

Modeling Topics and Sociolinguistic Variation in Code-Switched Discourse: Insights from Spanish-English and Spanish-Guaraní

Nemika Tyagi^{1*}, Nelvin Licona Guevara¹, Olga Kellert¹

¹Arizona State University, USA

{ntyagi8, nliconag, olga.kellert}@asu.edu

Abstract

This study presents an LLM-assisted annotation pipeline for the sociolinguistic and topical analysis of bilingual discourse in two typologically distinct contexts: Spanish-English and Spanish-Guaraní. Using large language models, we automatically labeled topic, genre, and discourse-pragmatic functions across a total of 3,691 code-switched sentences, integrated demographic metadata from the *Miami Bilingual Corpus*, and enriched the *Spanish-Guaraní* dataset with new topic annotations. The resulting distributions reveal systematic links between gender, language dominance, and discourse function in the Miami data, and a clear diglossic division between formal Guaraní and informal Spanish in Paraguayan texts. These findings replicate and extend earlier interactional and sociolinguistic observations with corpus-scale quantitative evidence. The study demonstrates that large language models can reliably recover interpretable sociolinguistic patterns traditionally accessible only through manual annotation, advancing computational methods for cross-linguistic and low-resource bilingual research.

Keywords: LLM-assisted annotation, code-switching, topic analysis

1. Introduction

Code-switching is the alternation between two or more languages within a single discourse and it is a pervasive feature of bilingual and multilingual communication. It plays a central role in how speakers negotiate identity, stance, and alignment in everyday interaction. Far from being random, language switching is systematically tied to discourse and pragmatic functions. Early interactional work (Gumperz 1982; Auer 1998) demonstrated that switches can index conversational frames, signal quotations or reported speech, and mark topic or participant shifts. Subsequent studies expanded this view by showing that switching contributes to narrative organization, topic management, and information flow (Bullock and Toribio 2009; Fricke and Kootstra 2016; Kootstra and Fricke 2020; Gardner-Chloros 2009). Pragmatic and cognitive approaches have also emphasized that code-switching reflects fine-grained speaker control over audience design, style, and footing (Gafaranga 2011; Myers-Scotton 1993; Matras 2009). Yet, these rich qualitative insights rely on manual annotation and small datasets, limiting their scalability and cross-linguistic coverage.

Computational linguists have sought to automate the detection and structural modeling of code-switching. Early NLP studies focused on identifying switch points or predicting the matrix language using surface and syntactic features (Solorio and Liu 2008a; Solorio and Liu 2008b; Jamatia et al. 2015; Molina et al. 2016). More recent work employs neural architectures to improve token-level language

identification and sequence tagging (Winata et al. 2018; Bhat et al. 2023), and shared tasks have promoted benchmark datasets across language pairs (Molina et al. 2016; Patwa et al. 2020). However, these methods primarily model *where* switches occur rather than *why*. Few computational studies explicitly represent the discourse, pragmatic, or topical motivations underlying switching (Liu et al. 2021; Zirn et al. 2023). Consequently, quantitative generalizations about the communicative functions of switching remain scarce. Sociolinguistic variation adds further complexity. Classic studies link code-switching to speaker demographics, community norms, and social meaning (Poplack 1980; Romaine 1995; Toribio 2004; Gumperz 1982). Yet, most available corpora lack demographic annotation, hindering systematic sociolinguistic analysis. Low-resource and endangered languages are especially underrepresented (Joshi et al. 2020; Ponti et al. 2020), creating a skewed empirical base that privileges shallow studies on the nature of code-switching for only high-resource bilingual pairs such as English-Spanish or Mandarin-English.

Large language models (LLMs) now present an opportunity to bridge linguistic and computational perspectives. Their ability to model contextual and cross-lingual semantics allows for new forms of discourse-level analysis (Ruder et al. 2019; Kirk et al. 2023). However, even the most advanced multilingual models still struggle with mixed-language input (Zhang et al. 2023; Potter and Yuan 2024; Cahyawijaya et al. 2021). Such errors arise because multilingual models are trained primarily on monolingual or parallel text, leaving conversational and dynamically switched data understudied.

*Main and Corresponding Author.

This study addresses this gap by exploring how topic modelling, pragmatic function, and speaker attributes shape bilingual discourse. We integrate LLM-based topic modeling with discourse-pragmatic annotation to examine *why* switches occur in addition to *where*. By focusing on both Spanish-English and Spanish-Guaraní, we extend typological and regional coverage beyond high-resource language pairs and include underrepresented bilingual contexts. This approach unites humanistic and computational perspectives, providing scalable yet interpretable analyses of bilingual communication.

Our main contributions are as follows:

- Propose a GPT-based topic classification pipeline with interpretable, linguistically grounded category refinement.
- Integrate sociolinguistic metadata (gender, age, language dominance) with bilingual datasets to enable demographic and cross-linguistic comparison.
- Release enhanced bilingual corpora annotated for topics and discourse functions, with visualization of switch and topic distributions¹.
- Provide comparative insights into discourse and social variables across high- and low-resource bilingual contexts.

In doing so, our study moves toward a more inclusive and linguistically grounded understanding of bilingual communication, highlighting how computational methods can capture the social and discourse complexity of multilingual speakers.

