解析切换：基于 LLM 的 UD 标注用于复杂代码切换和低资源语言

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

代码转换为句法分析带来了复杂的挑战，尤其是在注释数据稀缺的低资源语言环境中。虽然近期的研究探索了使用大型语言模型（LLM）进行序列级标注，但很少有方法系统性地调查这些模型在代码转换环境中捕捉句法结构的效果。此外，现有训练于单语树库的解析器常常未能推广到多语言和混合语言输入。为了弥补这一空白，我们引入了 BiLingua Parser，这是一种基于 LLM 的注释管道，旨在为代码转换文本生成通用依存（UD）注释。首先，我们为西班牙语-英语和西班牙语-瓜拉尼语数据开发了一个基于提示的框架，将少量样本的 LLM 提示与专家审评相结合。其次，我们发布了两个注释数据集，其中包括第一个进行西班牙语-瓜拉尼语 UD 解析的语料库。第三，我们对语言对和交流环境中的转换点进行了详细的句法分析。实验结果显示，BiLingua Parser 在专家修订后可以达到高达 95.29 % 的 LAS，显著优于先前的基线和多语言解析器。这些结果表明，当经过精心指导时，大型语言模型可以作为实用工具，用于在资源不足的代码转换环境中引导句法资源的构建 [1] 。

1 介绍

代码转换（CSW）是一种在世界各地多语言社区中普遍观察到的语言现象。尽管它在口语和非正式的数字交流中很常见，但对于自然语言处理（NLP）来说，尤其是句法解析，仍然是一个复杂的挑战。一个核心问题是，大多数最先进的解析模型是基于单语句法树库训练的，因此在应用于混合语言数据时缺乏鲁棒性 (Özate et al., 2022) 。
之前的 Özate et al. (2022); Rijhwani et al. (2017); Bhat et al. (2018) 的工作通过提出一种半监督依存句法分析框架向解决这一差距迈出了重要的一步，例如，该框架通过辅助序列标注任务增强训练 (Özate et al., 2022) 。他们的模型通过

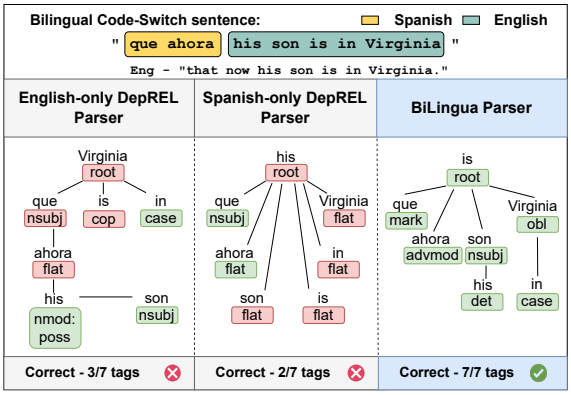


Figure 1: 对一个西班牙语-英语代码混合句子的依存关系预测（DepREL）在三个解析器中的比较。由于单语偏见，英语单独模型和西班牙语单独模型错误地分配了关键关系。相比之下，BiLingua 解析器能够正确地跨语言边界分析完整的结构。

在多语言环境中学习更好的句法结构表示，提高了在土耳其语-德语口语语料库上的解析准确性。然而，即使有这样的增强，现有的模型通常依赖于大量标注数据，这在资源匮乏的语言对中尤为局限。

由于资源缺乏的驱动，我们介绍了 BiLingua Parser，这是一款基于大型语言模型（LLM）的双语句法解析器，特别是使用 GPT-4.1 模型，来生成具有句法标注的 CSW 数据集。图 1 展示了当前单语解析器在一个被广泛研究的 CSW 语言对上表现明显低于 BiLingua Parser。接下来，我们专门针对两个语言对解决这一问题：西班牙语-英语（一个相对资源丰富的代码转换语言对）和西班牙语-瓜拉尼语（一个资源匮乏的语言对，多数语言学工具几乎不可用）(Chiruzzo et al., 2023) 。据我们所知，这些是西班牙语-英语和西班牙语-瓜拉尼语代码转换的首个基于 UD 的句法标注数据集，由母语人士审核。创建和使用 BiLingua Parser 的整个流程如图 2 所示。除了开发 BiLingua Parser 之外，我们还检验了当前句法解析评价指标的局限。为此，我们引入了额外的方法，通过语言学专家的帮助来评估我们的解析器的性能。此

