Modeling Cultural and Subcultural Variation in Code-Switched Discourse with Topic Annotation

Nemika Tyagi[*] **Nelvin Licona Guevara** **Olga Kellert**
Arizona State University, USA
{ntyagi8, nliconag, olga.kellert}@asu.edu

Abstract

Code-switching is often modeled in NLP as a structural or token-level phenomenon, overlooking its role as a discourse practice shaped by social and cultural context. In this work, we propose topic-based annotation as a framework for analyzing cultural and subcultural variation in bilingual discourse. Using large language models, we annotate 3,691 code-switched sentences from Spanish-English (Miami) and Spanish-Guaraní (Paraguay) corpora with topic and discourse-level information, integrating sociolinguistic metadata. Our analysis reveals systematic relationships between discourse topics, language choice, and social variables such as gender and language dominance. We observe subcultural variation within the Miami community and a clear diglossic distribution in Paraguay, where Guaraní is associated with formal domains and Spanish with informal communication. These findings suggest that modeling code-switching through discourse-level categories provides a more complete representation of multilingual communication and enables both cross-cultural and intra-cultural comparison at scale[1].

1 Introduction

Language use is fundamentally shaped by cultural and social context. Beyond linguistic form, what speakers choose to talk about, how they structure discourse, and which meanings they prioritize are influenced by shared norms, values, and communicative practices. Code-switching, the alternation between two or more languages within a single discourse, is not only a structural linguistic phenomenon but also a culturally structured discourse practice. Early interactional work (Gumperz 1982) demonstrated that switching is not random, but indexes conversational frames, quotation, and topic shifts. Subsequent research showed that code-switching contributes to narrative organization and information flow (Bullock and Toribio 2009), while pragmatic and sociolinguistic studies link it to identity, stance, and audience design (Gafaranga 2011; Myers-Scotton 1993). From this perspective, code-switching reflects how speakers organize meaning across social contexts, topics, and identities, rather than merely alternating between linguistic systems. Recent work in cultural and cross-cultural NLP highlights that language technologies must account for differences in communication, knowledge, and goals, while recognizing that culture cannot be reduced to fixed proxies such as language or nationality (Hershcovich et al. 2022; Zhou et al. 2025). Together, these perspectives point to a central challenge: cultural patterns are not always directly observable in text, but are also embedded in discourse.

In NLP, existing approaches to code-switching have largely treated it as a structural phenomenon, focusing on identifying switch points or predicting the matrix language (Solorio and Liu 2008; Jamatia et al. 2015). While effective for token-level tasks, these methods primarily capture *where* switches occur rather than *why*, overlooking its role as a culturally embedded discourse practice. In contrast, sociolinguistic approaches treat language as a reflection of social structure and interaction, showing that code-switching varies systematically with speaker attributes such as gender, age, and community norms (Poplack 1980; Toribio 2004). However, most available corpora lack demographic and contextual metadata, and low-resource bilingual settings remain underrepresented (Joshi et al. 2020), limiting large-scale analysis of culturally grounded language use. This motivates our focus on code-switched subsets of corpora, where discourse-level variation is most clearly expressed through language alternation.

Addressing these gaps, we argue that cultural analysis in NLP requires structured, discourse-

[*]Main and Corresponding Author.

[1]Data is available at https://github.com/N3mika/topicmodelling.

Proceedings of the 4th Workshop on Cross-Cultural Considerations in NLP (C3NLP 2026), pages 40–49

based comparative frameworks. Topic-based annotation provides such a framework by identifying recurring themes in language use and enabling comparison across contexts (Hou and Huang 2025; Egger and Yu 2022). Topics serve as a shared analytical space in which such patterns become observable. Crucially, culture is not equivalent to language: linguistic communities contain multiple subcultures shaped by social variables such as gender, age, and language dominance (Nguyen et al. 2016). These subcultures may diverge in topic preferences and discourse practices, making sociolinguistic metadata essential for interpreting discourse variation.