2. Background and Related Work

Discourse and Pragmatic Perspectives Foundational interactional studies positioned code-switching as a discourse-organizational and identity-constructing practice, linking language alternation to topic shifts, quotations, and stance (Gumperz 1982; Auer 1998; Bullock and Toribio 2009). Later pragmatic and cognitive accounts further showed that switches contribute to affective alignment, audience design, and narrative coherence (Myers-Scotton 1993; Matras 2009; Fricke and Kootstra 2016). Despite these insights, such analyses relied on manual coding, limiting systematic quantification of discourse functions across corpora. Sociolinguistic research emphasizes that code-switching varies with gender, age, and social networks (Poplack 1980; Romaine 1995; Toribio

2004; Sankoff 1998). However, most publicly available corpora lack demographic metadata and remain skewed toward high-resource bilingual pairs (Joshi et al. 2020; Ponti et al. 2020). Low-resource languages are particularly underrepresented, restricting comparative and inclusive analysis.

Computational Modeling of Code-Switching

Early computational work focused on token-level language identification and matrix language prediction (Solorio and Liu 2008a; Molina et al. 2016; Patwa et al. 2020). Neural architectures have since improved language tagging accuracy (Winata et al. 2018; Bhat et al. 2023), yet most systems do not incorporate pragmatic or topical cues. A few recent studies have begun exploring discourse motivations and contextual embeddings (Liu et al. 2021; Zirn et al. 2023), but these approaches are still limited in scope and coverage. Large language models (LLMs) offer new possibilities for discourse-level multilingual analysis (Ruder et al. 2019; Kirk et al. 2023). Nonetheless, empirical evaluations show that multilingual models often misinterpret mixed-language input, treating code-switching as noise rather than strategic discourse choice (Zhang et al. 2023; Potter and Yuan 2024; Cahyawijaya et al. 2021). These findings suggest the need for linguistically grounded evaluation resources that capture both structural and sociolinguistic aspects of bilingual communication. Our work integrates these strands by combining LLM-based topic modeling with discourse-pragmatic annotation and sociolinguistic metadata, linking functional interpretations of switching with scalable computational methods.

3. Experiments

3.1. Datasets

Miami Corpus (English-Spanish). The Miami corpus (Deuchar et al. (2014)) is part of the BilingBank repository and was collected between 2008 and 2011 by the ESRC Centre for Research on Bilingualism, Bangor University. It consists of transcripts of informal conversations among 84 bilingual speakers in Miami, USA. Recordings capture spontaneous interactions, later transcribed and pseudonymized. Participants provided demographic information through post-recording questionnaires, enabling sociolinguistic analysis. For the present study, we extract the subset of 2,825 sentences containing intra-sentential code-switching between English and Spanish.

GUASPA Corpus (Spanish-Guaraní). The Guarani dataset originates from the GUASPA shared task at IberLEF 2023 (Chiruzzo et al. (2023)), which contains Guarani-Spanish mixed

¹Data is available at <https://github.com/N3mika/topicmodelling>.

| Corpus | Sentences | Tokens | Avg token/sent. |
|---------|---|--------|-----------------|
| Miami | 2825 | 29.7k | 10.5 |
| Spa-Gua | 866 | 15.6k | 18.0 |
| Corpus | Lang. proportion (%) | | |
| Miami | spa 48.4; eng 40.1; punc 9.4; eng&spa 2 | | |
| Spa-Gua | spa 42.6; gn 38.7; other 16.6; gn&spa 2.1 | | |

Table 1: Summary of code-switched subsets and token-level language proportions. Language tags: *spa* = Spanish, *eng* = English, *gn* = Guaraní, *punc* = punctuation, *other* = punctuation, special characters, or emojis, *eng&spa/gn&spa* = ambiguous mixed-language tokens.

texts sourced from Paraguayan tweets and news articles. The full corpus comprises 1,500 texts and about 25k tokens. We use the training-set portion and further extract a subset containing sentences with intra-sentential switching, represented at the token level. All Spanish variants and named-entity tags were merged under *spa*, while foreign and unclassified tokens such as punctuations and emojis were grouped as *other*.

Subset Statistics. Table 1 summarizes the statistics of the two code-switched datasets and their corresponding subsets used in this study, including their token counts, language proportions, and average code-switching density measured as adjacent language changes per sentence.

3.2. Experimental Setup

The annotated datasets were created by inferring the `gpt-4.1-2025-04-14` model through the OpenAI API. Each sentence, together with speaker and situational metadata, was processed individually using deterministic parameters ($\text{temp}=0$, $\text{max_tokens}=200$). The pipeline constructed structured prompts containing sentence ID, language tag, and contextual information, and normalized the model’s outputs to canonical topic and function labels. Annotation was conducted in batches of 50-100 sentences, covering 2,825 code-switched sentences from the Miami corpus and 866 from the Spanish-Guaraní dataset (see Table 1).