*Equal contribution.
[1]数据和源代码可在 https://github.com/N3mika/ParsingProject 获取。

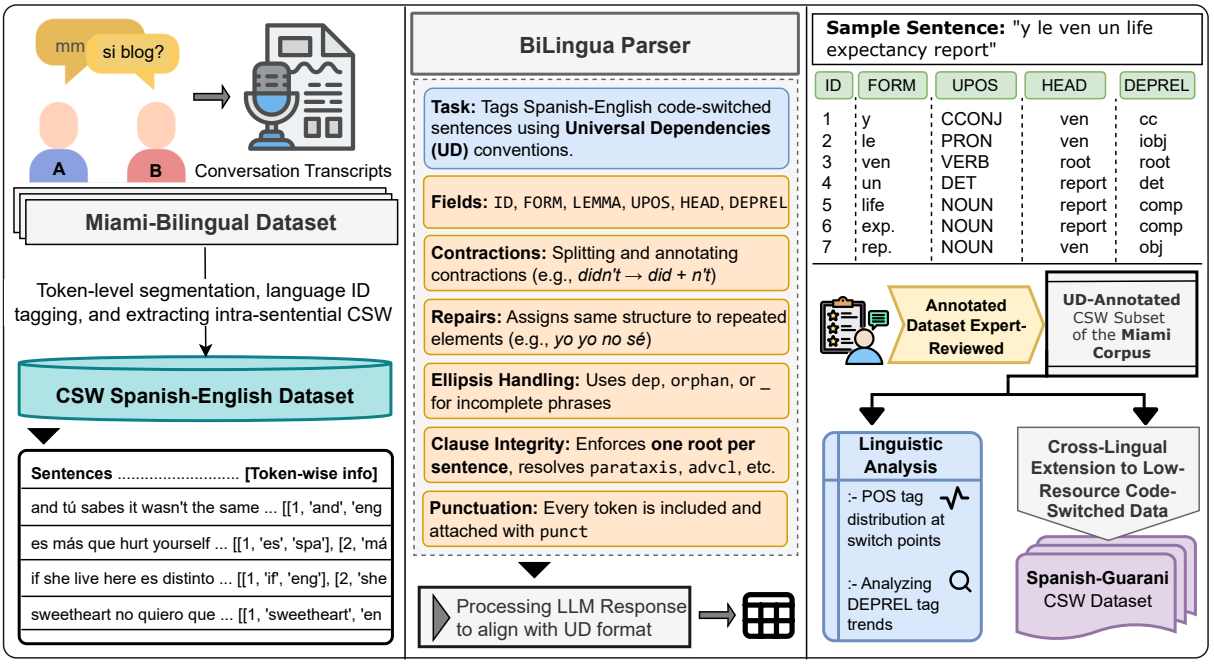


Figure 2: 西班牙语-英语语言转换的 BiLingua 解析器流水线概述。左：来自迈阿密双语语料库的对话记录经过词级分割、语言 ID 标注以及句内代码转换的过滤。中：BiLingua 解析器为 CSW 句子分配 UD 标签，处理缩写、重复、省略和从句结构。右：结果标注的数据集由语言学专家审查，并支持下游任务，如词性/依存关系分析，以及向包括西班牙语-瓜拉尼语在内的低资源环境扩展。

外，我们创建的标注数据集还支持涉及双语使用者的广泛的语言学分析，包括细粒度的转换点行为的理解。尽管之前大多数代码转换的 NLP 结构研究集中在词类（POS）标记上 (Martínez, 2020; Rijhwani and Solorio, 2016; Solorio and Liu, 2008)，但我们使用依存句法分析表明，在这两种语言对中，句法主语（nsubj）是最频繁的转换点之一，并且西班牙语-瓜拉尼语代码转换在转换点上表现出更高的变异。这一发现强调了依存句法分析在跨语言分析转换点中的价值。我们的主要贡献如下：总体而言，我们相信我们的发现和发布的资源将支持未来在代码转换的计算建模和语言分析方面的工作，特别是对于资源匮乏和类型学多样的语言对。

2 相关工作

解析代码转换文本比单语解析复杂得多，因为存在结构变化、混合语法规则和有限的标注语料库。先前的工作 (Rijhwani et al., 2017; Solorio and Liu, 2008; Solorio et al., 2014; Özate et al., 2022) 已证明，单独在单语数据上训练的模型在代码切换文本上表现不佳，除非进行特定的适应。Özate et al. (2022) 通过提出一个包含辅助序列标注任务的半监督解析框架来解决这个问题。他们的方法提高了土耳其语-德语代码切换文本的解析性能，标注依附分（LAS）达到 73 %。类似地，Bhat et al. (2018) 和 Rijhwani et al. (2017) 报告在使用适应模型的印地语-英语数据上的 LAS 分数在 70–72 % 范围内（表 1）。

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 %	–
Özate et al. (2022)	Hindi-English	–	71.93 %
Özate et al. (2022)	Turkish-German	–	73.0 %
Bhat et al. (2018)	Hindi-English	–	71.03 %

Table 1: 以往工作中针对代码转换数据集的词性标注和依存分析性能。

大多数先前的研究集中于词性标注而不是完整的句法解析，并且现有资源仍然只限于少数语言对。虽然存在用于西班牙语-英语代码转换的草案 UD 树库，但它并未公开发布，这进一步说明了句法标注 CSW 数据的稀缺性。为了提高解析速度的努力，例如将依存解析重新作为序列标注任务 (Strzyz et al., 2019; Roca et al., 2023)，在运行时间上有所改善，但仍然严重依赖于单语训练数据。

我们的工作以这些基础为依托，但转向基于 LLM 的注释。像 OpenAI GPT 这样的大型语言模型可以通过示例和语言规则进行提示，以生成注释，而无需广泛的监督训练。我们应用这种技术来生成和评估新的 CSW 数据集，包括

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: 西班牙语-英语和西班牙语-瓜拉尼语数据集的原始和代码转换子集中的句子和标记计数。

一个用于西班牙语-瓜拉尼语的数据集，从而扩大句法工具在服务不足的语言社区中的影响力。

3 数据集和实验

3.1 数据集

我们使用两个现有的数据集来实现我们的 BiLingua 解析器的管道，并创建具有语言学标注的代码转换数据集。我们使用的第一个数据集是迈阿密语料库 (Deuchar et al., 2014)，这是一个广泛用于双语研究和自然语言处理领域的著名西班牙语-英语数据集 (Fricke and Kootstra, 2016; Chi and Bell, 2024; Martínez, 2020)。第二个是来自共享任务 GUA-SPA 的西班牙语-瓜拉尼语数据集：瓜拉尼语-西班牙语代码转换分析 (Chiruzzo et al., 2023)，包括在资源匮乏环境中自发使用的社交媒体和新闻内容。