To operationalize this framework at scale, we use large language models as topic modeling and annotation tools. We apply this approach to code-switched subsets of bilingual corpora from two sociolinguistic contexts: Spanish-English discourse in Miami and Spanish-Guaraní discourse in Paraguay. By combining topic annotations with social variables such as gender and language dominance, we examine both cross-cultural and subcultural variation. Our results suggest that topic distributions reflect culturally structured discourse patterns, including diglossic language use in Paraguay and gendered differences in Miami. These findings indicate that code-switching is closely linked to discourse organization and that modeling it through discourse-level categories can provide a more complete representation of multilingual communication.

This paper makes four contributions to cross-cultural NLP research:

- It proposes topic-based annotation as a comparative framework for analyzing cultural and subcultural variation in discourse.
- It develops an LLM-assisted annotation pipeline with interpretable, linguistically grounded category design, enabling scalable analysis of discourse topics and functions.
- It integrates sociolinguistic metadata with bilingual corpora to support both cross-cultural and intra-cultural comparison, and releases enhanced datasets annotated for topics and discourse functions.
- It provides empirical evidence that sociolinguistic variables and language use systematically shape discourse patterns across high- and low-resource bilingual contexts.

In doing so, our study advances a more culturally grounded approach to NLP, showing how discourse-level analysis can reveal the social and cultural structures underlying language use and inform the development of models that better reflect the diversity of multilingual communication.

2 Background and Related Work

Culture, Discourse, and Sociolinguistic Variation Recent work in NLP has increasingly emphasized the role of culture in shaping language data, model behavior, and evaluation, while also highlighting the difficulty of defining and operationalizing cultural knowledge in computational settings (Hershcovich et al. 2022; Liu et al. 2025). Interactional and sociolinguistic research has long treated code-switching as a discourse-organizing and identity-constructing practice, linking language alternation to topic management, quotation, stance, and audience design (Gumperz 1982; Bullock and Toribio 2009; Myers-Scotton 1993). Research in computational sociolinguistics emphasizes that such variation reflects underlying social structures and subcultures (Nguyen et al. 2016; Blodgett et al. 2016; Kellert and Matlis 2022). However, most findings rely on manually annotated datasets, and large-scale corpora with rich sociolinguistic metadata remain scarce, particularly for low-resource languages (Joshi et al. 2020; Ponti et al. 2020). This limits the ability of NLP systems to model language as a culturally and socially embedded phenomenon.

Discourse Modeling and Annotation in NLP Computational approaches to code-switching have primarily focused on token-level language identification and matrix-language prediction (Solorio and Liu 2008; Patwa et al. 2020), with neural models improving performance in mixed-language settings (Winata et al. 2018). While recent work has begun to explore discourse-level representations and the use of large language models for multilingual analysis (Ruder et al. 2019; Kellert et al. 2025), these models still struggle with mixed-language input (Zhang et al. 2023; Potter and Yuan 2024). In parallel, topic modeling has been widely used in social sciences to identify recurring themes in discourse (Egger and Yu 2022). Our work builds on these directions by treating topic annotation as a form of cultural abstraction, enabling systematic comparison of discourse across cultures and subcultures at scale.

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.

3 Experiments

3.1 Datasets and Sociolinguistic Contexts

Miami Corpus (English-Spanish). We use the Miami corpus, a collection of transcribed informal conversations among 84 bilingual speakers in Miami, USA, with accompanying demographic metadata. From this corpus, we extract 2,825 sentences containing intra-sentential English-Spanish code-switching for sociolinguistic analysis. This corpus provides a setting for analyzing subcultural variation within a single bilingual community, given its rich sociolinguistic metadata.

GUA-SPA Corpus (Spanish-Guaraní). We use the Spanish-Guaraní dataset from the GUA-SPA shared task, consisting of mixed-language tweets and news texts from Paraguay. From the training portion (25k tokens), we extract 866 sentences containing intra-sentential code-switching. Spanish variants and named entities are merged under *spa*, while punctuation, emojis, and other non-classified tokens are grouped as *other*. This dataset reflects a distinct sociolinguistic context characterized by diglossic language use, enabling comparison of discourse patterns across sociolinguistic contexts.

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

We use LLMs (`gpt-4.1-2025-04-14` model) as annotation tools to assign topic and discourse labels to each code-switched sentence, enabling scalable analysis of discourse patterns. Each sentence, together with speaker and situational metadata, was processed individually using deterministic parameters (`temp`=0, `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 sentences from the Miami corpus and 866 from the Spanish-Guaraní dataset (see Table 1).

