English and Spanish-Guaraní data is quite challenging due to the absence of established gold-standard datasets. We measured the annotation quality of our resultant datasets using the LAS, which assesses both correct head assignment and dependency relation for each token, and we also report individual accuracy for UPOS and DEPREL tags. To compute these metrics, we compared model outputs against two reference sets:

1. **Manually annotated gold standard.** A small subset of 20 sentences (248 tokens) was selected at random and fully annotated by linguistic experts. Creating this bilingual gold standard is a tedious process, and it requires constructing complete parse trees and assigning UPOS, head indices, and dependency labels by hand. LAS was then calculated by comparing the LLM output to these expert annotations.

2. **Human-revised LLM output.** In a faster second round, two bilingual annotators reviewed and corrected the model's own parse outputs. Inter-annotator agreement on this

Functional Domain	Semantically Similar UD Tags
Verbal Core	root, aux, cop
Clausal Complements	xcomp, ccomp
Discourse/Clause Linking	parataxis, appos, conj, discourse, mark, advmod
Adjectival/Clausal Modifiers	amod, acl, acl:relcl
Nominal Modifiers	nmod, obl, advmod
Numeric/Adjectival Modifiers	nummod, amod
Referential/Appositional Structures	appos, nmod, conj

Table 8: Groups of semantically similar UD tags considered equivalent for evaluation purposes.

subset reached Cohen's Kappa of 0.85, indicating high consistency despite the structural ambiguity of CSW text. This approach accepts the LLM's annotations if they fall within a linguistically plausible range, even when differing from canonical UD labels.

One of the reasons for adopting the second evaluation approach is that it provides a rigorous benchmark for assessing LLMs' performance on CSW contexts, particularly in the absence of pre-existing gold annotations. Another key motivation for this method is that the traditional method of LAS calculation for UD parsers does not account for semantic similarity between dependency labels or POS categories. For example, while the distinction between AUX and VERB is clearly defined in the UD guidelines (copulas and auxiliary verbs are to be tagged as AUX only), there are other cases where tagging ambiguity is more justified. Consider the verb *"want"* in the sentence *"I want to ride my bicycle"*. Depending on the analysis, *"want"* may be treated as a main verb with a clausal complement (ccomp) or with an open clausal complement (xcomp), reflecting subtle differences in control and argument structure. Traditional LAS, however, penalizes such alternatives equally, even when both are linguistically reasonable. The expert review process accounts for such variation and tolerates plausible alternative annotations when they are linguistically motivated. To accommodate these subtleties, we treat sets of semantically related UD tags (see Table 8) as equivalent. Differences within each group are not counted as errors under our human-aligned evaluation (see Appendix A for annotator guidelines). As an additional baseline, we also trained a multilingual UD parser via sequence labeling (UDSL) to compare the results of Bilingua Pipeline; full experimental details are provided in Section 4.5.

Our experience with the evaluation process for Bilingua Pipeline's outputs suggests that the current UD evaluation metrics can be too rigid for complex, multilingual data such as dialogues or real-life conversational text. Developing a more flexible evaluation framework that systematically recognizes acceptable annotation variants would benefit future work on dependency parsing in CSW and other non-standard text genres.

4.4 Syntactic Analysis of Code-Switching

To demonstrate the utility of our LLM-based annotated datasets for linguistic research, we conducted a structural analysis of intra-sentential code-switching, a phenomenon in which two languages are used within a single sentence or utterance (Poplack, 1980), as it presents particularly interesting structural challenges for syntactic analysis. A switch point was defined as a token where the language tag differed from that of the preceding token. For each switch-in token, we extracted its part-of-speech (POS), dependency label, and language tag to study syntactic behavior at the boundary.

We aggregated switch-in tokens and examined which syntactic roles (e.g., determiners, objects, discourse markers) are most commonly involved in switching. This analysis helps answer questions such as whether switches occur more often in determiner positions or whether object slots are more flexible across languages. It also enables us to draw structural generalizations about how different language pairs manage code-switching syntactically, particularly with respect to typologically distinct pairs like Spanish-Guarani, where strong differences in grammatical structure (e.g., head-marking, word order, affix richness) may affect switch behavior. Consider the following example:

I bought un coche blanco.

[Eng-'I bought a white car.']

Here, the switch-in token *"un"* is labeled as a determiner (det), offering one instance of switching into a noun phrase. These cases are especially informative in Spanish-Guaraní, where mismatches such as article absence in Guaraní contrast with Spanish structures. To ensure a meaningful syntactic analysis, we filtered for CSW sentences containing at least three tokens. This yielded 1,711 annotated Spanish-English sentences and 877 annotated Spanish-Guaraní sentences suitable for more informed analysis.

4.5 Training a Universal Dependencies Parser with Sequence Labeling

In addition to generating LLM-based annotation, we trained a multilingual dependency parser using a sequence labeling approach. It can be used as an alternate baseline for the task undertaken by the Bilingua Pipeline. We used the CoDeLin framework and fine-tuned `bert-base-multilingual-cased` with two encoding strategies: Relative (REL) and Absolute (ABS), following Roca et al. (2023) to train this parser. The training data combined UD English EWT (Silveira et al., 2014) and Spanish AnCora (Taulé et al., 2008) datasets. These were merged, shuffled, and split into training, development, and test sets. The data was then encoded into sequence labels using CoDeLin. The parser was trained for 30 epochs using a learning rate of $1e{-}5$, batch size of 64, weight decay of 0.001, and Adam epsilon of $1e{-}7$. We decoded the predictions into CoNLL-U format for evaluation using the standard CoNLL 2018 script (Zeman et al., 2018). This parser (UDSL) serves as a supervised benchmark for parsing performance in monolingual and CSW contexts.

