Predictability Effects of Spanish-English Code-Switching: A Directionality and Part of Speech Analysis

Josh Higdon
University of Florida
j.higdon@ufl.edu

Valeria Pagliai
University of Florida
vpagliai@ufl.edu

Zoey Liu
University of Florida
liu.ying@ufl.edu

Abstract

Research on code-switching (CS), the spontaneous alternation between two or more languages within a discourse, remains relatively new and often limited by the use of elicited production tasks, with some exceptions leveraging naturalistic corpora. This study analyzes the effects of language directionality and part-of-speech (POS) tags on Spanish-English CS production between corpus modalities and speech communities. We use data from two spoken corpora: Miami Bangor Corpus (MBC; N = 261,711) and Spanish in Texas Corpus (STC; N = 416,784), as well as the written LinCE Corpus (N=278,093). Bootstrap analyses indicate that Spanish serves as the matrix language (i.e., the most used) for MBC and LinCE, while English is for STC. Logistic regression analyses show that the particle-coordinating conjunction combination was the strongest POS predictor of a CS. The results suggest that corpus modality and the speech community affect matrix language proportions and that both previously attested and unseen POS combinations modulate the production of Spanish-English CS.

1 Introduction

Code-switching (CS), or the fluid switching between languages in bilingual speech or text (Poplack, 1980), is ubiquitous in bilingual communities. This language alternation is known to be structured yet spontaneous (Myers-Scotton, 1993), as it is believed that two (or more) languages are active in the speaker's brain (Van Hell et al., 2015). CS is categorized as inter-sentential (e.g., "No saldré hoy [*I am not going out today*]. I'm too tired.") or intra-sentential (e.g., "She bought una casa en Florida [*a house in Florida*]"). Since languages are not used with the same frequency, the most dominant is considered the matrix language, strongly influencing the morphosyntactic structure of the discourse (Myers-Scotton, 1993).

Previous research has identified several characteristics of CS. For example, the presence of cognates seems to trigger a language switch (Kootstra et al., 2020), and the POS combination of determiner-noun has been noted as a frequent point of switch (Balam et al., 2020). Specifically in the case of Spanish-English CS, there is evidence suggesting that Spanish often serves as the matrix language in CS (Carter et al., 2010), and, regardless of this matrix, there is a consistent pattern where a Spanish determiner is followed by an English noun (e.g. *la house*) (Toribio, 2023). However, these studies mainly rely on (small) sets of controlled stimuli from psycholinguistic experimentation (cf. Soto et al. (2018); Winata et al. (2023)), lacking analysis of naturalistic data (e.g., corpora), meaning that it is unclear whether the mentioned patterns are present in naturalistic CS speech.

Our work addresses these limitations. Leveraging three Spanish-English CS corpora covering both spoken and written modalities, we ask: (1) To what degree is CS production constrained by the language directionality (i.e., Spanish-English vs. English-Spanish) of the CS? (2) To what degree can the POS combination of a pair of words predict whether a CS occurs between those words? (3) If directionality and POS effects are found, are these findings modulated by corpus modality (oral vs. written) and speech community?

Given that Spanish tends to occur as the matrix language in Spanish-English CS (Carter et al., 2010), we predict that Spanish-English CS occurs at higher proportions than in the opposite direction. We also predict that particular POSs will be able to predict the occurrence of CS, specifically with determiners (DETs, among other tags) occurring as the first word in a bigram containing a CS (Eichler et al., 2012; Balam et al., 2020) and conjunctions (CONJ) occurring as the second word in a bigram containing a CS (Soto et al., 2018).