4 Methods

4.1 Category Selection

To operationalize cultural and discourse variation in code-switched data, we developed corpus-specific annotation schemas through iterative manual review. Thirty randomly sampled sentences from each dataset were examined by two bilingual annotators to define initial categories. For the Miami corpus, we annotated *Topics* (content domains) and *Functions* (discourse-pragmatic roles), while for the Spanish-Guaraní dataset we defined *Formality*, *Genre*, and *Topic* dimensions to reflect its mixed conversational and institutional sources. The schemas were refined to ensure coverage and consistency, merging overlapping categories and clarifying label definitions, and were cross-validated prior to large-scale annotation. Because the corpora differ in communicative ecology, we developed topic inventories inductively within each dataset rather than enforcing a unified taxonomy. This design treats topic categories as context-sensitive abstractions of discourse, preserving culturally specific patterns while enabling comparison across communities. The resulting schemas provide an interpretable framework for analyzing how discourse topics and sociolinguistic variables interact in code-switched contexts, supporting both cross-cultural and subcultural analysis.

Below are the examples of multi-tiered annotation schemas for the Miami and Spanish-Guaraní corpora.

Miami Corpus: Functions

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., *you know, so*).
TopicShift: marking new topic/returning to one.
Narrative: embedding a story or recounting a past event.
Quotation: reproducing or stylizing another's voice.
TurnManagement: backchannels or acknowledgments (*mmhm, yeah*).
AddresseeShift: calling attention or changing addressee (*hey Bob*).
Directive: giving orders/requests/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.

Miami Corpus: Topics

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.

Spanish-Guaraní Corpus: Formality

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.

Spanish-Guaraní Corpus: Genre

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.

Spanish-Guaraní Corpus: Topics

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 preserva-

tion or archives.
Media_Broadcast_Notice: broadcast or program announcements.
Legal_Judicial: courts, rulings, or judicial matters.
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.

4.2 Topic Annotation Workflow

To scale the annotation of discourse-level categories, we designed a structured prompting workflow that uses an LLM as a proxy annotator. The workflow mirrors the instructions a human annotator would follow, ensuring that topic and function labels are assigned consistently across datasets. Each prompt consisted of three coordinated components: (1) a *system prompt* that defined the annotator's role 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 (*speaker*, *age*, *gender*, *situation*, *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 labels for each dataset. Few-shot exemplars were appended to the base prompt, illustrating correct labeling behavior for conversational and code-switched sentences, promoting consistency in discourse-level annotation. The model was instructed to assign a single label that best captured the primary topic and function of each utterance. Responses were required to follow strict JSON formatting, enabling automated parsing and normalization. This design ensures consistent and scalable annotation of discourse categories across sociolinguistic contexts. A schematic of this workflow:

Annotation Pipeline Structure

System Instructions:
"You are a careful Spanish-English discourse annotator. Given a sentence and short metadata, assign exactly one primary `topic` *and one primary* `function`*, and optionally a* `secondary_function` *if clearly present. Return* ***only*** *strict JSON — no extra text or explanations."*

Base Prompt:
"Input fields: `sent_id`, `filename`, `speaker`, `age`, `gender`, `situation`, `lang_tag`, `sentence`. *Select:* `topic` *(1 primary label),* `function` *(1 primary label),* `secondary_function` *(optional). Use exact category strings from the instruction lists."*

Example query and expected output:
Sentence: `ay ay yo vi los kneepads.`
Metadata: (Age 63, Gender F, LangTag spa+eng)
Output:
```
{"sent_id": 916, "topic":
"Casual_EverydayTalk", "function":
"TechnicalTermInsertion"}
```

4.3 Evaluation Metrics

To evaluate annotation quality, 100 randomly selected sentences from each dataset were reviewed by two bilingual linguists. They assessed the plausibility and fit of the assigned labels, validated them through rigorous linguistic and content-based analyses, and confirmed them through unanimous perceptual consensus (interrater reliability $> 90\%$).The Miami corpus achieved 97.8% accuracy for *topic*, *function*, and *secondary_function* labels, while 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 LLM-based annotation process, with minor variation in secondary or ambiguous cases.

