Where and How Do Languages Mix? A Study of Spanish-Guaraní Code-Switching in Paraguay

Olga Kellert[1*] **Nemika Tyagi**[1]

[1]Arizona State University
{olga.kellert, ntyagi8}@asu.edu

Abstract

Code-switching, the alternating use of multiple languages within a single utterance, is a widespread linguistic phenomenon that poses unique challenges for both sociolinguistic analysis and Natural Language Processing (NLP). While prior research has explored code-switching from either a syntactic or geographic perspective, few studies have integrated both aspects, particularly for underexplored language pairs like Spanish-Guaraní. In this paper, we analyze Spanish-Guaraní code-switching using a dataset of geotagged tweets from Asunción, Paraguay, collected from 2017 to 2021. We employ a differential distribution method to map the geographic distribution of code-switching across urban zones and analyze its syntactic positioning within sentences. Our findings reveal distinct spatial patterns, with Guaraní-dominant tweets concentrated in the western and southwestern areas, while Spanish-only tweets are more prevalent in central and eastern regions. Syntactic analysis shows that code-switching occurs most frequently in the middle of sentences, often involving verbs, pronouns, and adjectives. These results provide new insights into the interaction between linguistic, social, and geographic factors in bilingual communication. Our study contributes to both sociolinguistic research and NLP applications, offering a framework for analyzing mixed-language data in digital communication.

1 Introduction

Code-switching is the seamless alternation between languages within a single utterance. This phenomenon has long attracted linguists because it reveals how bilingual speakers manage diverse linguistic and social demands. It not only reflects language proficiency but also illustrates how speakers navigate their complex cultural identities and communication contexts. Over the past decades, researchers have explored the interplay of linguistic, cognitive, and social factors underlying code-switching, often relying on elicited data and structured interviews (Corvalán, 2005; Kallfell, 2011; Dietrich, 2002, 2010; Auer and Eastman, 2010; Bullock and Toribio, 2009; Myers-Scotton, 2002; Poplack, 1980, 1985). However, such methods frequently fall short in capturing the natural spontaneity and fluidity of everyday speech.

The advent of digital communication, particularly through social media, has opened new avenues for real-time code-switching studies. Platforms like Twitter provide detailed time stamps and geographic data that enable precise mapping of language use (Eleta and Golbeck, 2014; Grieve et al., 2019; Kellert, 2023c,a,b, 2022). Recent computational approaches have leveraged these datasets to enhance NLP applications, including machine translation and sentiment analysis in mixed-language contexts (Cerón-Guzmán and León-Guzmán, 2016; Guzmán et al., 2017; Agüero-Torales et al., 2021; Rijhwani et al., 2017). Despite these advances, most studies treat geographic localization and syntactic analysis as separate issues.

In Paraguay, where Spanish and Guaraní are both official languages and integral to daily communication, a critical research gap exists. Few studies have combined detailed geographic mapping with deep syntactic analysis of code-switching, particularly for less-studied language pairs. The unique blend of indigenous and non-indigenous language practices in Paraguay's digital spaces remains largely unexplored (Agüero-Torales et al., 2023; Jauhiainen et al., 2023; Muñoz-Ortiz and Vilares, 2023; Fricke and Kootstra, 2016; Kootstra et al., 2020) . Our work addresses this gap by using social media data to provide a comprehensive analysis of code-switching in digital communication in Paraguay. We introduce a novel method that employs precise GPS coordinates from social media to gen-

[*]First Author, OK collected the data, designed the methodology, performed the analysis, and wrote the substantial part of the article

Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching, pages 26–31
May 3, 2025 ©2025 Association for Computational Linguistics

erate detailed maps of code-switching patterns in Greater Asunción, and we conduct an in-depth syntactic analysis to determine the common positions and parts of speech involved in language switches. Understanding these patterns has broader implications: it can inform language policy, enhance NLP applications such as machine translation and sentiment analysis for mixed-language texts, and contribute to the preservation of indigenous languages like Guaraní. Our findings, compared with established patterns in other language pairs, highlight the unique sociolinguistic dynamics in Paraguay and pave the way for future research in the processing of digital communication of low-resource languages like Guaraní.