4.6 Extension of Bilingua Pipeline to Low-Resource Languages

We extended the Bilingua Pipeline to low-resource language pairs where no syntactically annotated CSW data is available to train supervised parsers. The Spanish-Guaraní dataset serves as a case study. Motivated by the scalability of LLMs in low-resource settings, we used prompt-based UD annotation combined with native speaker review to bootstrap syntactic resources without requiring large annotated corpora. Prompt and architectural details are provided in Appendix C. This dataset presented unique challenges due to its length and complexity. Nonetheless, the parser generated meaningful annotations that enable valuable syntactic analysis for code-switching in under-resourced, typologically diverse languages.

5 Results and Linguistic Analysis

5.1 Results of Bilingua Pipeline

Table 9 compares the performance of Bilingua Pipeline on the LAS metric on CSW datasets. For Spanish-English, the LLM-based annotation achieved **76.32%** score when compared with the Gold Annotation and **95.29%** when compared with Human reviewed outputs, outperforming earlier models (Sec.2) that reported LAS scores below 75%. In addition to the Spanish-English evaluation, we also report results for the Spanish-Guaraní dataset, which represents one of the first attempts to syntactically annotate CSW data involving this low-resource language. The Spanish-Guaraní dataset achieved LAS scores of **59.90%** and **77.42%**, respectively, on the two methods mentioned above. Notably, the Universal Dependencies Spanish-English model (UDSL), a general-purpose multilingual parser, achieved only **14.71%** LAS when compared with the Gold Annotation, highlighting the limitations of off-the-shelf models when applied to CSW data.

Dataset	Gold Annotation (LAS)	Human Review (LAS)
Spanish-English	76.32%	95.29%
Spanish-Guaraní	59.90%	77.42%
UDSL (Spa-Eng)	14.71%	–%

Table 9: Comparison of LAS before and after expert review on code-switched data.

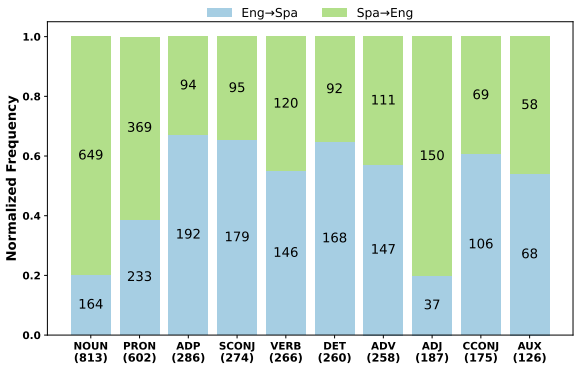
Dataset	UPOS	DEPREL	LAS
Spanish-English	99.54%	97.14%	95.29%
Spanish-Guaraní	84.21%	59.90%	77.42%

Table 10: UPOS, DEPREL, and overall LAS performance after expert revision.

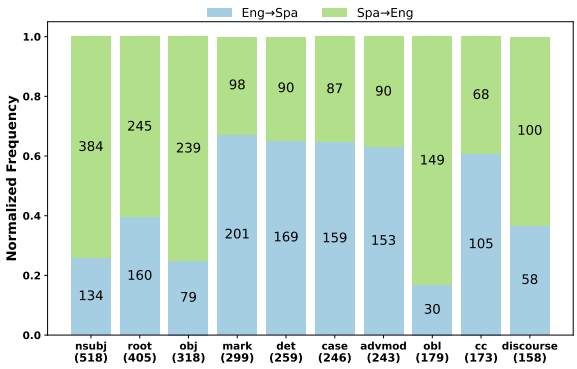
We also carried out a detailed analysis of the human-reviewed output results to showcase the overall accuracy of the UD tags generated by the Bilingua Pipeline. As shown in Table 10, the parser achieves high accuracy across UPOS, DEPREL, and LAS metrics (above **90%**), particularly in the Spanish-English dataset. While the Spanish-Guaraní dataset shows slightly lower performance in dependency parsing, UPOS tagging remains strong, which is a great result for developing linguistic resources for low-resource languages. These results suggest that LLMs can robustly handle syntactic analysis in bilingual contexts, outperforming both hybrid models and general-purpose multilingual parsers not specifically trained on code-switching.

5.2 Qualitative Error Analysis of LLM-based Dependency Parsing

Although the prompt provides explicit guidelines for handling repetitions and ellipsis, LLM's responses remain inconsistent in applying these rules.



(a) Normalized distribution of Universal Part-of-Speech (UPOS) tags at CSW points in the Miami Spanish-English dataset. Notably, NOUN, PRON, and ADP are the most common categories at switch points, with NOUN exhibiting a high rate of switching from Spanish to English.



(b) Normalized distribution of Universal Dependency Relations (DEPREL) at CSW points in the Miami Spanish-English dataset. The most common relations at switch points are nsubj, root, and obj, indicating that syntactic subjects and core verbal arguments are key sites for switching.

Figure 3: Normalized frequency distribution of syntactic categories (UPOS and DEPREL) at CSW points across both switching directions. Bars show Eng→Spa (blue) and Spa→Eng (green) proportions. Absolute counts are shown inside bars; totals in parentheses.