4. Methods

4.1. Category Selection

We began by sampling 30 random sentences from each dataset and collaboratively developed an initial categorization schema with two bilingual annotators. For the Miami corpus, two dimensions were defined: *Topics* (domain or content areas) and

Functions (discourse or pragmatic roles). For the Spanish–Guaraní dataset, three dimensions were introduced: *Formality*, *Genre*, and *Topic*. After initial annotation, the schemas were iteratively refined by reviewing an additional batch of 30 sentences to ensure coverage of the linguistic and contextual diversity of each corpus. Semantically overlapping categories were merged, and concise explanatory notes were added to clarify the scope of each category for future annotation consistency. All categories were cross-validated by both annotators before pipeline implementation. While these taxonomies were created through careful linguistic reasoning, we acknowledge that topic and discourse categorization is inherently subjective; different annotators or frameworks could yield alternative but equally valid interpretations. The goal was to establish a practical and interpretable schema for enhancing bilingual corpora and enabling downstream sociolinguistic analysis.

Below are the details of the multi-tiered Annotation Schemas that we developed for the Miami and SPA-GUA corpus.

Miami Corpus Annotation Schema

| Functions (choose one primary; secondary allowed if clearly two) |
|--|
| TechnicalTermInsertion: inserting domain-specific words or tool names. |
| ProperNounNamedEntity: naming a person, place, brand, or award. |
| PrecisionLexicalGap: switching for precise expression or lexical need. |
| DiscourseMarker: connective or organizing signals (e.g., <i>you know, so</i>). |
| TopicShift: marking a new topic or returning to one. |
| Narrative: embedding a story or recounting a past event. |
| Quotation: reproducing or stylizing another’s voice. |
| TurnManagement: backchannels or acknowledgments (<i>mmhm, yeah</i>). |
| AddresseeShift: calling attention or changing addressee (<i>hey Bob</i>). |
| Directive: giving orders, requests, or imperatives. |
| Repair: rephrasing, searching for a word, or self-correcting. |
| Agreement: affirming or echoing another speaker’s stance. |
| StanceEmphasis: expressing evaluation, certainty, or irony. |
| Humor: jokes, teasing, or playful language. |
| SolidarityIdentity: in-group markers or swearing showing closeness. |

Topics (choose one; if mixed, choose the dominant)

Workplace_Technical: technical terms, commissioning, CAD, architecture terms.

Education_YouthOrganizations: school, certificates, scouts, permission slips.

Architecture_Design: materials, styles, famous architects.

Office_Logistics: supplies, scheduling, file paths, emails.

Narratives_Quotations: recounting past events or reported speech.

Casual_EverydayTalk: greetings, jokes, small talk, banter.

Affect_Identity: swearing, nicknames, identity/solidarity markers.

ProperNouns_NamedEntities: sentences dominated by names, places, or awards.

Topics (choose one; if two, mark a secondary)

Government_Announcement: official statements from institutions.

Legislation_Policy: mentions laws, regulations, or legislative actions.

Protest_Report: reports describing protests or demonstrations.

Mobilization_Call: calls for strikes, activism.

Corruption_Donations_Procurement: references of such.

PublicAdministration_Changes: appointments or administrative shifts.

Procurement_Licitation: references to tenders or contract awards.

Infrastructure_Contract: mentions construction or development projects.

Transport_PublicSafety: transportation or safety-related content.

Agriculture_Reactivation: farming or agrarian reform.

Rural_Community_Issues: rural life or community concerns.

Indigenous_CommunityAid: Indigenous rights or aid programs.

Education_Policy_University: education reforms or student activism.

Cultural_Event_Festival: festivals or public celebrations.

Cultural_Heritage_Archive: heritage preservation or archives.

Media_Broadcast_Notice: broadcast or program announcements.

Legal_Judicial: courts, rulings, or judicial.

Crime_Investigation: mentions crimes or investigations.

Health_COVID: COVID-19, vaccines, or health effects.

PublicHealth_Services: hospitals or medical access.

Environment_NationalParks: conservation or protected areas.

Commercial_Product: product promotions or corporate content.

Shopping_PersonalPurchase: consumer life or buying habits.

Personal_Emotional: emotional reflections or personal states.

Humor_Rant: jokes, sarcasm, or venting.

Sports_Event: matches, scores, or athletes.

Entertainment_Music_Film: mentions music, artists, or movies.

Opinion_Commentary: subjective political or social commentary.

UserMention_Request_Response: direct replies, mentions, or user interactions.

Other: unclear or uncategorizable tweets.

Spanish-Guaraní Corpus Annotation Schema

Formality (choose one)

Formal: official or institutional tone; objective or procedural (e.g., announcements, reports, press releases).

Informal: conversational, personal, humorous, or emotional tone; includes slang, emojis, or direct address.

Genre (choose one)

News: objective reports or summaries of events.

Personal: emotions, reflections, or personal experiences.

Politics: mentions politicians, elections, or government affairs.

Activism_Protest: references to mobilizations or calls to action.

Culture_Arts: music, literature, art.

Education: covers schools, universities, or reforms.

Health: health, medicine, or COVID-19.

Environment: ecology, nature, conservation.

Sports: athletic events or teams.

Entertainment: celebs, humor, pop culture.

Commercial: ads, business, or products.