2 Related Works

The study of code-switching has evolved significantly from early qualitative approaches to more sophisticated, data-driven analyses. Foundational research using elicited speech and interviews offered critical insights into the social and cognitive dimensions of bilingual language use (Corvalán, 2005); (Kallfell, 2011)), yet these methods often struggled to capture the fluidity of spontaneous communication in the digital space. The recent proliferation of social media as a data source has enabled researchers to overcome these limitations by analyzing naturally occurring, high-resolution datasets. Studies leveraging content from social media platforms have advanced our understanding of code-switching in mixed-language environments by addressing key NLP challenges such as machine translation, sentiment analysis, and language modeling (Cerón-Guzmán and León-Guzmán, 2016; Guzmán et al., 2017; Agüero-Torales et al., 2021; Rijhwani et al., 2017). However, most previous work has examined geographic and syntactic aspects separately, missing the chance to explore their interaction. Moreover, while many studies have focused on well-known language pairs such as English-Spanish, Paraguay's unique bilingual environment—where Spanish and Guaraní intermingle—remains largely unexplored. Our study fills this gap by combining detailed geographic mapping with an in-depth syntactic analysis of code-switching, offering fresh insights into both the computational and sociolinguistic dimensions of mixed-language communication.

3 Dataset and Methods

We acquired tweets via the Twitter API from 2017 to 2021 (Kellert and Matlis, 2022) and filtered them using the Spanish language tag "es". Code-switching was defined as Spanish tweets that contained Guaraní words (e.g., *Nde* "hey!"), where the selected Guaraní words were chosen based on their frequency in the literature.

3.1 Geographic Localization of Code-Switching

To map code-switching geographically, we first extracted tweets from the city of Asunción using its defined geographic extent. We then applied a binning algorithm that partitions the city into equal zones (100 x 100 bins), with each bin roughly corresponding to a city block. This fine-grained partitioning enables us to capture localized variations in language use that might be missed with coarser methods. In each zone, we computed the relative frequency of Guaraní and Spanish words using a metric called the *Differential Distribution*. This metric calculates the difference in the proportion of tweets containing Guaraní words versus those containing only Spanish words in each bin. Positive values indicate a higher presence of Guaraní words, while negative values reflect a higher presence of Spanish words (Kellert and Matlis, 2022). This approach quantifies the degree of code-switching in specific urban areas.

To reduce the impact of sparse data, the normalization step in our method suppresses noise from low-count bins and ensures that the overall sum of differences across all bins is zero, making results comparable across zones. We visualized these results using Cartopy in Python, where red markers denote zones with more Guaraní words and blue markers indicate zones with more Spanish words. The size of each marker reflects the magnitude of the differential value, offering an immediate visual cue to the strength of language preference in each area. Base maps were generated using OpenStreetMap data under the Open Database License. These detailed visualizations highlight distinct patterns of language use in Asunción and provide a replicable framework for analyzing code-switching in other urban contexts. For the underlying tweet data, the reader can refer to the first author of this paper.

3.2 Syntactic Localization of Code-Switching

To determine the syntactic position of code-switching within a sentence, we first segment each sentence into words and then divide it into three parts: initial, middle, and final segments. For a sentence

$$S = [w_1, w_2, \ldots, w_n],$$

the initial segment S_{init} comprises the first $\lfloor 0.3n \rfloor$ words, the middle segment S_{mid} includes the words from $\lfloor 0.3n \rfloor + 1$ to $\lfloor 0.7n \rfloor$, and the final segment S_{end} consists of the remaining words. Code-switching is identified when consecutive words are assigned different language labels.

For example, consider the sentence with Spanish and English words where the Spanish word ’Oye’ is used at the beginning of the sentence:

"Oye, I don’t know what to do."

This sentence consists of 6 words. We assign language labels as follows:

$$L(S) = [\mathrm{ES}, \mathrm{EN}, \mathrm{EN}, \mathrm{EN}, \mathrm{EN}, \mathrm{EN}].$$

This sentence can then be divided into an initial segment containing the words [“Oye,” and “I”], a middle segment containing [“don’t” and “know”], and a final segment containing [“what” and “to do”]. Code-switching is detected between “Oye,” (ES) and “I” (EN), which places the switch in the initial segment. This simple segmentation method allows us to categorize the syntactic positions where code-switching occurs and to analyze their distribution across our corpus.