It sometimes analyzes repetitions as coordinations, while in other cases it repeats the dependency structure for each clause. While both analyses are linguistically plausible, this inconsistency may affect the reliability of syntactic generalizations about CSW points. We also observed inconsistencies in the analysis of functional verbs, including auxiliaries, modal verbs, and light verbs such as Spanish *ser* ('to be'). The LLM-based annotation oscillates between assigning these functional elements the syntactic role of root and attaching them to other verbal heads, indicating variability in the treatment of verbal dependency structures. Notably, native speaker feedback on the Guaraní data often identified morphologically complex words that should be split into multiple tokens for accurate syntactic and POS annotation. For instance, the token *noñeguahëi* ('not come') was suggested to be split into a negational adverb and a verb, each with its own POS tag. This observation underscores the need for language-specific morphological preprocessing in low-resource and agglutinative languages and suggests that future improvements to LLM annotation pipelines may benefit from integrating language-specific morphological analyzers or token-splitting mechanisms. Detailed examples investigating specific errors are provided in App D.

5.3 Syntactic Generalizations at CSW Points in the English-Spanish dataset

While prior research on the structural characteristics of CSW points in NLP has largely focused on part-of-speech (POS) tags (Martínez, 2020), our analysis advances this work by leveraging UD to capture syntactic roles at switch sites. Figures 3a and 3b show the normalized distributions of UPOS and DEPREL tags across both switch directions. We find that subject positions (nsubj) are among the most frequent loci of switching, particularly in English-to-Spanish segments. Although this pattern is not consistently emphasized in the literature, it aligns with broader findings that permit switching at major syntactic boundaries, including clause-initial positions. Classic studies such as (Poplack, 1980; Myers-Scotton, 2002) highlight noun phrases, especially determiners (det), modifiers (amod), and prepositions (case), as common switch sites when structural equivalence holds. Our results support this, showing frequent switches in the nominal domain and at clause boundaries (mark, cc, discourse). The prominence of nsubj may reflect language-pair-specific traits or discourse patterns, such as topic-prominence or left dislocation. These findings suggest that dependency relations uncover fine-grained switching patterns not captured by POS tags alone (Martínez, 2020), and motivate the need for richer syntactic annotation in bilingual corpora.

Switching within the main verb or root predicate (i.e., the root in UD) has been considered highly constrained in the Spanish-English CSW literature. Early studies such as Poplack (1980) and models like the Matrix Language Frame (Myers-Scotton, 1993) argue that verb phrase boundaries are typi-

cally resistant to switching due to morphosyntactic incompatibilities between the two languages. Corpus-based studies (e.g., Toribio, 2001; Bullock and Toribio, 2009; Parafita Couto et al., 2015) confirm that switching at or within the main verb is rare, with bilingual speakers favoring switches at clause boundaries. When switches do occur within the verbal domain, they tend to involve semantically transparent structures or frequent bilingual patterns. Our findings suggest that this restriction of the linguistic theory needs to be reexamined. The high frequency of code-switches at the root level may partly reflect parser errors, such as incorrectly analyzing modals or auxiliaries as roots. We acknowledge this limitation and plan to address it in future work by incorporating manual validation or model calibration strategies.

5.4 Syntactic Generalizations at CSW Points in the Spanish-Guaraní dataset

Our analysis of Spanish-Guaraní code-switching reveals broader syntactic flexibility than is typically observed in Spanish-English bilingualism. As shown in Figure 7 (see Appendix E), switch points in the Spanish-Guaraní data occur not only at canonical noun phrase boundaries, such as subjects (`nsubj`), objects (`obj`), and determiners (`det`), but also at clause-internal positions, including auxiliaries, modals, and root-level verbs. These sites are generally more resistant to switching in other language pairs. In contrast, the Spanish-English data (Figure 3) exhibit a more constrained switching pattern, largely centered on nominal boundaries and functional markers such as `mark` and `case`, with verbal heads showing lower susceptibility. The relative openness of Guaraní to verbal integration appears to license a wider range of switch locations. Further analysis, including a breakdown of emoji vs. non-emoji subsets and the role of discourse-level cues, is presented in Appendix E.

6 Conclusion

This work introduces BiLingua Pipeline, a novel pipeline for syntactic annotation of code-switched data using LLMs, supported by expert human validation. By leveraging GPT-4.1 and linguistically informed prompting, we produced high-quality UD annotations for Spanish-English and Spanish-Guaraní code-switching. Our results show that LLM-based annotations outperform conventional parsers in syntactic accuracy, particularly at switch points where monolingual models typically fail. This performance gap is especially pronounced under our second evaluation method, which compares LLM outputs against human-revised annotations and does not penalize linguistically plausible variation. By incorporating groups of semantically similar dependency labels, this evaluation provides a more realistic benchmark for parsing in multilingual settings. Importantly, we release the first publicly available UD-annotated datasets for Spanish-English and Spanish-Guaraní CSW, addressing a critical gap in multilingual NLP resources. These datasets and our annotation methodology not only enable fine-grained analysis of code-switching behavior but also provide a foundation for advancing low-resource dependency parsing.

Limitations

The UD framework provides a cross-linguistically consistent approach to syntactic annotation, but its complexity poses challenges for annotators unfamiliar with formal linguistic parsing conventions. Without such training, annotation quality may vary, and comparisons with other UD-based datasets may be less reliable. To ensure consistency and interoperability, we emphasize the importance of equipping native Guaraní speakers with detailed UD guidelines and hands-on annotation practice. This will support the creation of high-quality, linguistically grounded resources for low-resource languages.