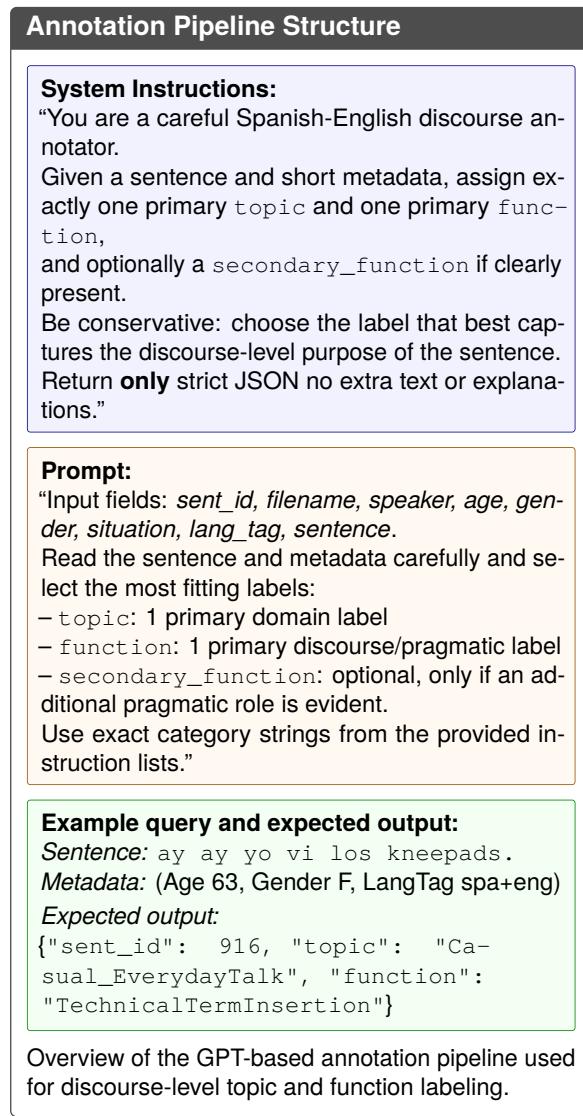
Announcement: schedules, program info.

Opinion: commentary or evaluation of public issues.

Other: fallback for unclear categories.

4.2. Topic Annotation Workflow

To operationalize the annotation schemas, we designed a structured prompting workflow for automatic labeling using the `gpt-4.1` model. Each prompt simulated the detailed instructions a human annotator would follow and consisted of three coordinated components: (1) a *system prompt* that defined the annotator's role ("You are a careful Spanish-English discourse annotator") and constrained the model to output only a single JSON object; (2) a *base prompt* that described the input structure and labeling procedure, specifying that each sentence accompanied by speaker and situational metadata such as `speaker`, `age`, `gender`, `situation`, and `lang_tag` must be tagged with exactly one `topic` and one `function`, and an optional `secondary_function`; and (3) a set of *instruction lists* enumerating the available topic and function labels defined for each dataset. A schematic representation of this workflow is shown below, summarizing the process from data input and prompt construction to annotation and post-processing.



Few-shot exemplars were appended to the base prompt to illustrate correct labeling behavior for short conversational and code-switched sentences, ensuring consistent adherence to the pragmatic scope of the task. The model was instructed to be conservative, choosing the single label that best captured the topical or discourse-level purpose of each utterance. Each response was required to follow strict formatting, which facilitated automated parsing and normalization.

4.3. Evaluation Metrics

To evaluate annotation quality, 30 randomly selected sentences from each annotated dataset were reviewed by bilingual linguists. Each verifier assessed the plausibility and fit of the assigned labels. The Miami corpus achieved 100% accuracy for `topic` and `function` labels, and 60% accuracy for `secondary_function`. The Spanish-Guaraní corpus obtained 94.17% accuracy across its four fields (*Formality*, *Genre*, *Topic*, and *Secondary_Topic*). These results indicate high reliability of the GPT-based labeling pipeline for both corpora, with minor variation in secondary or ambiguous cases.

5. Results and Discussion

5.1. Miami Corpus

5.1.1. Sociolinguistic Topic Modelling

Table 2, 3 presents the gender-normalized topic and function distributions for the Miami corpus. The topic distribution (Table 2) shows the relative proportions of topics across male and female speakers, while the function distribution (Table 3) highlights pragmatic differences across genders. Both male and female speakers predominantly engage in "*Casual_EverydayTalk*" (60%), reflecting the conversational and informal character of the recordings. Narrative and quotation contexts are the next most frequent, followed by work-related and technical discussions. Minor yet notable differences emerge: female speakers contribute proportionally more to "*Office_Logistics*" and "*Education_YouthOrganizations*", while male speakers show slightly higher proportions of named-entity references. Pragmatically, both genders favor "*Narrative*" functions, though women exhibit marginally higher rates of directive and solidarity-related functions, consistent with earlier sociolinguistic observations that female bilinguals employ switching for interpersonal alignment (cf. Poplack 1980; Romaine 1995). The current large-scale annotation provides quantitative support for these qualitative observations across nearly 3,000 bilingual utterances.

| Topic | Men (%) | Women (%) | Tot. (n) |
|------------------------------|---------|-----------|----------|
| Casual_EverydayTalk | 59.8 | 60.1 | 1694 |
| Narratives_Quotations | 20.5 | 18.5 | 536 |
| Workplace_Technical | 4.8 | 4.8 | 135 |
| Office_Logistics | 1.6 | 5.7 | 130 |
| ProperNouns_NamedEntities | 7.0 | 3.7 | 130 |
| Education_YouthOrganizations | 2.5 | 3.4 | 90 |
| Affect_Identity | 2.8 | 3.3 | 89 |
| Architecture_Design | 1.1 | 0.5 | 18 |
| Total Sentences | 757 | 2065 | 2822 |