4 Results

4.1 Geographic Distribution of Code-Switching

A clear pattern emerges when examining the map of Asunción (see Figure 1), which illustrates the relative prominence of Guaraní words in Spanish tweets (red) versus Spanish tweets without Guaraní words (blue) based on tweets from 2017–2021. Red markers cluster in the western and southwestern parts of the city, indicating higher Guaraní usage, while blue markers dominate the central and eastern areas, suggesting a stronger preference for Spanish. Some regions show an overlap of red and blue, implying zones with more balanced bilingual practices.

These patterns may stem from various social, economic, and historical factors. Red-heavy areas could reflect neighborhoods with deeper indigenous roots or informal communicative settings, whereas blue-dominant zones might correspond to commercial or governmental districts where Spanish is the default. Transitional neighborhoods and culturally diverse districts often display both colors, indicating regular use of both languages. Overall, this distribution underscores how local context can shape language preferences and highlights the multifaceted nature of code-switching in Asunción.

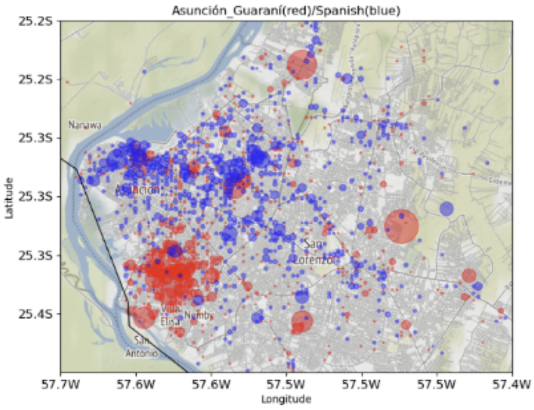

Figure 1: Relative prominence of Guaraní (red) vs. Spanish (blue) in Greater Asunción in the Twitter corpus collected from 2017-2021. Map produced using Cartopy[*] on OpenStreetMap[†] data.

4.2 Syntactic Distribution of Code-Switching

An analysis of sentence-level code-switching (see Figure 2) reveals that switches most frequently occur in the middle of sentences, often within complement phrases (e.g., *Nadia ya le dijo que Si a Marc Anthony, opa la ore amor’i con eso...*). Verbs, pronouns, and nouns are common points of transition, with verbs emerging as the most frequent category for Guaraní-Spanish switches. Tweets containing code-switching also tend to express a range of themes, including emotions, invitations to celebrations, sports-related discussions, and everyday experiences, indicating that bilingual usage is woven into many facets of daily life.

5 Discussion

The geographic results suggest that social, economic, and historical factors may shape language preferences in Asunción. Areas with a strong Guaraní presence might be linked to communities

[*] https://scitools.org.uk/cartopy/

[†] http://wiki.openstreetmap.org/wiki/Open_Database_License

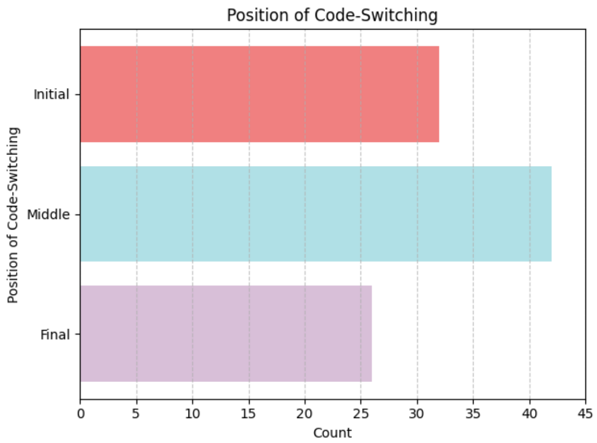

Figure 2: Position of Code-Switching within sentences. Bottom line indicates final sentence position, Top line indicates initial position and middle line indicates middle sentence position