迈阿密西班牙语-英语语料库。 此数据集包括双语说话者的转录口语互动。每个句子都经过标记和注释，包含语言标签、词性标签和形态特征（例如，`be.V.3S.PRES`）。元数据包括标记索引、句子和话语 ID、说话者身份和文件名。下面展示了一个切换到英语的示例话语：

> 说话者 A: la composición es increíblemente associada a Joachim because 最先在那里演奏。
> [英语–" 这首作品与 Joachim 紧密相连，因为他最先在那演奏过。"]

西班牙语-瓜拉尼语数据集。 该数据集包含社交媒体和新闻背景下的西班牙语-瓜拉尼语话语。每个例子都是一个分词后的句子，其中每个词都标注了语言标签或命名实体标签（例如，`gn` 代表瓜拉尼语，`es-b-ul` 代表西班牙语的起始词，`ne-b-org` 代表组织实体的起始）。下面显示了数据集中的一个示例：

> @USER: Movilización kakuaa opu'ãva tiranía venezolana rehe .
> [英语–"一次对抗委内瑞拉暴政的重大动员。"]

在这个例子中，`@USER` 被标记为一个命名实体（`ne-b-per`），而像 kakuaa opu'ãva 和 rehe 这样的词元被标记为瓜拉尼语，其余则为西班牙语。瓜拉尼语和西班牙语的混合展示了自然的代码切换行为。

代码切换子集 为了分析混合语言环境中的句法行为，我们自动过滤了这两个数据集中代码切换的句子。如果一个句子包含至少两个来自不同语言标签的标记（例如，一个英语和一个西班牙语），它就被归类为代码切换句。表 2 总结了每个数据集的完整和代码切换子集中的句子和标记的数量。

3.2 实验设置

为了生成这些数据集的句法标注，我们开发了一个轻量级流程，利用 GPT-4.1（版本 `gpt-4.1-2025-04-14`）驱动。我们使用 OpenAI API，配置为确定性：`temperature=0`、`top_p=1` 和 `max_tokens=3000`。每个提示由系统指令组成，接着是包含 CSW 句子及请求 UD 格式词级标注的用户消息。该流程在 Section 4 中有详细说明。西班牙语-英语数据集还包括一些典型的自发语音对话特征，如省略、感叹词、重复和犹豫，这些都是非正式或口语体的已知标志 (Georgi et al., 2021)。我们通过使用二元列 `SPEC` 标记这些例子，以促进未来可能需要对这些结构进行单独处理的句法和语篇级研究。

我们的处理流程仅针对每个数据集的 CSW 子集进行处理，并输出一个类似 CoNLL 的表格，该表格有八列：词元索引（ID）、词元形式（FORM）、语言标签（LANG）、词干（LEMMA）、通用词性标签（UPOS）、句法主语索引（HEAD ID）、句法主语词元（HEAD）、和依存关系（DEPREL）。相关语言对的母语人士对模型输出进行了审查和修正，以确保标注的准确性。基于来自西班牙语-英语数据集的代码切换句子的这种格式的一个例子如表 3 所示。最终的数据集以宽松的开源许可证发布，以鼓励对低资源和多语言解析的进一步研究。

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: 代码转换句子 "and tú sabes it wasn't the same"（英语："and you know it wasn't the same"）的 UD 风格标注。

4 方法论

我们的方法整合了四个组成部分来构建和分析具有句法标注的代码混合数据：（1）开发用于生成 UD 标注的双语解析器；（2）通过专家审查验证标注并评估准确性；（3）对句内切换点进行结构分析；（4）将该框架扩展到低资源语言。图 2 提供了整个流程的概览。

4.1 BiLingua 解析器的开发

为了创建 BiLingua Parser，我们使用 OpenAI API 上的 GPT-4.1 生成 CSW 句子的 UD 标注。对于单语数据来说，生成准确的 UD 标注已经是一个复杂且耗时的任务，而在双语或 CSW 输入的情况下，这一挑战变得更加艰巨。因此，BiLingua Parser 的提示通过少量示例精心设计，并通过反复测试进行优化，结合了熟悉目标语言对的语言学专家的反馈。我们的提示专门设计用于处理口语和非正式语言中典型的非规范结构，如缩略、重复、不完整句子和省略协调。模型被指示基于传统的 CoNLL-U 格式生成标注，包括 ID、FORM、LANG、LEMMA、UPOS、HEAD ID、HEAD 和 DEPREL。提示结构的完整细节在附录 B 和 C 中提供。

在对话和代码混杂的语音中，非规范的结构如省略词、犹豫和合并词经常出现。这些现象对自动依存解析提出了挑战，因为许多 UD 解析器假设句子是结构良好且完整的。我们在此解释 BiLingua Parser 的提示如何考虑这些非正式的构造，从而使生成的标注在语言上保持连贯性。

不完整或省略的句子。 多人之间的会话语音通常会有对话中断，导致不完整的句子或省略。我们区分真正不完整的句子和那些省略句法元素但仍可解释的省略句。表格 4 和 5 展示了我们如何在这种情况下使用 `dep` 、`orphan` 或 _ 分配依赖关系。

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: 标记不完整句子 [‘It’s the end of the...’] 的缺失最终名词短语的 UD 标注。

在口语交流中，重复通常是由于犹豫或自我修正引起的。当出现重复时，为了保持结构对齐，两个实例被分配相同的句法角色和中心词。我们在提示中使用类似的方法来处理数据集中的重复。请参见表 6 了解示例。