Ethical Considerations

Our work investigates the use of LLMs for syntactic annotation of code-switched language data, with a focus on Spanish-English and Spanish-Guaraní. While this research contributes to the development of more inclusive and multilingual NLP tools, it also raises several ethical considerations. The application of LLM-based syntactic annotation involves the risk of propagating model biases and structural inaccuracies, especially in under-resourced language contexts where gold-standard syntactic annotations are scarce. If such annotations are used for downstream tasks without human oversight, there is a danger of entrenching erroneous linguistic assumptions about bilingual speakers and their language practices. Naive or unsupervised deployment of LLMs in multilingual settings could unintentionally reinforce dominant-language structures or misrepresent code-switching norms. Be-

fore deploying such tools in real-world contexts, appropriate measures should be taken to ensure reliability and linguistic expertise. We have used AI assistants (Grammarly and ChatGPT) to address the grammatical errors and rephrase the sentences.

Acknowledgement

We thank the anonymous annotators and reviewers for their constructive suggestions and help. We extend our gratitude to the Research Computing (RC) and Enterprise Technology at ASU for providing computing resources and access to the ChatGPT enterprise version for experiments. We acknowledge grants GAP (PID2022-139308OA-I00) funded by MICIU/AEI/10.13039/501100011033/ and ERDF, EU; LATCHING (PID2023-147129OB-C21) funded by MICIU/AEI/10.13039/501100011033 and ERDF, EU. CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01). Furthermore, this research was supported by the International, Interdisciplinary and Intersectoral Information and Communications Technology PhD programme (3-i ICT) granted to CITIC and supported by the European Union through the Horizon 2020 research and innovation programme under a Marie Skłodowska-Curie agreement (H2020-MSCA-COFUND), GA 101034261.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Irshad Bhat, Riyaz A. Bhat, Manish Shrivastava, and Dipti Sharma. 2018. Universal Dependency parsing for Hindi-English code-switching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 987–998, New Orleans, Louisiana. Association for Computational Linguistics.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City. Association for Computational Linguistics.

Barbara E Bullock and Almeida Jacqueline Toribio. 2009. Trying to hit a moving target: On the sociophonetics of code-switching. *International Journal of Bilingualism*, 13(2):165–193.

Jie Chi and Peter Bell. 2024. Analyzing the role of part-of-speech in code-switching: A corpus-based study. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1801–1811, St. Julian's, Malta. Association for Computational Linguistics.

Luis Chiruzzo, Marvin Agüero-Torales, Gustavo Giménez-Lugo, Aldo Alvarez, Yliana Rodríguez, Santiago Góngora, and Thamar Solorio. 2023. Overview of gua-spa at iberlef 2023: Guarani-spanish code switching analysis.

Margaret Deuchar, Peter Davies, Judith Herring, María C. Parafita Couto, and Dan Carter. 2014. Building bilingual corpora. In Enlli M. Thomas and Ineke Mennen, editors, *Advances in the Study of Bilingualism*, pages 93–110. Multilingual Matters, Bristol.

Melinda Fricke and Gerrit Jan Kootstra. 2016. Primed codeswitching in spontaneous bilingual dialogue. *Journal of Memory and Language*, 91:181–201.

Ryan Georgi, Yating Wang, and Fei Xia. 2021. Evaluating dependency parsers on spoken language transcripts. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 33–43, Online. Association for Computational Linguistics.

Víctor Soto Martínez. 2020. *Identifying and Modeling Code-Switched Language*. Ph.D. thesis, Columbia University, New York, NY.

Carol Myers-Scotton. 1993. *Duelling languages: Grammatical structure in code-switching*. Oxford University Press.

Carol Myers-Scotton. 2002. *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford University Press, Oxford, New York.

Maria Carmen Parafita Couto, Marianne Gullberg, and Pieter Muysken. 2015. Subject positioning in spanish–english code-switching. *Linguistic Approaches to Bilingualism*, 5(3):277–300.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en español: Toward a typology of code-switching. *Linguistics*, 18(7-8):581–618.

Shruti Rijhwani and Thamar Solorio. 2016. Estimating code-switching on twitter with a novel generalized word-level classification model. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 33–42, Austin, Texas. Association for Computational Linguistics.

Shruti Rijhwani, Lawrence Wolf-Sonkin, Victor Kuperman, Timothy Baldwin, and Thamar Solorio. 2017. Analyzing code-switched social media text. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Diego Roca, David Vilares, and Carlos Gómez-Rodríguez. 2023. A system for constituent and dependency tree linearization. *Kalpa Publications in Computing*.

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-2014)*.

Thamar Solorio, Emily Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Di Lin, et al. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*.

Thamar Solorio and Yang Liu. 2008. Part-of-speech tagging for english-spanish code-switched text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060, Honolulu, Hawaii. Association for Computational Linguistics.

Michalina Strzyz, David Vilares, and Carlos Gómez-Rodríguez. 2019. Viable dependency parsing as sequence labeling. 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 717–723, Minneapolis, Minnesota. Association for Computational Linguistics.

Hakyung Sung and Gyu-Ho Shin. 2025. Second language korean universal dependency treebank v1. 2: Focus on data augmentation and annotation scheme refinement. *arXiv preprint arXiv:2503.14718*.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Lrec*, volume 2008, pages 96–101.

Almeida Jacqueline Toribio. 2001. Accessing bilingual code-switching competence. *International Journal of Bilingualism*, 5(4):403–436.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Şazi Murat Özateş, Özlem Çetinoğlu, Reut Tsarfaty, Dilek Küçük, and Olcay Taner Yıldız. 2022. Improving code-switching dependency parsing with semi-supervised auxiliary tasks. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1126–1141, Seattle, United States. Association for Computational Linguistics.