It is difficult to hypothesize the answers to the

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 83–89
August 29, 2025 ©2025 Association for Computational Linguistics

third question, since there is a lack of literature on Spanish-English CS in the written domain, and also that the effect of speech community on CS has been generally neglected (Chan, 2009; Couto et al., 2021). Regarding the effect of corpus modality on CS production, it might not be surprising to see differences in matrix language selection between oral and written corpora, taking into account various differences (including register variations) between modalities (Rabinovich et al., 2019). In one of the few studies examining the effects of speech community on Spanish-English CS, Blokzijl et al. (2017) compared determiner-noun (DET-N) CS production in the Miami Bangor Corpus (MBC) to a corpus of interviews conducted in Nicaragua. The authors found significantly higher rates of Spanish-English DET-N CSs in the MBC, while English-Spanish DET-N rates were higher in the Nicaragua Corpus (Blokzijl et al., 2017). As such, we conjecture that differences in regional norms may manifest in directionality effects. For example, Spanish-English CS may occur at higher rates in the MBC than the Spanish-English proportions of the Spanish Texas Corpus (STC), since Spanish has been documented as the matrix language in Spanish-English CS in the MBC (Carter et al., 2010), but not in STC.

2 Related Work

Recent experiments have used computational analyses of linguistic corpora to uncover Spanish-English CS trends (Winata et al., 2023), but they face limitations in two respects. First, most experiments focus on modeling distributional trends in CS (e.g., the preference for the *estar* + English gerund switch (Tsoukala et al., 2019)) or improving a model's ability to diagnose the presence of a CS (Iliescu et al., 2021) through semi-supervised language identification methods that leverage monolingual data and models like Viterbi decoding. Second, much of the corpus-based research on Spanish-English CS is limited to one corpus, such as the MBC (Deuchar et al., 2009) or the STC (Bullock and Toribio, 2024), raising the question of whether the trends discovered would be representative of those across speech modalities and communities. Furthermore, given that non-computational analyses have found regional differences in DET-N CS patterns, it stands to reason that these differences would appear in CS behavior (Blokzijl et al., 2017). However, this analysis did not account for regional differences in other structural properties of Spanish-English CS (e.g., other POS pairings or trends in matrix language usage).

Closer to this research, through a cross-linguistic analysis of Spanish-English and Mandarin-English corpora, Chi and Bell (2024) highlight POS tags as strong predictors of CS and expand this idea by concluding that this predictive strength decreases the farther a word is from the CS point. While their study underscores the value of POS based approaches, their Spanish-English analysis was also limited to the MBC. We extend overall previous research by comparing CS patterns in three corpora of different modalities, identifying a range of POS tag combinations that predict the presence or absence of CS, and testing prior assumptions about Spanish as the default matrix language. We highlight the role of modality and speech community as key factors shaping CS behavior.

3 Experiments

3.1 Datasets and preprocessing

In contrast to prior corpus-based studies on Spanish-English CS, which typically used only one of these corpora, we selected three corpora that enabled us to include both spoken and written data across different speaking communities.

Miami Bangor Corpus The Miami Bangor Corpus (MBC) (Deuchar et al., 2009), previously used in studies of bilingual speech (Soto et al., 2018; Soto and Hirschberg, 2017), consists of transcriptions of conversations from 86 speakers (33 male, 53 female) based in Miami, Florida. The mean speaker age was 33 years old, and 91.6% reported having at least a college education. On average, speakers acquired Spanish between 2 and 4 years old and English between 4 years old and primary school age. The MBC corpus comprises of a total of 242,475 words, of which 63% is in English, 34% in Spanish, and the rest is undetermined.

Spanish in Texas Corpus The Spanish in Texas Corpus (STC) (Bullock and Toribio, 2024) consists of transcribed interviews and conversations with speakers based in different cities across Texas. It contains approximately 500,000 words from 96 speakers (36 male and 60 female) whose mean age was 39.1 years upon corpus creation. Although information about age of acquisition was not directly provided, 78.1% of speakers reported speaking primarily Spanish with their parents. Additionally, 92.7% of speakers reported having received at least

a high school education. As such, we highlight the similarities between the MBC and the STC in terms of speaker backgrounds. However, in terms of the linguistic makeup of the corpus, the STC is 96% Spanish and 4% English, making it the corpus with the most pronounced language imbalance (and the highest proportion of Spanish) in this study.

LinCE Corpus Unlike the previous corpora, which consist of spoken data, the LinCE Corpus (LinCE) (Aguilar et al., 2020) is the only one based on written texts retrieved from X (formerly Twitter). No background information regarding the speakers of the LinCe corpus is publicly available. The LinCe corpus contains 390,953 words, of which approximately 33% are in Spanish, 64% in English, and the remaining portion is undetermined.