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: 带有省略的省略句的 UD 标注 [‘Me gusta comer y a ella bailar’ (英文为：‘I like eating and she dancing.’)]。

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: 对句子 [‘Yo yo no sé.’ (英文 - ‘I I don’t know.’)] 进行 UD 标注，包含犹豫和主语重复。

在传统语言解析器中，缩写词（例如，don’t, they’re）通过将它们拆分为其组成部分进行标记。BiLingua Parser 的提示指示 LLM 对英语标记采取相同的方法，并为每个部分分配适当的依存角色。此外，标点符号始终使用 `punct` 标签附着于根或主句动词。请参见表 7 以获得典型输出。

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: 句子 [‘She didn’t go.’] 中的 UD 标注，展示了缩合的分开。

值得注意的是，由于缺乏公认的金标准数据集，在 CSW 西班牙语-英语和西班牙语-瓜拉尼语数据上评估 BiLingua Parser 生成的 UD 标注相当具有挑战性。我们使用标签依附得分（LAS）来衡量我们生成的数据集的标注质量，该得分评估每个标记的正确头部分配和依存关系，同时我们还报告了 UPOS 和 DEPREL 标签的个体准确率。为了计算这些指标，我们将模型输出与两个参考集进行了比较：

1. 人工标注的黄金标准。一个小的句子子集被随机选择，并由语言专家进行完整标注。创建这种双语黄金标准是一个繁琐的过程，且需要手动构建完整的句法树，并分配 UPOS、中心词索引和依存标签。然后通过将 LLM（大语言模型）的输出与这些专家注释进行比较来计算 LAS。

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: 为评估目的而被视为等价的语义相似 UD 标签组。

2. 人工修订的 LLM 输出。在较快的第二轮中，两名通晓双语的注释者审查并纠正了模型自身解析输出。在这个子集上的注释者间一致性达到 Cohen’ s Kappa 系数 0.85，表明即使在 CSW 文本具有结构性歧义的情况下也具有高度一致性。即使与传统的 UD 标签不同，该方法也接受 LLM 的注释，只要它们落在语言学上合理的范围内。

采用第二种评估方法的原因之一是它为评估 LLM 在 CSW 背景下的表现提供了一个严格的基准，特别是在缺乏现有的金标准注释的情况下。这种方法的另一个主要动机是，传统的 UD 解析器的 LAS 计算方法没有考虑依存标签或词性类别之间的语义相似性。例如，虽然在 UD 指南中 AUX 和 VERB 之间的区别是明确的（系动词和助动词只能被标记为 AUX），但在其他情况下，标记歧义更为合理。考虑句子“我想骑我的自行车”中的动词“want”。根据分析不同，“want”可以被视为具有从句补语（ccomp）的主要动词，或者具有开放从句补语（xcomp），反映了在控制和论元结构上的细微差别。然而，传统的 LAS 对这些替代方案一视同仁，即使它们在语言学上都是合理的。专家评审过程则考虑到了这种变异并容忍在语言学上有依据的合理替代注释。为了适应这些微妙之处，我们将一组语义相关的 UD 标签视作等同（见表 ~ 8）。在我们以人为本的评估中，每组内的差异不被计为错误（见附录 ~ A 中的注释者指南）。作为附加的基线，我们还通过序列标注训练了一个多语种 UD 解析器（UDSL），以比较 BiLingua 解析器的结果；全部实验细节见第 4.3 节。

我们在评估 BiLingua Parser 输出的过程中发现，目前的 UD 评估指标对于如对话或真实会话文本等复杂的多语言数据来说可能过于严格。开发一个更加灵活的评估框架，用以系统地识别可接受的注释变体，将有利于未来在代码切换及其他非标准文本类型上的依存分析研究。

4.2 语码转换的句法分析

为了展示我们基于大型语言模型标注的数据集在语言研究中的实用性，我们进行了句内代码转换的结构分析。这种现象是指在一个句子或话语中使用两种语言，这对句法分析提出了极具趣味性的结构性挑战。我们将一个切换点定义为语言标签与前一个标记不同的标记。对于每个切入标记，我们提取了其词性、依赖标签和语言标签，以研究边界处的句法行为。

我们汇总了切换中的词语并检查了哪种句法角色（例如，限定词、宾语、话语标记）最常涉及到语言切换。此分析有助于回答诸如切换是否更常发生在限定词位置或者宾语位置在语言间是否更具灵活性等问题。它还使我们能够就不同语言对如何在句法上处理语言切换提出结构性概括，特别是对于像西班牙语-瓜拉尼语这样的类型学上明显不同的语言对，其语法结构（例如，头标语言特征、词序、词缀丰富性）上的显著差异可能会影响切换行为。请看以下示例：

I bought un coche blanco.
[Eng-‘I bought a white car.’]

这里，切换符“un”被标记为限定词（det），提供了切换到名词短语的一个实例。这些情况在西班牙语-瓜拉尼语中特别有意义，因为在瓜拉尼语中没有冠词与西班牙语结构对比。为了确保有意义的句法分析，我们筛选出包含至少三个词的语码转换句子。这产生了 1,711 个注释过的西班牙语-英语句子和 877 个注释过的西班牙语-瓜拉尼语句子，适合进行更为深入的分析。