A Annotation Guidelines

Tag	Label	Example(s)
NOUN	Noun	*house, tree*
VERB	Verb	*to run, to speak*
ADJ	Adjective	*big, pretty*
PRON	Pronoun	*I, they*
ADV	Adverb	*quickly, well*
ADP	Adposition	*in, under*
DET	Determiner	*the, his/her*
PROPN	Proper noun	*Spain, Juan*
NUM	Numeral	*three, twenty*
CCONJ	Coordinating conjunction	*and, but*
SCONJ	Subordinating conjunction	*because, although*
PART	Particle	*not, yes*
INTJ	Interjection	*Hello!, Ugh!*
PUNCT	Punctuation	*., ?*
other	Miscellaneous	*Context-dependent*

Table 11: Common UPOS tags provided during annotator training.

Tag	Label	Example
nsubj	Nominal subject	*She ran* → *She* is the nsubj of *ran*
obj	Direct object	*I saw him* → *him* is the obj of *saw*
iobj	Indirect object	*I gave her a book* → *her* is the iobj
root	Sentence root	*He left* → *left* is the root
det	Determiner	*The book* → *The* is the det of *book*
case	Case marker	*in the house* → *in* is the case of *house*
amod	Adjectival modifier	*big house* → *big* is the amod
advmod	Adverbial modifier	*He ran quickly* → *quickly* is the advmod
conj	Conjunct in coordination	*tea and coffee* → *coffee* is the conj
cc	Coordinating conjunction	*tea and coffee* → *and* is the cc

Table 12: Key UD dependency relations introduced to annotators.

We provided native speakers of Guaraní from Paraguay with a linguistic background with an overview of UD annotation scheme before annotation. This included explanations and examples for POS tags and DEPREL labels. A subset of the most relevant tags is listed in the Tables 11 and 12. For Spanish-English annotations, native speakers of English and Spanish with a linguistic background were instructed to use the official Universal Dependencies documentation at https://universaldependencies.org/ as a reference for POS and DEPREL labels during annotation.

B Prompts for the Bilingua Pipeline

The figures below show the full prompt design used to guide the Bilingua Pipeline. Figure 4 presents a reference sheet of dependency relation definitions aligned with Universal Dependencies (UD) conventions. Figure 5 shows the full base prompt given to the model, specifying the expected token-level format, as well as tailored instructions for handling special cases in CSW input.

Instructions for identifying Dependency Relations

```
Definitions and further instructions for applicable
Dependency tags for Spanish-English sentences:
Core Syntactic Relations
nsubj: Nominal subject – The syntactic subject of a
clause.
obj: Object – The direct object of a verb.
iobj: Indirect object – A secondary object, often
marked with a preposition.
csubj: Clausal subject – A clause functioning as the
subject of another clause.
ccomp: Clausal complement – A clause functioning as
the object of a verb.
xcomp: Open clausal complement – A non-finite clause
that shares its subject with the main verb.

Modifiers and Complements
amod: Adjectival modifier – An adjective modifying a
noun.
nmod: Nominal modifier – A noun phrase modifying
another noun, often introduced by a preposition.
advmod: Adverbial modifier – An adverb modifying a
verb, adjective, or other adverb.
obl: Oblique nominal – A nominal dependent
introduced by a preposition.
vocative: Vocative – A noun used for direct address.

Function Words and Connectors
det: Determiner – An article or quantifier modifying
a noun.
case: Case marking – A preposition or postposition
introducing a nominal.
mark: Marker – A subordinating conjunction
introducing a clause.
cc: Coordinating conjunction – A word that connects
two coordinated elements.
conj: Conjunct – An element in a coordination.

Structure and Function Management
cop: Copula – A linking verb (typically "ser" or
"estar").
aux: Auxiliary – An auxiliary verb used to form
tense, aspect, or mood.
punct: Punctuation – Punctuation marks.

Discourse and Pragmatic Elements
discourse: Discourse element – Words or phrases used
to structure discourse (e.g., "pues", "bueno").
parataxis: Parataxis – Loosely connected clauses or
phrases.
dislocated: Dislocated element – Preposed or
postposed element related anaphorically to the
clause.
```

Figure 4: Dependency relation reference sheet provided to the model in the system prompt. The definitions follow UD conventions and include core **syntactic relations, modifiers, function words, clause-level structures, and discourse-related dependencies**. These definitions help constrain the model's predictions to syntactically valid options for Spanish-English code-switch contexts.