5 Results

5.1 Miami Corpus

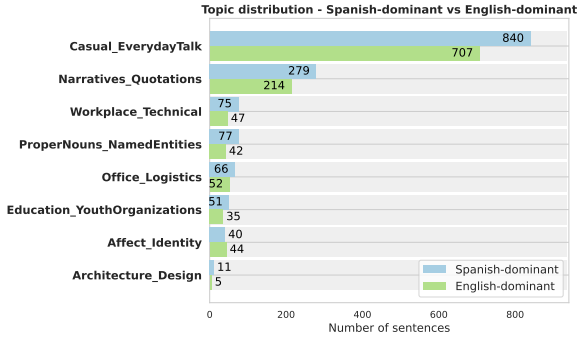
Subcultural Variation in Topic Distributions Table 2, 3 presents the gender-normalized topic and function distributions for the Miami corpus. The topic distribution (Table 2) shows the relative proportions of topics across male and female speakers, while the function distribution (Table 3) high-

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

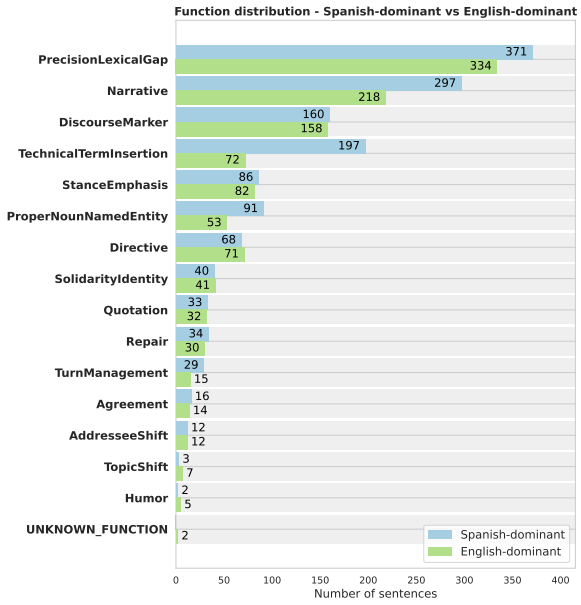
Table 2: Topic distribution by gender (normalized by gender totals) in the Miami corpus. Percentages are normalized within each gender.

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.



(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: topic and function distributions by dominant language. Sentences with equal Spanish and English token counts are excluded; totals may not sum to overall counts.

lights pragmatic differences across genders. Both male and female speakers predominantly engage in *"Casual_EverydayTalk"* (60%), reflecting the conversational nature of the data, with narrative and quotation contexts as the next most frequent categories. Notable differences emerge: female speakers contribute more to *"Office_Logistics"* and *"Education_YouthOrganizations"*, while male speakers show slightly higher proportions of named-entity references. Pragmatically, both groups favor *"Narrative"* functions, though women exhibit higher rates of directive and solidarity-related functions (Poplack 1980). These patterns reflect subcultural variation within the same bilingual community, where gender is associated with differences in discourse orientation and interactional style. The large-scale annotation thus makes these subcultural patterns observable in topic distributions across nearly 3,000 bilingual utterances.

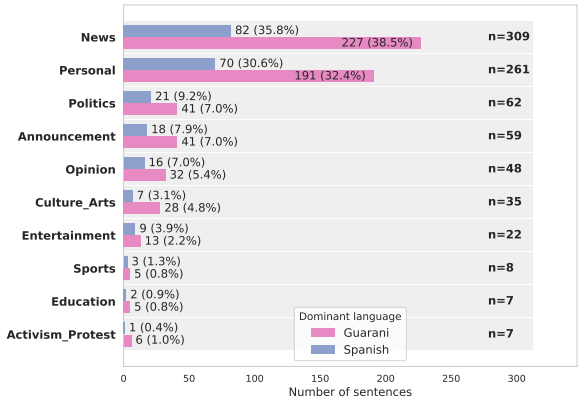
Bilingual Asymmetries Figure 1 illustrates bilingual asymmetry patterns in the Miami corpus. Figure 1a compares topic distributions across Spanish- and English-dominant code-switched sentences, while Figure 1b presents the corresponding discourse functions. Spanish-dominant segments cluster around casual, affective, and narrative domains, while English-dominant spans show higher frequencies in technical and precision-related categories. 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 (Bullock and Toribio 2009; Toribio 2004). Overall, the results indicate that language choice is associated with culturally meaningful discourse domains, with Spanish used more for interpersonal topics and English for technical and referential functions.