4.3 用序列标注训练一个通用依赖关系解析器

除了生成基于 LLM 的注释之外，我们还使用序列标注方法训练了一个多语言依赖分析器。它可以作为 BiLingua 分析器任务的替代基线。我们使用 CoDeLin 框架，并采用两种编码策略：相对（REL）和绝对（ABS）微调 bert-base-multilingual-cased，按照 Roca et al. (2023) 来训练这个分析器。训练数据结合了 UD 英语 EWT (Silveira et al., 2014) 和西班牙语 AnCora (Taulé et al., 2008) 数据集。这些数据集被合并、打乱，并分为训练集、开发集和测试集。然后使用 CoDeLin 将数据编码为序列标签。分析器在训练时使用学习率 1e−5 、批次大小 64、权重衰减 0.001 和 Adam epsilon 1e−7 进行了 30 轮训练。我们用标准 CoNLL 2018 脚本 (Zeman et al., 2018) 将预测解码为 CoNLL-U 格式进行评估。这个分析器作为单语言和 CSW 环境下的解析性能的监督基准。

4.4 BiLingua 解析器到低资源语言的扩展

我们将 BiLingua Parser 扩展到那些没有可用的句法标注代码转换数据以训练监督解析器的低资源语言对。西班牙语-瓜拉尼语数据集被用作案例研究。受大型语言模型在低资源环境中可扩展性的启发，我们使用基于提示的 UD 标注结合母语者审查，以在不需要大规模标注语料的情况下引导句法资源。提示和架构细节见附录 C 。本数据集由于其长度和复杂性呈现出独特的挑战。尽管如此，解析器生成了有意义的标注，使对资源不足、类型学多样的语言的代码转换进行有价值的句法分析成为可能。

5 结果与语言分析

5.1 BiLingua 解析器的结果

表格 9 显示了 BiLingua Parser 在标注附加分数（LAS）指标上对代码切换数据集的性能对比。对于西班牙语-英语，这种基于 LLM 的标注在与金标注相比时达到了 76.32 % 的得分，且与人类审阅的输出相比时达到了 95.29 % 的得分，超越了此前报告的低于 75 % 的 LAS 分数的模型（见节 2 ）。除了西班牙语-英语的评估，我们还报告了西班牙语-瓜拉尼语数据集的结果，该数据集代表了首次尝试对涉及这种低资源语言的代码切换数据进行句法标注。在上述两种方法中，西班牙语-瓜拉尼语数据集分别达到了 59.90 % 和 77.42 % 的 LAS 分数。值得注意的是，通用多语言解析器 Universal Dependencies 西班牙语-英语模型（UDSL）在与金标注对比时仅达到了 14.71 % 的 LAS，突显了现成模型在应用于代码切换数据时的局限性。

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: 代码切换数据中专家审查前后的 LAS 比较。

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

Table 10: UPOS、DEPREL 和整体 LAS 性能在专家修订后的表现。

我们还进行了详细分析，对人工审核的输出结果进行了分析，以展示 BiLingua 解析器生成的 UD 标记的整体准确性。如表 10 所示，解析器在 UPOS、DEPREL 和 LAS 指标上均取得了高准确率（超过 90 % ），尤其是在西班牙语-英语数据集上。虽然西班牙语-瓜拉尼语数据集在依存关系解析上表现略低，但 UPOS 标记仍然保持强劲，对于低资源语言的语言资源开发来说，这是一个很好的结果。这些结果表明，大型语言模型能够在双语环境中稳健地进行句法分析，性能超越了既不是专为代码切换训练的混合模型也不是通用多语言解析器。

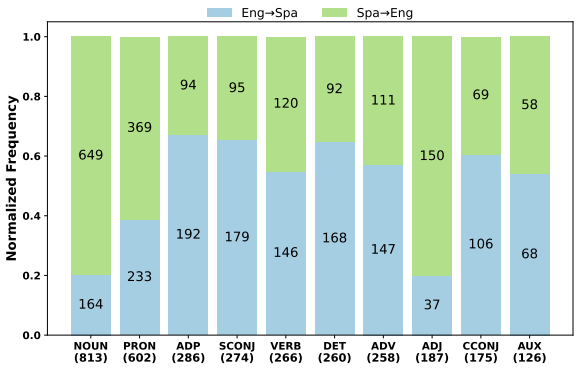
5.2 基于 LLM 的依存解析的定性错误分析

尽管提示提供了处理重复和省略的明确指南，但 LLM 的响应在应用这些规则时仍然不一致。有时将重复分析为并列关系，而在其他情况下，它为每个子句重复依存结构。虽然这两种分析在语言学上都是合理的，但这种不一致可能会影响关于语码转换点的句法泛化的可靠性。我们还观察到在功能动词的分析上存在不一致性，包括助动词、情态动词和西班牙语“ser”（“是”）等轻动词。基于 LLM 的标注在指定这些功能元素的句法角色为 `root` 和将它们附加到其他动词中心之间摇摆不定，显示出在处理动词依存结构上的变化。

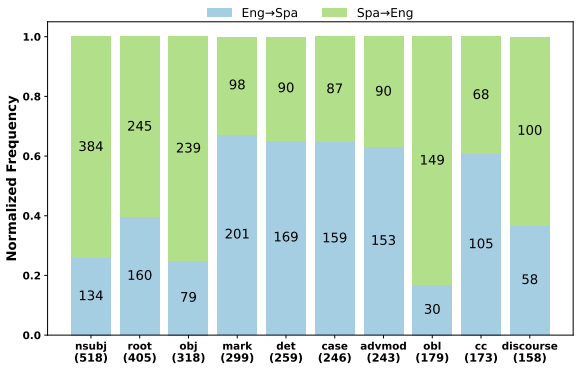
值得注意的是，母语者对瓜拉尼语数据的反馈经常会识别出形态复杂的词，这些词需要拆分为多个标记，以便进行准确的句法和词性标注。例如，标记 noñeguahëi（“没有来”）被建议拆分为一个否定副词和一个动词，每个都有其自己的词性标记。这一观察强调了在资源贫乏和黏着性语言中需要进行语言特定的形态预处理，并建议未来对 LLM 标注管道的改进可能会受益于整合形态分析器或针对这些语言的标记拆分机制。对具体错误进行调查的详细例子在附录 D 中提供。