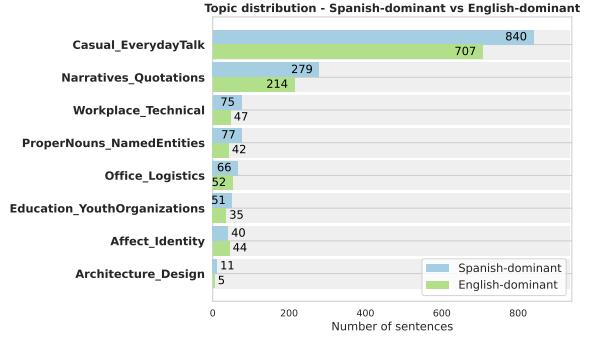
Table 2: Topic distribution by gender (normalized by gender totals) in the Miami corpus. Percentages are normalized within each gender.

| Function | Men (%) | Women (%) | Total (n) |
|------------------------|---------|-----------|-----------|
| PrecisionLexicalGap | 24.3 | 28.1 | 765 |
| Narrative | 19.6 | 19.8 | 556 |
| DiscourseMarker | 12.0 | 12.4 | 348 |
| TechnicalTermInsertion | 10.6 | 10.5 | 296 |
| StanceEmphasis | 6.9 | 6.1 | 178 |
| ProperNounNamedEntity | 8.5 | 4.5 | 156 |
| Directive | 4.0 | 6.0 | 153 |
| SolidarityIdentity | 2.4 | 3.5 | 90 |
| Repair | 3.4 | 2.3 | 73 |
| Quotation | 3.3 | 2.2 | 70 |
| TurnManagement | 2.4 | 1.6 | 52 |
| Agreement | 1.3 | 1.1 | 33 |
| AddresseeShift | 0.7 | 1.2 | 30 |
| Humor | 0.7 | 0.2 | 10 |
| TopicShift | 0.1 | 0.4 | 10 |
| UNKNOWN_FUNCTION | 0.0 | 0.1 | 2 |
| Total Sentences | 757 | 2065 | 2822 |

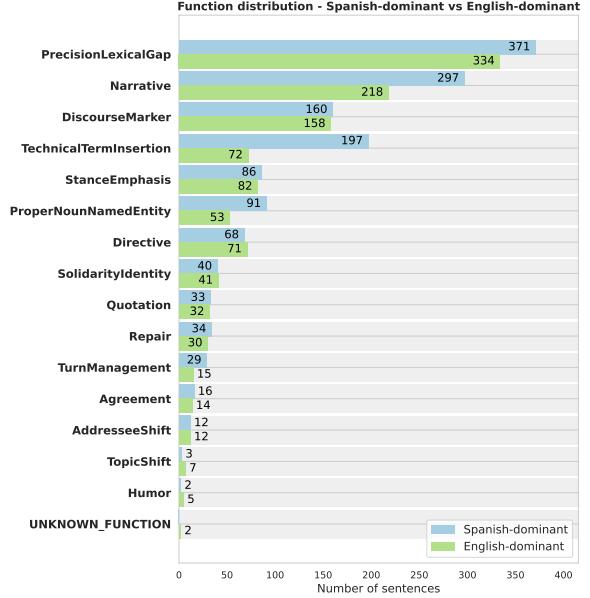
Table 3: Function distribution by gender (normalized by gender totals) in the Miami corpus. Percentages are normalized within each gender.

5.1.2. Bilingual Asymmetries

Figure 1 illustrates bilingual asymmetry patterns in the Miami corpus. Figure 1a compares topic distributions across Spanish- and English-dominant sentences, while Figure 1b presents the corresponding discourse functions. Spanish-dominant segments cluster around casual, affective, and narrative domains, reflecting conversational and stance-oriented uses. English-dominant spans show slightly higher frequencies in technical and precision-related categories, whereas both languages are similarly represented in discourse-marker functions. These tendencies align with interactional analyses suggesting that Spanish indexes personal stance and social proximity, while English supports informational precision and referential clarity (Bullock and Toribio 2009; Toribio 2004). Such distributional asymmetries may also be influenced by lexical-level factors that shape bilingual sentence planning. Further evidence for this comes from Fricke and Kootstra (2016), who found that bilinguals often code-switch in response to lexical triggers such as cognates or shared lexical items. In this view, code-switching operates as a strategy to access otherwise unavailable expressions or to optimize lexical retrieval. Lexical



(a) Topic distribution across Spanish- and English-dominant sentences.



(b) Function distribution across Spanish- and English-dominant sentences.

Figure 1: Bilingual asymmetries in the Miami corpus, showing variation in topic and function distributions between Spanish- and English-dominant contexts.

overlap, especially in the non-default or less dominant language, heightens the likelihood of switching, reflecting how bilinguals dynamically manage cross-linguistic resources based on contextual and lexical accessibility.