Although the corpora contain POS information, this is not directly comparable across all three due to differences in coding schema. To ensure consistency and scalability, we automatically annotated the POS tags using tools aligned with the Universal Dependencies (UD) annotation scheme (de Marneffe et al., 2021), instead of the tags provided within the data. This decision was also motivated by our observations of several inconsistencies in the existing annotations. We did retain the language tags present in all three corpora and based the subsequent analysis on them.

We used Stanza (Qi et al., 2020) to parse every sentences with the English model. Each word was then processed based on the language tag they had in the original data. Spanish words were reanalyzed with spaCy's es_core_news_sm model, which is trained on the AnCora corpus (Taulé et al., 2008), while unrecognized English words (POS tagged as "X") were reanalyzed with the spaCy model en_core_web_sm to resolve this. We applied spaCy's models at the token level to ensure accurate annotation of individual words within mixed-language utterances.

To investigate the effects of POS production on CS, we analyzed the relationship between the occurrence (or lack thereof) of CS and POS tag bigrams. This was motivated by Soto et al. (2018) who also examined the roles of POS tags on CS production. After filtering punctuation at the sentence level for each corpus, we extracted POS bigrams sequences and included their corresponding language tags to diagnose the presence of a CS. If the two words in a bigram belonged to different languages, we categorized it as a CS.

3.2 Statistical analyses

To examine language directionality effects on CS production, we subjected each corpus to bootstrapping analysis (Efron and Tibshirani, 1994). We calculated the number of bigrams in which the language of the first word was Spanish and that of the second word was English, indicating a Spanish-English CS. The reverse was true for calculating the number of English-Spanish CSs. We divided the count of each CS type (Spanish-English, English-Spanish) by the number of total CSs to calculate its proportions for both CS types. We conducted 10,000 iterations of this process to derive 95% confidence intervals (CIs). Mean proportions and CIs were calculated using a) exclusively the POS bigrams labeled as CSs, and b) all three corpora combined to understand the quantity of each CS type relative to the entire corpus.

We used a logistic regression model (Cox, 1958) to analyse POS effects on CS production across data. Soto et al. (2018) found that a variety of tags, such as determiners (DET), nouns (NOUN), pronouns (PRON), subordinating conjunctions (SCONJ), tend to be associated with CS occurrence; however, their analyses were derived from descriptive statistics with Chi-squared tests. In contrast, our usage of logistic regression models enables us to reliably assess whether a certain combination of POS tags can predict the presence of a CS. In detail, for each POS bigram, which was used as the fixed effect, a binary variable was created to indicate whether a CS was ('1') or was not ('0') present; this variable was the dependent variable. To control for the potential effect of the individual corpus on CS production, we included the corpus as a fixed effect, with the MBC corpus as the reference level (see the formula below).

$$\text{CS\_OCCURRENCE} \sim \text{POS BIGRAM} + \text{CORPUS}$$

All regression models were fit using the glm function in R version 4.3.3 (R Core Team 2018).

4 Results

4.1 Bootstrapping analysis

The results of the bootstrapping analyses can be seen in Table 1. Reliable differences were found for each corpus; these were consistent even when looking at the entirety of MBC, STC and LinCe combined. For the MBC and LinCe corpora, the mean proportion of Spanish-English CS was no-

Bootstrap estimation	Corpus	ES-EN Proportion	EN-ES Proportion
CS tokens exclusively	MBC	0.55 [0.54, 0.57]	0.44 [0.42, 0.46]
	LinCE	0.55 [0.54, 0.57]	0.44 [0.42, 0.45]
	Texas Spanish	0.47 [0.44, 0.48]	0.52 [0.51, 0.53]
Entire corpus	MBC	0.00980 [0.00939, 0.0120]	0.00784 [0.00747, 0.00822]
	LinCE	0.0100427 [0.0096, 0.01045]	0.00797 [0.00763, 0.00834]
	Texas Spanish	0.0129 [0.0125, 0.0132]	0.0143 [0.0139, 0.0146]

Table 1: Proportions of Spanish-English and English-Spanish CSs, measured from a) exclusively the POS bigrams labeled as CSs, and b) all three corpora combined.