5.3 英语-西班牙语数据集中 CSW 点的句法概括

此前对 NLP 中代码切换点的结构特征的研究主要集中在词性（POS）标记 (Martínez, 2020) 上，而我们的分析通过利用 UD 在切换位置捕捉句法角色来推进这项工作。图 3a 和 3b 显示了 UPOS 和 DEPREL 标签在两种切换方向上的归一化分布。我们发现主语位置（`nsubj` ）是最常见的切换位置之一，尤其是在英语到西班牙语的片段中。尽管文献中对此模式没有一致强调，但它与允许在主要句法边界切换的更广泛发现一致，包括小句初始位置。经典研究如 (Poplack, 1980; Myers-Scotton, 2002) 突出显示了在结构等价存在时，名词短语，尤其是限定词（`det` ）、修饰语（`amod` ）和介词（`case` ），作为常见切换位置。我们的结果支持这一点，显示在名词领域和小句边界（`mark` ，`cc` ，



(a) 在迈阿密西班牙语-英语数据集中，代码切换点的通用词类 (UPOS) 标签的归一化分布。值得注意的是，NOUN 、PRON 和 ADP 是切换点上最常见的类别，其中 NOUN 显示出从西班牙语切换到英语的高频率。



(b) Miami 西班牙语-英语数据集中代码转换点的通用依赖关系（DEPREL）的归一化分布。转换点最常见的关系是 nsubj 、root 和 obj ， 表明句法主语和核心动词论元是切换的关键位置。

Figure 3: 在代码转换点的句法类别（UPOS 和 DEPREL）的归一化频率分布中，包括两个转换方向。条形图显示了 Eng → Spa（蓝色）和 Spa → Eng（绿色）的比例。绝对计数显示在条形图内；总数在括号中。

discourse ）存在频繁的切换。nsubj 的突出可能反映了语言对特定的特性或话语模式，例如主题突出或左边离置。这些发现表明依赖关系揭示了词性标记单独未能捕捉到的细粒度切换模式 (Martínez, 2020) ，并激励需要在双语语料库中进行更丰富的句法标注。

在西班牙语-英语代码转换（CSW）文献中，主要动词或根谓词（即 UD 中的 root ）内的转换被认为是受到高度限制的。早期的研究如 Poplack (1980) 和像矩阵语言框架 (Myers-Scotton, 1993) 这样的模型认为，由于两种语言之间的形态句法不兼容，动词短语边界通常抗拒代码转换。基于语料库的研究 (例如， **?**) 证实，在或在主要动词内的转换是罕见的，双语者倾向于在从句边界进行转换。当转换确实发生在动词域内时，它们往往涉及语义透明的结构或频繁的双语模式。我们的研究结果表明，这一语言理论的限制需要重新审视。根层面的代码转换高频率可能部分反映了解析错误，如错误地将情态动词或助动词分析为根。我们承认这一局限性，并计划在未来的工作中通过手动验证或模型校准策略来解决这个问题。

5.4 西班牙语-瓜拉尼数据集中的 CSW 点语法概括

我们对西班牙语-瓜拉尼语代码转换的分析显示出比西班牙语-英语双语中通常观察到的更广泛的句法灵活性。如图 7 所示（见附录 E ），西班牙语-瓜拉尼语数据中的转换点不仅出现在典型的名词短语边界，比如主语（nsubj ）、宾语（obj ）和限定词（det ），还出现在子句内部位置，包括助动词、情态动词和根级动词。这些位置通常在其他语言对中更难以转换。相比之下，西班牙语-英语数据（图 3 ）表现出更受限制的转换模式，主要集中在名词界限和功能标记如 mark 和 case 上，动词核心显示出较低的易受影响性。瓜拉尼语对动词整合的相对开放性似乎允许更广泛的转换位置。进一步的分析，包括表情符号与非表情符号子集的分解和话语层面线索的作用，见附录 E 。

6 结论

这项工作介绍了 BiLingua Parser，这是一种利用大型语言模型（LLMs）进行代码转换数据句法注释的新型流程，该流程得到专家人工验证的支持。通过利用 GPT-4.1 和语言学知识驱动的提示，我们为西班牙语-英语和西班牙语-瓜拉尼语代码转换生成了高质量的 UD 注释。我们的结果表明，基于 LLM 的注释在句法准确性上优于传统解析器，特别是在单语模型通常失败的转换点上。这种性能差距在我们的第二种评估方法下尤为显著，该方法将 LLM 输出与人工修订的注释进行比较，并且不惩罚语言学上合理的变异。通过整合语义相似的依存标签组，该评估为多语言环境中的解析提供了更现实的基准。重要的是，我们发布了首个公开可用的西班牙语-英语和西班牙语-瓜拉尼语 CSW 的 UD 注释数据集，填补了多语言自然语言处理资源的关键空白。这些数据集和我们的注释方法不仅使代码转换行为的细粒度分析成为可能，还为推进低资源依存解析提供了基础。

7

局限性

UD 框架提供了一种跨语言一致性的句法标注方法，但其复杂性给不熟悉形式语言解析惯

例的标注者带来了挑战。没有这样的训练，标注质量可能会有所不同，与其他基于 UD 的数据集的比较可能不那么可靠。为了确保一致性和互操作性，我们强调为母语为瓜拉尼语的人提供详细的 UD 指导和动手标注练习的重要性。这将支持为低资源语言创造高质量、语言学基础的资源。