5.2. Spanish-Guaraní Corpus

5.2.1. Formality-driven topic and genre modelling

We first inspect how formality (Formal vs Informal) conditions topical and genre distributions in the Spanish-Guaraní dataset. Table 4 and Table 5 present the same statistics shown as proportions normalized by formality totals (i.e., each formality column is normalized), where categories have been trimmed to the most frequent items (topics:

| Genre | Form.(%) | Inform. (%) | Total (n) |
|-----------------|----------|-------------|-----------|
| News | 65.1 | 0.3 | 320 |
| Personal | 0.0 | 72.1 | 271 |
| Politics | 12.9 | 0.3 | 64 |
| Announcement | 12.2 | 0.3 | 61 |
| Opinion | 1.2 | 11.7 | 50 |
| Culture_Arts | 5.7 | 2.9 | 39 |
| Entertainment | 0.0 | 5.9 | 22 |
| Sports | 0.0 | 2.7 | 10 |
| Others | 3.8 | 7.2 | 29 |
| Total Sentences | 490 | 376 | 866 |

Table 4: Genre formality split (aggregated). Values for all categories after top 8 are summed into the “Others” row. Percent columns are proportions normalized by formality totals.

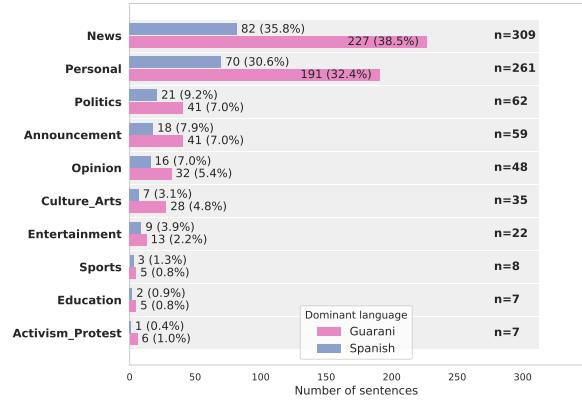
| Topic | For. (%) | Inf. (%) | Tot. (n) |
|----------------------------------|----------|----------|----------|
| UserMention_Request_Response | 0.0 | 30.6 | 115 |
| Humor_Rant | 0.0 | 21.8 | 82 |
| Personal_Emotion | 0.0 | 19.9 | 75 |
| Government_Announcement | 14.7 | 0.0 | 72 |
| Opinion_Commentary | 4.3 | 9.6 | 57 |
| Cultural_Event_Festival | 9.8 | 1.6 | 54 |
| PublicAdministration_Changes | 9.6 | 0.3 | 48 |
| Legislation_Policy | 7.8 | 0.3 | 39 |
| Corruption_Donations_Procurement | 4.7 | 1.3 | 28 |
| Education_Policy_University | 4.1 | 1.1 | 24 |
| Protest_Report | 4.1 | 0.5 | 22 |
| Sports_Event | 1.0 | 4.0 | 20 |
| Crime_Investigation | 3.5 | 0.8 | 20 |
| Transport_PublicSafety | 3.7 | 0.3 | 19 |
| Legal_Judicial | 3.9 | 0.0 | 19 |
| Indigenous_CommunityAid | 3.5 | 0.3 | 18 |
| PublicHealth_Services | 3.3 | 0.3 | 17 |
| Infrastructure_Contract | 3.5 | 0.0 | 17 |
| Cultural_Heritage_Archive | 2.9 | 0.8 | 17 |
| Others | 15.8 | 6.7 | 103 |
| Total Sentences | 490 | 376 | 866 |

Table 5: Topic formality split (aggregated). All categories with 15 or less total sentences are combined under “Others.” Percent columns are proportions normalized by formality totals.

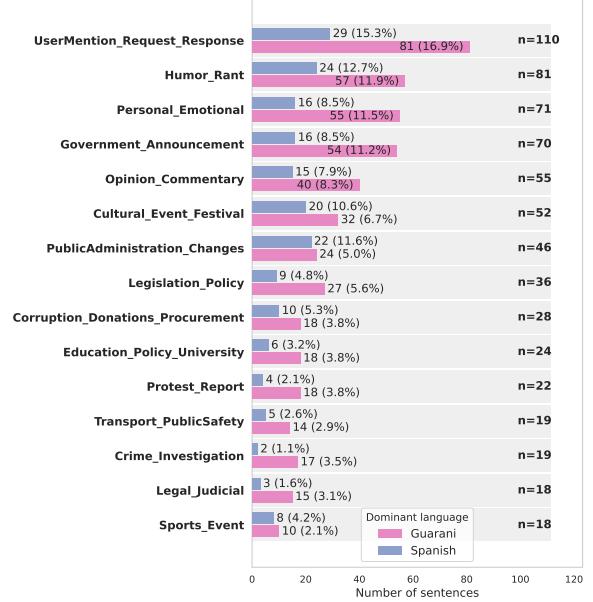
top 15; genres: top 10). These plots highlight register differences (formal institutional uses vs. informal/personal use) across categories. As shown in the above tables, the corpus contains nearly equal proportions of formal and informal sentences; however, their distribution varies substantially by topic. This variation provides a useful basis for examining linguistic phenomena such as code-switching within distinct communicative contexts. Moreover, the observed balance between formal and informal registers appears to correlate with language dominance, suggesting that register and language choice are interdependent dimensions of bilingual discourse.