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

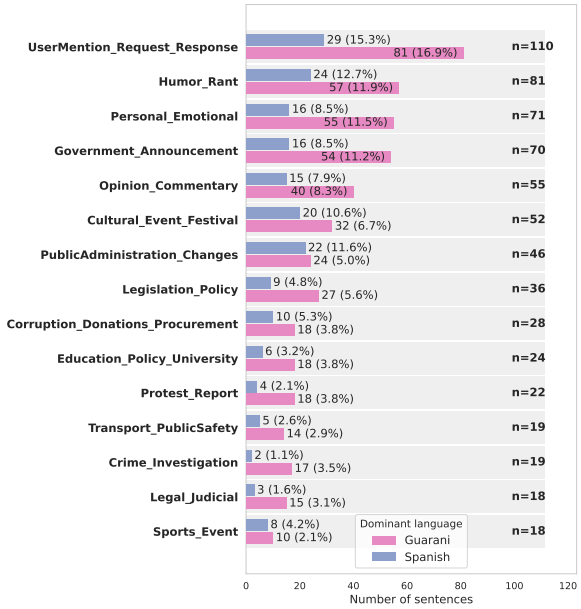
Table 4: Genre formality split (aggregated). Values for all categories after top 8 are summed into the "Others" row. Percent columns are proportions normalized by formality totals.

Topic	For. (%)	Inf. (%)	Tot. (n)
UserMention_Request_Response	0.0	30.6	115
Humor_Rant	0.0	21.8	82
Personal_Emotional	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.



(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 shows Guaraní- and Spanish-dominant counts; sentences with equal token counts are excluded, and categories are trimmed to the most frequent items for clarity.

5.2 Spanish-Guaraní Corpus

Formality-driven topic and genre modelling We examine how formality (Formal vs. Informal) shapes topic and genre distributions in the Spanish-Guaraní dataset. Tables 4 and 5 report proportions normalized within each formality class, restricted to the most frequent categories (top 15 topics; top 10 genres). Although the corpus contains nearly balanced formal and informal sentences, their distributions differ across topics. Formal texts cluster around institutional and policy-related domains, whereas informal texts concentrate in personal and expressive categories. This distribution reflects culturally structured discourse domains, where language use aligns with distinctions between institutional and interpersonal communication.

Language-dominance asymmetries Next, we compare category counts across dominant-language splits (Guaraní-dominant vs Spanish-dominant) in the code-switched sentences. Figure 2b and Figure 2a present category counts for each dominant-language class. These plots reveal which topics and genres are more commonly associated with Guaraní- vs Spanish-dominant sentences. Guaraní-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_Emotional", *"Humor_Rant"*, and *"UserMention_Request_Response"*, highlighting interpersonal and expressive use. The resulting division between formal Guaraní and informal Spanish supports long-standing observations of diglossic role distribution in Paraguay (Rubin 1962) but now emerges from corpus-scale quantitative evidence, showing that language choice systematically aligns with culturally defined discourse domains.

6 Discussions

This study shows that topic-based annotation enables scalable analysis of cultural and subcultural variation in bilingual discourse. By combining discourse-level categories with sociolinguistic metadata, we identify systematic differences in how communities organize discourse. The results reveal distinct cultural dynamics, including subcultural variation in Miami and complementary language use across discourse domains in Spanish-Guaraní.

Methodological and Resource Implications. The proposed framework provides a scalable and interpretable approach to enriching bilingual corpora with discourse-level annotations. By leveraging LLMs as proxy annotators, the method reduces annotation cost while maintaining consistency across datasets. The resulting topic-annotated resources expand the empirical basis for studying culturally embedded language use and support the development of more representative multilingual NLP systems, particularly for low-resource settings.

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 a purely structural phenomenon. At the lexical level, our results also resonate with psycholinguistic evidence that lexical accessibility and cognate activation influence switch likelihood (Wintner et al. 2023). 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.