8

伦理考虑

我们的工作研究了使用大语言模型（LLMs）对代码转换语言数据进行句法标注，重点关注西班牙语-英语和西班牙语-瓜拉尼语。虽然这项研究有助于开发更具包容性和多语言的自然语言处理工具，但也带来了几项伦理考虑。基于 LLM 的句法标注的应用涉及到传播模型偏见和结构不准确的风险，尤其是在资源匮乏的语言环境中，标准句法标注稀缺。如果在没有人类监督的情况下将这种标注用于后续任务，就有可能导致对双语使用者及其语言实践的错误语言假设。LLM 在多语言环境中的简单或不受监督的应用可能会无意中强化主导语言结构或误解代码转换规范。在将这些工具应用于实际环境之前，应采取适当措施以确保其可靠性和语言专家的参与。我们使用人工智能助手（Grammarly 和 ChatGPT）来解决语法错误并改写句子。

9

致谢 我们感谢匿名的标注者和审稿人给予的建设性建议和帮助。我们感谢亚利桑那州立大学的研究计算中心（RC）和企业技术提供计算资源及访问 ChatGPT 企业版用于实验。我们承认 GAP (PID2022-139308OA-I00) 的资助由 MICIU/AEI/10.13039/501100011033 和 ERDF, EU 提供；LATCHING (PID2023-147129OB-C21) 的资助由 MICIU/AEI/10.13039/501100011033 和 ERDF, EU 提供。作为加利西亚大学系统卓越中心获得认证的 CITIC 是 CIGUS 网络的成员，并且从加利西亚大区的教育、科学、大学和职业培训部获得补贴。此外，它还获得由欧盟提供的联邦 2021-27 行动计划（Ref. ED431G 2023/01）的联合资助。

References

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.

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.

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. *ArXiv*, abs/1902.10505.

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

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.

azi Murat Özate, Özlem Çetinolu, Reut Tsarfaty, Dilek Küçük, and Olcay Taner Yldz. 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 标注指南

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: 在标注者培训期间提供的常见 UPOS 标签。

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: 向标注者介绍的关键 UD 依存关系。

我们在注释之前向来自巴拉圭的瓜拉尼语母语者提供了一个包含语言背景的 UD 标注方案概述。这包括 POS 标签和 DEPREL 标签的解释和例子。表格 11 和 12 列出了最相关标签的一个子集。对于西班牙语-英语的注释，具有语言背景的英语和西班牙语母语者被指导在注释期间使用在 https://universaldependencies.org/ 的官方通用依赖文档作为 POS 和 DEPREL 标签的参考。

B BiLingua 解析器的提示

下图展示了用于指导 BiLingua 解析器的完整提示设计。图 4 展示了依赖关系定义的参考表，与通用依赖 (UD) 约定对齐。图 5 展示了提供给模型的完整基础提示，具体说明了预期的标记级格式，以及对于处理代码转换输入中的特殊情况的定制指令。

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: 系统提示中提供给模型的依存关系参考表。定义遵循 UD 约定，包括核心句法关系、修饰语、功能词、句子级结构和与话语相关的依存关系。这些定义帮助限制模型的预测，使其在西班牙语-英语代码转换背景下符合句法有效的选项。

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: 用于指导 GPT 生成西班牙语-英语代码转换句子词汇级别 UD 标注的提示。提示概述了所需的输出格式，包括标准 UD 字段（ID、FORM、LEMMA、UPOS、HEAD ID、HEAD、DEPREL），并结合了针对代码转换特定现象的指引。这些现象包括处理英语缩略形式的规则（例如，把 didn't 拆分为 → did + n't），不流利和修复现象（例如，像 yo yo 这样的重复词），省略或不完整的结构，以及标点符号的附加。提示通过要求每个句子必须有一个根词和提供处理从句及语篇级依赖关系的规则来保证句子结构有效。一个完整格式化的例子说明了符合 UD 惯例的 GPT 期望输出的结构。

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: 用于指导 GPT 生成西班牙语-瓜拉尼语代码转换句子中词级 UD 注释的提示。该提示定义了所需的通用依存关系（UD）输出字段，ID、FORM、LEMMA、UPOS、HEAD ID、HEAD 和 DEPREL，并且通过要求每个句子恰好有一个句法根来强制结构有效性。说明明确地涉及如何附加额外的动词或从句（例如，使用 conj 或 parataxis 而不是第二个 root）以及如何使用 discourse 或 other 标签处理标点符号和非标准符号（例如表情符号或语篇小品词）。两个完全格式化的例子展示了这些习惯用法如何应用于混合语言句子，包括瓜拉尼语动词和西班牙语名词短语。该提示旨在处理多类型、低资源输入而无需预处理或形态切分。

C 西班牙语-瓜拉尼语数据集的架构和提示

在构建西班牙语-瓜拉尼语 UD 标注时，我们保留了来自源数据集 (Chiruzzo et al., 2023) 的原始标记和句子分割。模型没有被指示去拆分形态复杂的标记或简化数据。因此，许多句子超过了 50 个标记，并具有复杂的、富含从句的结构，这与平均约 5 个标记的较短的西班牙语-英语句子形成了对比。这为解析准确性带来了额外的挑战。为解决这些挑战，我们为西班牙语-瓜拉尼语的代码转换输入设计了特定的任务提示，如图 6 所示。该提示概述了预期的 UD 输出格式，包括处理依存结构有效性的针对性指令、话语元素的处理，以及类型学上多样的结构。它使得 LLM 能够在不需要预处理或形态分析的情况下生成结构良好的标注，这对于资源匮乏和形态丰富的语言环境是适用的。