5.2.2. Language-dominance asymmetries

Next, we compare category counts across dominant-language splits (Guaraní-dominant vs Spanish-dominant). Figure 2b and Figure 2a present category counts with two grouped bars per category (one bar for each dominant-



(a) Genre distribution broken down by dominant language (Guaraní-dominant vs Spanish-dominant). Top 10 genres shown.



(b) Topic distribution broken down by dominant language (Guaraní-dominant vs Spanish-dominant). Top 15 topics shown.

Figure 2: Language-dominance comparisons for topics and genres in the Spanish-Guaraní dataset. Each row displays two bars (Guaraní-dominant and Spanish-dominant counts); category lists are trimmed to the most frequent items for clarity.

language class). These plots reveal which topics and genres are more commonly associated with Guarani- vs Spanish-dominant sentences. Guarani-dominant texts are concentrated in “Government_Announcement”, “PublicAdministration_Changes”, and “Indigenous_CommunityAid”, reflecting the language’s institutional and communal authority. Spanish-dominant texts emphasize “Personal_Emotion”, “Humor_Rant”, and “UserMention_Request_Response”, highlighting interpersonal and expressive use. The resulting division between formal Guarani and informal Spanish sup-

ports long-standing observations of diglossic role distribution in Paraguay (Rubin 1968; Gynan 2001; Zajícová 2019) but now emerges from corpus-scale quantitative evidence.

6. Discussions

This study demonstrates that LLMs can serve as practical tools for enriching bilingual corpora with topic and sociolinguistic annotations. By combining GPT-based inference with interpretable category schemas, we achieved high annotation accuracy across both high- and low-resource code-switched datasets. Beyond methodological validation, the analysis revealed distinct community-level dynamics: Miami speakers showed sharper gender- and language-dominant contrasts, whereas the Spanish-Guaraní community displayed complementary language use across registers, with Guarani favored in formal discourse and Spanish in informal, affective interaction.

Methodological and Resource Implications.

The proposed pipeline provides a scalable and interpretable approach for semi-automatic corpus enrichment. By leveraging structured instructions and few-shot examples, the method reduces annotation cost while maintaining linguistic transparency. The resulting topic- and function-annotated corpora expand the empirical base for analyzing bilingual discourse and improve the accessibility of training resources for multilingual NLP, particularly in underrepresented languages. Future refinements could integrate adaptive prompt optimization to assess and mitigate potential contextual biases in LLM-generated labels.

Sociolinguistic and Theoretical Insights.

The integration of sociolinguistic metadata with discourse-pragmatic annotation provides a quantitative basis for examining how language choice reflects social and cognitive constraints. Gender- and dominance-based asymmetries in the Miami data, alongside register-based differentiation in Spanish-Guaraní, align with sociolinguistic theories of stance, identity, and alignment (Poplack 1980; Toribio 2004). These findings illustrate that code-switching functions as a socially strategic resource rather than linguistic noise. At the lexical level, our results also resonate with psycholinguistic evidence that lexical accessibility and cognate activation influence switch likelihood (Fricke and Kootstra 2016). Future modeling could operationalize these mechanisms by incorporating measures of lexical overlap and semantic similarity, thereby linking discourse-level switching patterns with cognitive processes of bilingual word retrieval.

Advancing Topic Modeling for Bilingual Data.

Current annotation schemes rely on fixed topic taxonomies that promote comparability but may constrain discovery. Dynamic topic generation by LLMs offers a way to capture fluid, context-dependent thematic structure in spontaneous discourse. Because topics in natural interaction frequently overlap, such as politics intersecting with education or identity, proto-ablation-based and embedding-based representations may better capture semantic proximity and topic coherence (Bianchi et al. 2021). Incorporating these representations can disambiguate semantically adjacent categories and yield richer, more flexible bilingual topic models.

Toward Integrated Multilevel Modeling. The next step is to integrate the annotation of discourse and pragmatics with syntactic, semantic, and sociolinguistic analyzes to construct a multilevel model of bilingual production. Linking topic and function annotations with grammatical and dependency structures will allow researchers to trace how discourse roles interact with structural switch points and grammatical constraints. Such integration can connect item-level accessibility with discourse-level planning, offering a cognitively informed and computationally tractable account of how bilingual speakers manage alignment and coherence across typologically distinct languages. Ultimately, this approach points toward a unified framework for modeling social bilingual discourse across a variety of multilingual communities.

7. Conclusion

This paper introduced an LLM-assisted annotation framework for topic and sociolinguistic labeling of bilingual corpora, evaluated on Spanish-English and Spanish-Guaraní datasets. The method achieved high annotation reliability while maintaining interpretability, enabling scalable enrichment of code-switched data. The resulting resources and framework contribute to expanding the empirical and typological scope of bilingual discourse research, particularly for underrepresented language pairs. Beyond the specific corpora released, this work demonstrates how LLMs can serve as practical instruments for linguistically grounded resource development. Future research will extend this framework to integrate syntactic and affective layers, such as dependency relations, sentiment, and stance, toward building comprehensive, socio-computational models of code-switching.

8. Acknowledgements

This work was supported by the School of International Letters and Cultures (SILC) at Arizona State

University. Special thanks go to the Universidad Nacional de Asunción and especially to Eliodora Verón and Ricardo Peloso for their annotation reviews and contributions. We also thank the organizers of IberLEF 2023 for access to the GUA-SPA dataset.