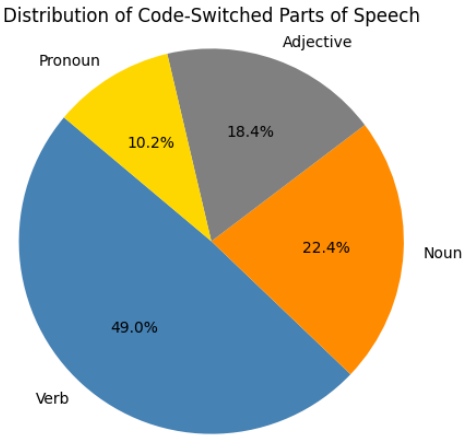

Figure 3: Distribution of parts of speech in Spanish-Guaraní code-switching. Verbs form the largest portion (49.0%), followed by nouns (22.4%), adjectives (18.4%), and pronouns (10.2%).

preserving indigenous linguistic heritage, whereas the dominance of Spanish in central and eastern zones could reflect formal, commercial, or governmental settings. Mixed regions underscore the fluid boundaries where both languages are regularly used.

From a syntactic standpoint, the tendency for code-switching to cluster in the middle of sentences highlights the role of complement phrases and specific parts of speech (especially verbs) in bilingual discourse. This pattern aligns with observations in other code-switching contexts, suggesting that grammatical constraints and discourse functions heavily influence where switches occur. Future work could investigate whether similar patterns emerge in other bilingual communities and how they correlate with social or cultural factors.

Figure 3 illustrates the distribution of parts of speech involved in code-switching, revealing that verbs make up the largest portion (49.0%), followed by nouns (22.4%), adjectives (18.4%), and pronouns (10.2%). This dominance of verbs and other content words supports previous findings that content-rich elements are more likely to be switched than function words, possibly due to their communicative salience in bilingual contexts.

Thematically, the prevalence of code-switching in tweets related to emotions, social interactions, and popular culture suggests that bilingual speakers employ both languages for expressive and affective purposes. This supports the idea that Guaraní serves as a marker of identity and intimacy in informal communication (Estigarribia, 2020). The presence of code-switching in digital discourse also indicates that social media provides a unique space for bilingual expression, free from the constraints of formal linguistic norms. The methodological approach employed in this study, particularly the use of differential distribution for geographic analysis and syntactic segmentation for linguistic analysis, provides a replicable framework for future research on code-switching. By leveraging large-scale social media data, our approach overcomes the limitations of traditional survey and interview methods, offering real-time insights into bilingual language use.

6 Conclusion

In conclusion, this study sheds new light on the spatial and syntactic characteristics of Spanish-Guaraní code-switching, demonstrating its strong ties to geographic, social, and communicative factors. Our findings enhance the broader understanding of bilingual language use and offer valuable implications for sociolinguistic research, computational linguistics, and language policy in Paraguay and beyond.

Limitations and Biases

While our approach offers significant insights, some limitations should be noted. First, although Twitter provides a rich corpus of spontaneous bilingual communication, it may not fully capture language use across all demographic groups, particularly older or less digitally active populations. Second, our study has focused primarily on geographic and syntactic dimensions, leaving other aspects such as sentiment and discourse dynamics to be further explored. Third, our current method approximates the syntactic position of code-switching but does not pinpoint the exact location within complex sentence structures. These limitations present opportunities for refinement without detracting from our study's overall contributions.

Future Work

Building on our findings, future research can integrate data from additional social media platforms to achieve a more complete picture of code-switching trends. Incorporating sentiment analysis and advanced syntactic parsing techniques will provide deeper insights into the emotional and structural dimensions of code-switching, enabling more rigorous testing of linguistic theories such as the Noun Phrase Constraint (Berk-Seligson, 1986). Finally, applying our methods to other bilingual communities will help assess the generalizability of our approach and further enrich our understanding of bilingualism as a global phenomenon.

Acknowledgements

Olga Kellert would like to thank Nicholas Hill Matlis for supporting the design and software development of the geographic analysis of geolocated tweets (see (Kellert and Matlis, 2022)).

References

Marvin M. Agüero-Torales, Antonio G. López-Herrera, and David Vilares. 2023. Multidimensional affective analysis for low-resource languages: A use case with guarani-spanish code-switching language. *Cognitive Computation*, (4).

Marvin M. Agüero-Torales, David Vilares, and Antonio G. López-Herrera. 2021. On the logistical difficulties and findings of jopara sentiment analysis. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 95–102. Association for Computational Linguistics.