D CSW 结果的扩展定性分析

表 13 显示了一个包含重复和省略的句法结构的例子。短语 “they’re high enough so that él no se…” 包含在两个重叠的分句中重复出现的主语-系动词结构（“they’re”）。LLM 对这些重复形式进行了不一致的分析，有时是平行连接，有时是重复附加词头。它也将西班牙语分句 “él no se...” 视为一个省略结构，而没有解析出最终的动词。这个例子展示了模型在管理话语层结构和保持长篇、代码切换的语句中的句法连贯性方面的挑战。表 14 显示了另一个经常出现的问题：对重复动词的不一致处理。在语句 “hay hay que dice o’clock somewhere” 中，动词 “hay”（‘there is’）出现了两次，这是自发讲话中的常见现象。虽然两个实例都是有效的，但 LLM 将第二个实例赋予了一个连词（conj）标签，而不是将其视为不流畅或根的重复。这在句法解释中创造了模糊性，表明需要制定指南或预处理策略来应对代码切换输入中的重复标记。

E 西班牙语-瓜拉尼语数据集中的基于表情符号的变化结果

为了进一步了解西班牙语-瓜拉尼语转码中的话语变化，我们将数据集分为两个子集：带有表情符号的信息和不带表情符号的信息。这种划分近似于形式和表达性上的差异，包含表情符号的子集代表更非正式或更具情感表达的交流。图 7 展示了两个子集在转码点的主要 UPOS 和 DEPREL 标签。在表情符号丰富的子集（图 7a 、7c ）中，转码频繁发生在话语敏感的句法角色，如 discourse 、parataxis 和与

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	–

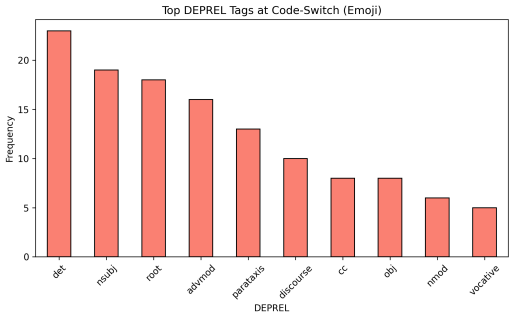
Table 13: “但我认为它们已经足够高了以至于 él no se...” 的依赖分析。高亮显示的行展示了重复的主语-系动词结构以及一个省略的状语从句。

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

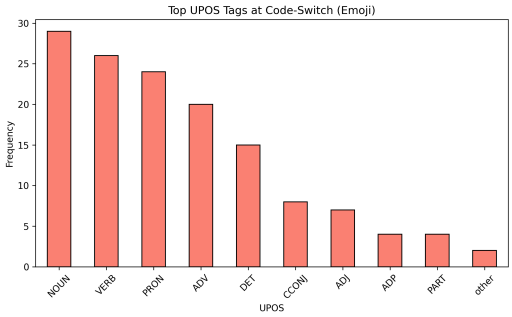
Table 14: 对 “hay hay que dice o’clock somewhere” 的简化 UD 分析。高亮行显示了 “hay” 动词处理不一致的问题。

立场相关的动词，除此之外还有像 det 、nsubj 和 root 这样的传统位置。这表明非正式的信息允许更多的句法灵活性，而语用上下文在转换位置中发挥重要作用。

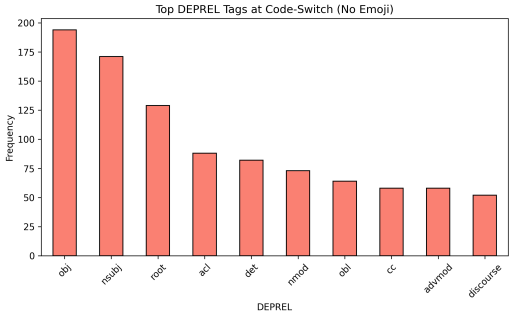
相反，非表情符号子集（图 7b , 7d ）显示了一个更稳定的转换模式，集中在如 obj , nsubj , acl 和 det 等规范名词位置，并且在子句级话语功能或动词中心发生转换的情况较少。这些结果一起支持了这样一个观察，即转换的结构模式并不是固定的，而是根据交流的上下文而变化。表情符号的使用似乎允许更大的句法流动性，特别是在话语级转换和双语讲话中语用标记段落中。



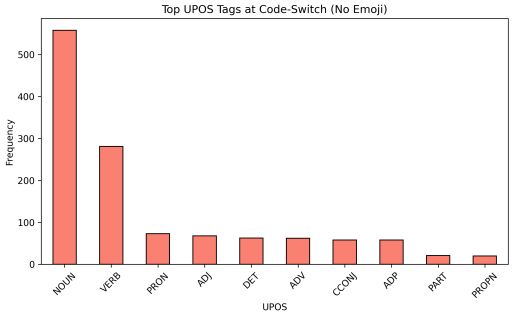
(a) DEPREL + Emoji



(c) UPOS + 表情符号



(b) DEPREL –Emoji



(d) upos——表情符号

Figure 7: 在西班牙语-瓜拉尼语句子中的代码转换点，比较包含表情符号和不包含表情符号子集的 DEPREL 和 UPOS 标签的分布。(+ Emoji) 指的是包含表情符号的数据集子集，(- Emoji) 指的是不含表情符号的子集。