9. Ethical Considerations

All data used are publicly available or anonymized (Bangor Miami Corpus, IberLEF GUA-SPA). No personally identifiable information was processed. We acknowledge potential cultural bias in LLM outputs and have incorporated manual verification to ensure representational fairness, especially for Indigenous languages. All examples respect community norms and licensing agreements.

10. Bibliographical References

- Peter Auer. 1998. *Code-Switching in Conversation: Language, Interaction and Identity*. Routledge, London.
- Isha Bhat et al. 2023. Understanding multilingual code-switching via large language models. In *EACL*.
- Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.
- Barbara E. Bullock and Almeida Jacqueline Toribio. 2009. Themes in the study of code-switching. In Barbara E. Bullock and Almeida Jacqueline Toribio, editors, *The Cambridge Handbook of Linguistic Code-Switching*, pages 1–17. Cambridge University Press, Cambridge.
- Samuel Cahyawijaya et al. 2021. Cross-lingual and multilingual automatic speech recognition: Challenges and opportunities. In *Findings of ACL*.
- Melinda Fricke and Gerrit Jan Kootstra. 2016. Primed codeswitching in spontaneous bilingual dialogue. *Journal of Memory and Language*, 91:181–201.
- Joseph Gafaranga. 2011. Transition space medium repair: Language shift talked into being. *Journal of Pragmatics*, 43(1):118–135.
- Penelope Gardner-Chloros. 2009. *Code-switching*. Cambridge University Press.
- John J. Gumperz. 1982. *Discourse Strategies*. Cambridge University Press, Cambridge.
- Shaw N. Gynan. 2001. Language planning and policy in paraguay. *Current Issues in Language Planning*, 2(1):53–118.
- Anupam Jamatia, Amitava Das, Björn Gambäck, and Shibamouli Ghosh. 2015. Part-of-speech tagging for code-mixed english–hindi–bengali social media text. In *Proceedings of RANLP 2015*, page 239–248.
- Pratik Joshi et al. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *ACL*.
- Hannah Kirk et al. 2023. Beyond translation: Cross-lingual abilities of large language models. *arXiv preprint arXiv:2304.09106*.
- Gerrit Jan Kootstra and Melinda Fricke. 2020. Interactive alignment and code-switching in bilingual dialogue. *International Journal of Bilingualism*, 24(6):1200–1216.
- Wei Liu et al. 2021. Analyzing the discourse functions of code-switching with neural models. In *EMNLP*.
- Yaron Matras. 2009. *Language Contact*. Cambridge University Press.
- Gustavo Molina et al. 2016. Overview for the second shared task on language identification in code-switched data. In *EMNLP Workshop on Code-Switching*.
- Carol Myers-Scotton. 1993. *Duelling Languages: Grammatical Structure in Code-Switching*. Clarendon Press, Oxford.
- Parth Patwa et al. 2020. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of SemEval*.
- Edoardo Maria Ponti et al. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *EMNLP*.
- 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.
- Thomas Potter and Zhen Yuan. 2024. Llm-based code-switched text generation for grammatical error correction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16957–16965. Association for Computational Linguistics.

- Suzanne Romaine. 1995. *Bilingualism*, 2nd edition. Blackwell, Oxford.
- Joan Rubin. 1968. National bilingualism in paraguay. *Southwestern Journal of Anthropology*, 24(2):153–168.
- Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.
- David Sankoff. 1998. A quantitative and sociolinguistic model for code-switching. In Wei Li, editor, *The Bilingualism Reader*, pages 75–89. Routledge.
- Thamar Solorio and Yang Liu. 2008a. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 973–981.
- Thamar Solorio and Yang Liu. 2008b. Part-of-speech tagging for english–spanish code-switched text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 1051–1060. Association for Computational Linguistics.
- Almeida Jacqueline Toribio. 2004. Convergence as an optimization strategy in bilingual speech: Evidence from code-switching. *Bilingualism: Language and Cognition*, 7(2):165–173.
- Genta Indra Winata et al. 2018. Code-switching language modeling using syntax-aware multi-task learning. In *ACL*.
- Lenka Zajícová. 2019. Diglossia in paraguay revisited: New trends in spanish–guaraní bilingualism. *Journal of Multilingual and Multicultural Development*, 40(4):325–340.
- R. Zhang, S. Cahyawijaya, J. C. B. Cruz, G. I. Winata, and A. F. Aji. 2023. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582. Association for Computational Linguistics.
- Cristoph Zirn et al. 2023. Modeling the pragmatic motivations for code-switching in multilingual nlp. In *ACL*.
- Rodríguez, Yliana and Góngora, Santiago and Solorio, Thamar. 2023. *GUASPA: Guaraní-Spanish Code-Switching Corpus (IberLEF 2023)*. IberLEF 2023 Shared Task on Guaraní-Spanish Code-Switching Analysis.
- Deuchar, Margaret and Davies, Peredur and Herring, Jon and Parafita Couto, M. Carmen and Carter, Diana. 2014. *Bangor Miami Corpus of Spanish-English Bilingual Speech*. ESRC Centre for Research on Bilingualism, Bangor University. PID <https://bangortalk.org.uk>.

11. Language Resource References

Chiruzzo, Luis and Agüero-Torales, Marvin and Giménez-Lugo, Gustavo and Alvarez, Aldo and