Base Prompt

```
Given a Spanish-English code-switched sentence, tag each token with the following fields, using Universal
Dependencies-style annotation conventions:

- "ID" (number): The index of the token in the sentence, starting from 1.
- "FORM" (string): The surface form of the word as it appears in the sentence.
- "LEMMA" (string): The base or dictionary form of the word (e.g., infinitive for verbs, singular for nouns).
- "UPOS" (string): The Universal Part-of-Speech tag (e.g., VERB, NOUN, ADJ).
- "HEAD ID" (number): The ID of the token's syntactic head.
- "HEAD" (string): The FORM of the head token.
- "DEPREL" (string): The dependency relation linking the token to its head (e.g., nsubj, obj, root, aux, cc).

Please follow these additional guidelines:

1. Only one root per sentence. Only one token may have `"HEAD ID": 0`, and that should be the syntactic root of
the sentence. Any additional finite verbs should be connected using `conj`, `parataxis`, or similar relations.

2. Contractions: When a token appears as a contraction (e.g., "wasn't", "they're", "can't"), split the
contraction into two rows sharing the same "ID" and "FORM", but with different lemmas and syntactic roles.

Example - Sentence: "She didn't go ."
<Formatted Output>

3. Repetition:
   - If a word is repeated due to hesitation or repair (e.g., "yo yo no sé"), assign the same dependency label
and head to both repeated tokens.

Example - Sentence: "Yo yo no sé ."
<Formatted Output>

4. Incomplete Sentences or Ellipses:
   - Grammatically incomplete sentence (e.g., "It's the end of the") → tag known words and assign `dep` or use
`_` in HEAD fields where no head exists.
   - Elliptical constructions (e.g., "Me gusta comer y a ella bailar") → use `orphan` to attach a promoted
dependent.

Example - Sentence: "It's the end of the ."
<Formatted Output>

Final Reminders:
- HEAD ID values must match the correct ID of the referenced head token.
- The FORM in the "HEAD" field must exactly match the FORM of the token referenced by the HEAD ID.
- Every token in the sentence (including punctuation) must be included in the output.
- Always use the `punct` relation to attach punctuation (e.g., ., ?, !) to the main clause verb or root.
- Do not omit any token — even emojis, filler words, or interjections should be annotated with `"UPOS":
"other"` and `"DEPREL": "discourse"` or similar where appropriate.

Example - Sentence: "and if you're not doing quality work para qué te van a pagar ?"

Output:
[
  {"ID": 1, "FORM": "and", "LEMMA": "and", "UPOS": "CCONJ", "HEAD ID": 6, "HEAD": "doing", "DEPREL": "cc"},
  {"ID": 2, "FORM": "if", "LEMMA": "if", "UPOS": "SCONJ", "HEAD ID": 6, "HEAD": "doing", "DEPREL": "mark"},
  {"ID": 3, "FORM": "you", "LEMMA": "you", "UPOS": "PRON", "HEAD ID": 6, "HEAD": "doing", "DEPREL": "nsubj"},
  {"ID": 3, "FORM": "'re", "LEMMA": "be", "UPOS": "AUX", "HEAD ID": 6, "HEAD": "doing", "DEPREL": "aux"},
  {"ID": 4, "FORM": "not", "LEMMA": "not", "UPOS": "PART", "HEAD ID": 6, "HEAD": "doing", "DEPREL": "advmod"},
  {"ID": 5, "FORM": "doing", "LEMMA": "do", "UPOS": "VERB", "HEAD ID": 0, "HEAD": "root", "DEPREL": "root"},
  {"ID": 6, "FORM": "quality", "LEMMA": "quality", "UPOS": "ADJ", "HEAD ID": 8, "HEAD": "work", "DEPREL":
"amod"},
  {"ID": 7, "FORM": "work", "LEMMA": "work", "UPOS": "NOUN", "HEAD ID": 6, "HEAD": "doing", "DEPREL": "obj"},
  {"ID": 8, "FORM": "para", "LEMMA": "para", "UPOS": "ADP", "HEAD ID": 10, "HEAD": "qué", "DEPREL": "case"},
  {"ID": 9, "FORM": "qué", "LEMMA": "qué", "UPOS": "PRON", "HEAD ID": 14, "HEAD": "pagar", "DEPREL": "obj"},
  {"ID": 10, "FORM": "te", "LEMMA": "tú", "UPOS": "PRON", "HEAD ID": 14, "HEAD": "pagar", "DEPREL": "iobj"},
  {"ID": 11, "FORM": "van", "LEMMA": "ir", "UPOS": "AUX", "HEAD ID": 14, "HEAD": "pagar", "DEPREL": "aux"},
  {"ID": 12, "FORM": "a", "LEMMA": "a", "UPOS": "PART", "HEAD ID": 14, "HEAD": "pagar", "DEPREL": "mark"},
  {"ID": 13, "FORM": "pagar", "LEMMA": "pagar", "UPOS": "VERB", "HEAD ID": 6, "HEAD": "doing", "DEPREL":
"advcl"},
  {"ID": 14, "FORM": "?", "LEMMA": "?", "UPOS": "PUNCT", "HEAD ID": 6, "HEAD": "doing", "DEPREL": "punct"}
]

Sentence:
```

Figure 5: Prompt used to guide GPT in generating token-level UD annotations for **Spanish-English** code-switched sentences. The prompt outlines the required output format, including standard UD fields (`ID`, `FORM`, `LEMMA`, `UPOS`, `HEAD ID`, `HEAD`, `DEPREL`), and incorporates targeted instructions to address code-switching-specific phenomena. These include rules for handling English **contractions** (e.g., *didn't* → *did* + *n't*), **disfluencies and repairs** (e.g., repeated tokens like *yo yo*), **elliptical or incomplete** constructions, and **punctuation** attachment. The prompt ensures the sentence structure is valid by requiring **one root** token per sentence and giving rules for handling clausal and discourse-level dependencies. A fully formatted example illustrates the desired structure of GPT's response, aligning with UD conventions.