POS tag before CS	Coef.	*N*
INTJ	0.13	2,815
PROPN	0.08	1,305
NOUN	0.06	2,960
PART	0.04	60
ADJ	0.02	1,062
VERB	0.009	2,338
NUM	0.005	139
DET	0.004	1,843
ADP	0.0007	1,636
ADV	-0.0002	1,349
AUX	-0.003	822
CCONJ	-0.005	749
SCONJ	-0.01	643
PRON	-0.01	912

Table 2: Part of Speech (POS) tags of words *preceding* a CS with their coefficient estimates and counts, ordered from the most predictive to the least.

POS tag after CS	Coef.	*N*
INTJ	0.12	1,767
PROPN	0.06	1,370
PART	0.04	185
CCONJ	0.024	1,168
SCONJ	0.021	982
ADJ	0.016	1,364
NOUN	0.012	2,838
ADV	0.006	1,509
ADP	0.003	1,527
AUX	-5.20E-07	753
PRON	-0.0005	2,561
DET	-0.001	906
VERB	-0.003	1,535
NUM	-0.01	168

Table 3: Part of Speech (POS) tags of words *following* a CS with their coefficient estimates and counts, ordered from the most predictive to the least.

tably higher, while the mean proportion of English-Spanish CS was higher in STC. STC's higher proportion of English-Spanish CS goes against the prediction that Spanish-English CS proportions would be higher than English-Spanish CS proportions. Moreover, it is unexpected given that previous literature (Carter et al., 2010; Couto et al., 2021) has demonstrated that Spanish tends to be the matrix language in Spanish-English CS.

4.2 Logistic regression analysis

Of 195 POS combinations detected across the three corpora, 150 POS combinations were statistically significant predictors of CS. Estimate values from the model were consulted to determine whether a given POS combination predicted that a CS would occur (POSitive estimate) or not (negative estimate). 60 POS combinations predicted the occurrence of a CS, while 90 combinations predicted that a CS would not occur. We calculated mean estimate values across all statistically significant POS combinations to derive the values found in Table 2 and Table 3.

As presented in Table 2, among all POS tag bigrams that indicate the occurrence of a CS, interjections (INTJ) and proper nouns (PROPN) were the most predictive of a CS occurring (please see Example 1 in Table 4 for an example of a CS where INTJ precedes the CS). However, when compared to all bigrams predicting that a CS would not occur, the subordinating conjunction (SCONJ) and pronoun (PRON) tags were the least effective at predicting a CS would occur. DET, which was predicted to be the first word in a CS pair due to the prevalence of the DET-N CS (e.g., Valdés Kroff et al. (2018); Balam et al. (2020)), had the second-smallest estimate value of the tags that were predictive of CS. See Example 2 (Table 4) for an example of a DET-N CS.

Similarly, the estimate values provided in Table 3 indicate that INTJ, PROPN, and PART tags being the second tag in a bigram effectively diagnose that a CS has occurred. See Example 3 in Table 4 for an example of a CS containing a PART tag. The auxiliary (AUX), PRON, and DET tags (among others) effectively determine that a CS does *not* occur in the same context. Interestingly, we found that SCONJ and CCONJ happen regularly as the second, but not the first, word in a CS bigram, contrasting with Soto et al. (2018), who found that both tags occur at significantly higher rates as the first word in a bigram containing a CS. Examples

Number	Corpus	Tag preceding CS with Lang ID	Tag following CS with Lang ID	Example
1	LinCe	INTJ (EN)	AUX (ES)	**Like**, hay que vacilar (*like, you have to stagger*)
2	STC	DET (ES)	NOUN (EN)	... para el **deadline** (*for the deadline*)
3	MBC	CCONJ (ES)	PART (EN)	**Pero** to test for lead (*but to test for lead*)
4	LinCe	ADJ (EN)	SCONJ (ES)	... how **funny que** I'm sitting there (*how funny that I'm sitting there*)

Table 4: Specific examples of code-switching contexts from the three different corpora.

of some of the pairs mentioned can be found in Table 4.