Peter Auer and Carol M. Eastman. 2010. Code-switching. In J. Jaspers, J.-O. Östman, and J. Verschueren, editors, *Society and Language Use*, pages 84–112. Benjamins, Amsterdam, Philadelphia.

Susan Berk-Seligson. 1986. Linguistic constraints on intrasentential code-switching: A study of spanish/hebrew bilingualism. *Language in Society*, 15(3):313–348.

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 1–18. Cambridge University Press, Cambridge, UK.

Jhon Adrián Cerón-Guzmán and Elizabeth León-Guzmán. 2016. Lexical normalization of spanish tweets. In *WWW'16 Companion*, Montréal, Québec, Canada. ACM.

Graziella Corvalán. 2005. La vitalidad de la lengua guaraní en el paraguay. *Población y Desarrollo*, 30:9–27.

Wolf Dietrich. 2002. Guaraní criollo y guaraní étnico en paraguay, argentina y brasil. In M. Crevels, S. van de Kerke, S. Meira, and H. van der Voort, editors, *Current studies on South American languages*, pages 31–41. Leiden, Netherlands.

Wolf Dietrich. 2010. Lexical evidence for a redefinition of paraguayan 'jopará'. *STUF- Language Typology and Universals / Sprachtypologie und Universalienforschung*, 63(1):39–51.

Irene Eleta and Jennifer Golbeck. 2014. Multilingual use of twitter: Social networks at the language frontier. *Computers in Human Behavior*, 41:424–432.

Bruno Estigarribia. 2020. *A Grammar of Paraguayan Guarani*. UCL Press.

M. Fricke and G. J. Kootstra. 2016. Primed codeswitching in spontaneous bilingual dialogue. *Journal of Memory and Language*, 91:181–201.

Jack Grieve, Chris Montgomery, Andrea Nini, Akira Murakami, and Diansheng Guo. 2019. Mapping lexical dialect variation in british english using twitter. *Frontiers in Artificial Intelligence*, 2(11).

Gualberto A. Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2017. Moving code-switching research toward more empirically grounded methods. In *Proceedings of the Workshop on Corpora in the Digital Humanities (CDH)*, pages 1–9.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2023. Tuning heli-ots for guarani-spanish code switching analysis. In *Proceedings of the Workshop on Computational Approaches to Multilingual Code-Switching*.

Guido Kallfell. 2011. *Grammatik des Jopara: Gesprochenes Guaraní und Spanisch in Paraguay*. Peter Lang, Frankfurt am Main.

O. Kellert and N. H. Matlis. 2022. Geolocation of multiple sociolinguistic markers in buenos aires. *PLoS ONE*, 17(9):e0274114.

Olga Kellert. 2022. Gender neutral language in (greater) buenos aires, (greater) la plata, and córdoba: An analysis of social context information using textual and temporal features. *Frontiers in Sociology*.

Olga Kellert. 2023a. Linguistic variation in twitter: a case study of italian loanwords in spanish of south america. In Natascha Pomino, Eva-Maria Remberger, and Julia Zwink, editors, *From Formal Linguistic Theory to the Art of Historical Editions: The Multifaceted Dimensions of Romance Linguistics*, pages 347–359. V&R unipress.

Olga Kellert. 2023b. Probing sociodemographic influence on code-switching and language choice in quebec with geolocation of tweets. *Frontiers in Psychology / Language Sciences*, 14.

Olga Kellert. 2023c. Using geolocated tweets for probing language geography and migration. In Sandra Issel-Dombert, Ignacio Andrés Soria, and Laura Morgenthaler García, editors, *Language, Migration and Multilingualism in the Age of Digital Humanities*, pages 129–137. De Gruyter.

Gerrit Jan Kootstra, Joost Schilperoord, and Janet G. van Hell. 2020. Interactive alignment and lexical triggering of code-switching in bilingual dialogue. *Frontiers in Psychology*, 11:1747.

Alberto Muñoz-Ortiz and David Vilares. 2023. Guarani-spanish code-switching analysis. *Manuscript*.

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.

Shana Poplack. 1985. Contrasting patterns of codeswitching in two communities. In Monica Heller, editor, *Codeswitching. Anthropological and sociolinguistic perspectives*, pages 215–243. Mouton De Gruyter, Berlin.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on Twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1971–1982, Vancouver, Canada. Association for Computational Linguistics.