Base Prompt

```
Given a Spanish-Guarani code-switched sentence, tag each token with the following fields, following Universal
Dependencies-style conventions:


- "ID" (number): The index of the token in the sentence, starting from 1.
- "FORM" (string): The surface form of the word as it appears in the sentence.
- "LEMMA" (string): The base or dictionary form of the token (e.g., infinitive for verbs, singular for nouns).
- "UPOS" (string): The Universal Part-of-Speech tag (e.g., VERB, NOUN, ADJ).
- "HEAD ID" (number): The ID of the token's syntactic head.
- "HEAD" (string): The FORM of the head token.
- "DEPREL" (string): The dependency relation linking the token to its head (e.g., nsubj, obj, root, aux, cc).


Please follow these core instructions:

- Only one token should have `"HEAD ID": 0`, which represents the syntactic root of the sentence.
- If another verb or clause seems to behave like a root, it should instead be connected using `conj` or
`parataxis`, not as another root.

For example, in the sentence "Leave me and stay away from me", the first verb "Leave" is the root, and the
second verb "stay" should be tagged as `conj`, not as another root.

---

Final Reminders:
- `HEAD ID` values must exactly match the `ID` of the referenced token.
- The `HEAD` field must match the `FORM` of the referenced token.
- Every token in the sentence (including punctuation, emojis, and discourse particles) must be included in the
output.
- Always attach punctuation marks (e.g., `.`, `,`, `?`, `!`) using the `punct` relation, usually to the root
verb or main clause.
- For emojis, fillers, or interjections, use `"UPOS": "other"` and an appropriate `"DEPREL"` such as
`discourse` or `other`.

---

Example 1
Sentence: "Mbae sentido oreko las olimpiadas sin basket 😭"

Output:
[
  {"ID": 1, "FORM": "Mba'e", "LEMMA": "Mba'e", "UPOS": "PRON", "HEAD ID": 2, "HEAD": "sentido", "DEPREL":
"det"},
  {"ID": 2, "FORM": "sentido", "LEMMA": "sentido", "UPOS": "NOUN", "HEAD ID": 3, "HEAD": "oreko", "DEPREL":
"nsubj"},
  {"ID": 3, "FORM": "oreko", "LEMMA": "oreko", "UPOS": "VERB", "HEAD ID": 0, "HEAD": "root", "DEPREL": "root"},
  {"ID": 4, "FORM": "las", "LEMMA": "el", "UPOS": "DET", "HEAD ID": 5, "HEAD": "olimpiadas", "DEPREL": "det"},
  {"ID": 5, "FORM": "olimpiadas", "LEMMA": "olimpiada", "UPOS": "NOUN", "HEAD ID": 3, "HEAD": "oreko",
"DEPREL": "obj"},
  {"ID": 6, "FORM": "sin", "LEMMA": "sin", "UPOS": "ADP", "HEAD ID": 7, "HEAD": "basket", "DEPREL": "case"},
  {"ID": 7, "FORM": "basket", "LEMMA": "basket", "UPOS": "NOUN", "HEAD ID": 3, "HEAD": "oreko", "DEPREL":
"obl"},
  {"ID": 8, "FORM": "😭", "LEMMA": "😭", "UPOS": "other", "HEAD ID": 0, "HEAD": "other", "DEPREL": "other"}
]

---

Example 2
Sentence: "Calmate nde ridicula , cuida de tu novio mba'e pq está siendo comidita del pueblo y ni cuenta gua'u
te das 🙂"

Output:
<Formatted Output>

Sentence:
```

Figure 6: Prompt used to guide GPT in generating token-level UD annotations for **Spanish-Guaraní** code-switched sentences. The prompt defines the required Universal Dependencies (UD) output fields, `ID`, `FORM`, `LEMMA`, `UPOS`, `HEAD ID`, `HEAD`, and `DEPREL`, and enforces structural validity by requiring exactly **one syntactic root** per sentence. The instructions explicitly address how to attach additional verbs or clauses (e.g., using `conj` or `parataxis` rather than a second `root`) and how to treat punctuation and nonstandard tokens such as emojis or discourse particles using the `discourse` or `other` labels. Two fully formatted examples demonstrate how these conventions apply to mixed-language sentences, including Guaraní verbs and Spanish noun phrases. The prompt is designed to handle typologically diverse, low-resource input without preprocessing or morphological segmentation.

C Architecture and Prompts for Spanish-Guaraní Dataset

In constructing the Spanish-Guaraní UD annotations, we retained the original tokenization and sentence segmentation from the source dataset (Chiruzzo et al., 2023). The model was not instructed to split morphologically complex tokens or simplify the data. Consequently, many sentences exceeded 50 tokens and featured complex, clause-rich structures, in contrast to the shorter Spanish-English sentences which average around 5 tokens. This presented additional challenges for parsing accuracy. To address these challenges, we designed a task-specific prompt for Spanish-Guaraní CSW input, shown in Figure 6. The prompt outlines the expected UD output format and includes targeted instructions for dependency structure validity, handling of discourse elements, and typologically diverse constructions. It enables the LLM to produce well-structured annotations without requiring preprocessing or morphological analysis, making it suitable for low-resource and morphologically rich language contexts.

D Extended Qualitative Analysis on CSW Results

Table 13 shows an example of a syntactic structure containing both repetition and ellipsis. The phrase "they're high enough so that él no se…" features repeated subject–copula constructions ("they're") across two overlapping clauses. The LLM inconsistently analyzes these repeated forms, sometimes attaching them in parallel, sometimes duplicating heads. It also treats the Spanish clause "él no se..." as an elliptical construction without resolving the final verb. This example illustrates the model's challenges in managing discourse-level structures and maintaining syntactic coherence across long, CSW utterances. Table 14 illustrates another recurrent issue: inconsistent handling of repeated verbs. In the utterance "hay hay que dice o'clock somewhere," the verb "hay" ('there is') appears twice, a common phenomenon in spontaneous speech. While both instances are valid, the LLM assigns the second instance a conjunct (`conj`) label instead of treating it as a disfluency or repetition of the root. This creates ambiguity in syntactic interpretation and points to the need for guidelines or preprocessing strategies for repeated tokens in CSW input.