Topic Annotation as Cultural Abstraction. Our results suggest that topic annotation can serve as a form of cultural abstraction, capturing recurring discourse patterns that reflect underlying social and cultural structures. Unlike fixed taxonomies, topic categories derived from discourse allow for context-sensitive representation of language use across communities. While traditional topic modeling approaches emphasize comparability, they may constrain discovery in spontaneous discourse. In contrast, more flexible representations can better capture overlapping and context-dependent thematic structure, where topics such as politics, identity, and education intersect (Bianchi et al. 2021). This supports more nuanced and comparative analysis of multilingual data, where cultural patterns are expressed through differences in discourse organization rather than solely through surface form.

Toward Multilevel Modeling. Future work can extend this framework by integrating discourse-level annotations with syntactic and semantic representations, enabling multilevel models of bilingual language use. Such approaches would allow researchers to link discourse patterns with structural constraints, providing a more comprehensive account of how cultural and linguistic factors interact in multilingual communication.

7 Conclusion

We presented a topic-based annotation framework for analyzing cultural and subcultural variation in code-switched bilingual discourse, applied to Spanish-English and Spanish-Guaraní data. By combining discourse-level topic annotation with sociolinguistic metadata, we showed how cultural patterns become observable through differences in discourse organization across and within communities. Our results indicate that topic distributions capture culturally meaningful structures, including subcultural variation in Miami and diglossic language use in Paraguay. More broadly, this work highlights the value of discourse-level representations for modeling code-switching as a culturally embedded practice. By using LLMs as annotation tools, we enable scalable and interpretable analysis of multilingual data while preserving sensitivity to social and cultural context. These findings point toward more culturally grounded NLP approaches

that move beyond surface-level language modeling to better capture the diversity of discourse practices in multilingual communities.

Limitations

The annotation schemas rely on predefined category inventories, which may constrain the representation of fluid or overlapping discourse phenomena. The study focuses on two bilingual contexts, and can be extended to typologically diverse language pairs. Future work could extend this framework by incorporating syntactic and affective layers, such as dependency relations, sentiment, and stance, toward developing integrated socio-computational models of code-switching. Future research should examine the transferability of this initial framework to other language pairs or domains and determine whether unsupervised topic modeling yields comparable patterns.

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.

References

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.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Barbara E. Bullock and Almeida Jacqueline Toribio. 2009. *Themes in the study of code-switching*, page 1–18. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Roman Egger and Joanne Yu. 2022. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in Sociology*, Volume 7 - 2022.

Joseph Gafaranga. 2011. Transition space medium repair: Language shift talked into being. *Journal of Pragmatics*, 43(1):118–135.

John J. Gumperz. 1982. *Discourse Strategies*. Studies in Interactional Sociolinguistics. Cambridge University Press.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Yuxin Hou and Junming Huang. 2025. Natural language processing for social science research: A comprehensive review. *Chinese Journal of Sociology*, 11(1):121–157.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2015. Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages. In *Proceedings of the international conference recent advances in natural language processing*, pages 239–248.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Olga Kellert and Nicholas Hill Matlis. 2022. Social context and user profiles of linguistic variation on a micro scale. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 14–19, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Olga Kellert, Nemika Tyagi, Muhammad Imran, Nelvin Licona-Guevara, and Carlos Gómez-Rodríguez. 2025. Parsing the switch: Llm-based ud annotation for complex code-switched and low-resource languages. *arXiv preprint arXiv:2506.07274*.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Carol Myers-Scotton. 1993. *Social Motivations For Codeswitching: Evidence from Africa*. Oxford University Press.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online). International Committee for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espaÑol: toward a typology of code-switching1. *Linguistics*, 18(7-8):581–618.

Tom Potter and Zheng 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.

Joan Rubin. 1962. Bilingualism in paraguay. *Anthropological Linguistics*, 4(1):52–58.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981.

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, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Code-switching language modeling using syntax-aware multi-task learning. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 62–67, Melbourne, Australia. Association for Computational Linguistics.

Shuly Wintner, Safaa Shehadi, Yuli Zeira, Doreen Osmelak, and Yuval Nov. 2023. Shared lexical items as triggers of code switching. *Transactions of the Association for Computational Linguistics*, 11:1471–1484.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri 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, Singapore. Association for Computational Linguistics.

Naitian Zhou, David Bamman, and Isaac L Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural nlp. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25869–25886.