ID	FORM	LEMMA	HEAD	DEPREL	LANG
1	but	but	3	cc	eng
2	I	I	3	nsubj	eng
3	think	think	0	root	eng
4	that	that	7	mark	eng
5	they're	they	7	nsubj	eng
5	they're	be	7	cop	eng
6	they're	they	7	nsubj	eng
6	they're	be	7	cop	eng
7	high	high	3	ccomp	eng
8	enough	enough	7	advmod	eng
9	so	so	12	mark	eng
10	that	that	12	mark	eng
11	él	él	12	nsubj	spa
12	no	no	13	advmod	spa
13	se	se	7	advcl	spa
14	.	.	3	punct	–

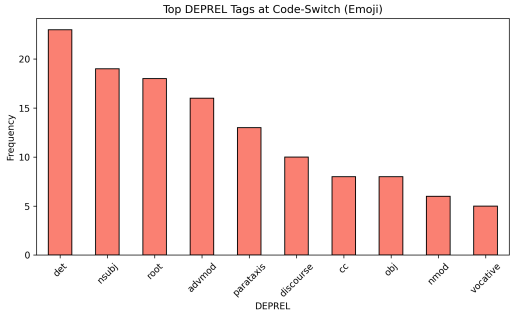
Table 13: Dependency analysis of "But I think that they're high enough so that él no se..." Highlighted rows show repeated subject–copula constructions and an elliptical adverbial clause.

ID	FORM	LEMMA	UPOS	HEAD	DEPREL
1	hay	haber	VERB	0	root
2	hay	haber	VERB	1	conj
5	que	que	PRON	2	obj
6	dice	decir	VERB	5	acl:relcl
10	o'clock	o'clock	NOUN	6	ccomp
11	somewhere	somewhere	ADV	10	advmod

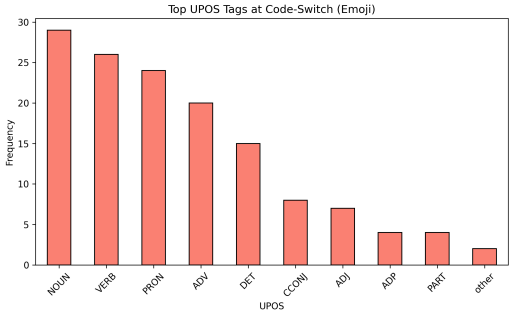
Table 14: Condensed UD analysis of "hay hay que dice o'clock somewhere." Highlighted rows show the repeated verb "hay" handled inconsistently.

E Results for Emoji-Based Variation in Spanish-Guaraní dataset

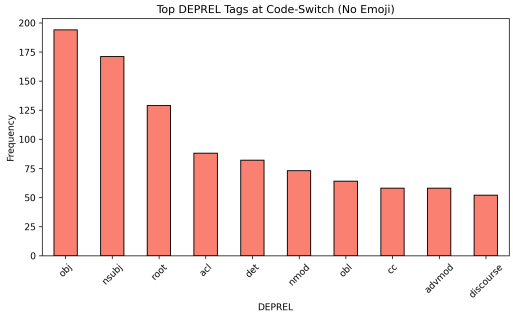
To further understand discourse variation in Spanish-Guaraní code-switching, we divided the dataset into two subsets: messages with emojis and those without. This split approximates a difference in formality and expressiveness, with the emoji-containing subset representing more informal or emotionally expressive communication. Figure 7 presents the top UPOS and DEPREL tags at CSW points for both subsets. In the emoji-rich subset (Figures 7a, 7c), switching occurs frequently at discourse-sensitive syntactic roles such as `discourse`, `parataxis`, and stance-related verbs, in addition to traditional sites like `det`, `nsubj`, and `root`. This suggests that informal messages allow for more syntactic flexibility and that pragmatic context plays an important role in switch placement.



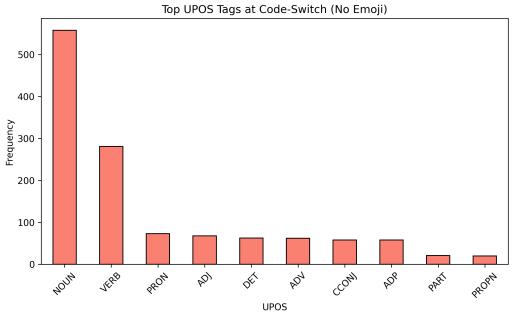
(a) DEPREL + Emoji



(c) UPOS + Emoji



(b) DEPREL – Emoji



(d) UPOS – Emoji

Figure 7: Distribution of DEPREL and UPOS tags at CSW points in Spanish-Guaraní sentences, comparing emoji-containing and non-emoji subsets. (+ Emoji) refers to the subset of the dataset containing emojis, and (- Emoji) refers to the subset without the emojis.

In contrast, the non-emoji subset (Figures 7b, 7d) reveals a more stable switching pattern, with concentration at canonical nominal positions such as `obj`, `nsubj`, `acl`, and `det`, and fewer instances of switching at clause-level discourse functions or verb heads. Together, these results support the observation that structural patterns of switching are not fixed but vary depending on the communicative context. Emoji usage appears to license greater fluidity in syntax, particularly at discourse-level transitions and pragmatically marked segments of bilingual speech.