5 Discussion and Limitation

After analyzing three Spanish-English spoken and written corpora, we expanded previously known patterns of directionality and POS in CS. Bootstrapping showed a tendency in Spanish-English switches throughout the MBC and LinCE corpora, while the opposite was found in the STC. The logistic regression analysis concluded a robust number of POS combinations that predicted CS: tags like INTJ and PROPN were commonly present during switches, while PRON and SCONJ were not.

This paper contributes to the field by highlighting the complex role that POS plays in Spanish-English CS. Similarly to Soto et al. (2018), we found that a variety of tags predicts the event that precedes (INTJ, PROPN, NOUN and PART, among others) or follows (INTJ, PROPN, PART, CCONJ and SCONJ, among others) a given CS. Phrases like Example 1 in Table 4 illustrate how frequent INTJs are more characteristic of naturalistic language use than of elicited stimuli, which may explain why this POS had not been previously identified as statistically significant.

The fact that Soto et al. (2018) did not find predictive validity of the SCONJ and CCONJ tags following a CS but our experiment did, can be explained by a difference in the corpora analyzed. Soto et al. (2018) examined solely a portion of the MBC, while our work covered two additional corpora, besides the entirety of MBC; the CS behavior in Soto et al. (2018)'s analyses may be qualitatively and quantitatively different from the data used in this study. Moreover, a switch occurring before a CCONJ and SCONJ could be expected since these POS tags often introduce new clauses or sentence-like units, creating opportunities for a language switch (as seen in Example 4, Table 4).

To our knowledge, this is the first research to empirically establish a higher proportion of English-Spanish CS relative to Spanish-English CS with the written STC corpus. This finding may be motivated by pragmatic differences in CS across written and oral corpora. However, it could also be the case that speech community and corpus modality combine to affect distributional trends of Spanish-English and English-Spanish CS. Further investigation of this would require more written Spanish corpora, which is lacking at the moment; however, we hope that future work can mitigate this gap.

Despite the implications of this study, it is not without limitations. First, only three corpora were analyzed due to the lack of more publicly available data. We look forward to expanding our experiments to additional speech communities in the future. Second, we only examined Spanish-English CS, two languages considered relatively high-resource; we hope that our analyses can be extended to different language pairings, particularly those that involve understudied languages, to probe the generalizability of our findings.

One final limitation of our analyses is that we used POS tags and not dependency relations to investigate Spanish-English CS. Our decision, which is line with with previous CS research (i.e. Soto et al. (2018), Balam et al. (2020), inter alia), was motivated by the fact that POS tags are (albeit a somewhat simplistic) way to gain information about structural characteristics of Spanish-English CS. For example, our analyses provided novel insight into the predictability of INTJ, NOUN, and PART occurring before a CS and INTJ, PART, CCONJ, and SCONJ occurring after a CS. However, we recognize the value of including dependency relations in our analyzes, given that in German-English CS, two adjacent words have been reported to be more likely to be in the same language if they are more directly related in terms of dependency relations (Eppler, 2010). We expect that future analyses consider both POS tags and dependency relations to further enrich the field's understanding of Spanish-English CS.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Osmer Balam, María del Carmen Parafita Couto, and Hans Stadthagen-González. 2020. Bilingual verbs in three Spanish/English code-switching communities. *International Journal of Bilingualism*, 24(5-6):952–967.

Jeffrey Blokzijl, Margaret Deuchar, and M Carmen Parafita Couto. 2017. Determiner asymmetry in mixed nominal constructions: The role of grammatical factors in data from Miami and Nicaragua. *Languages*, 2(4):20.

Barbara E. Bullock and Almeida Jacqueline Toribio. 2024. Spanish in Texas Corpus.

Diana Carter, Peredur Davies, Ma Carmen Parafita Couto, and Margaret Deuchar. 2010. A corpus-based analysis of codeswitching patterns in bilingual communities. *Revista Española de Lingüística*.

Brian Hok Shing Chan. 2009. *Code-switching between typologically distinct languages*. Cambridge University Press.

Jie Chi and Peter Bell. 2024. Analyzing the role of part-of-speech in code-switching: A corpus-based study. In *Findings of the Association for Computational Linguistics*, page 1712–1721. Association for Computational Linguistics. The 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 ; Conference date: 17-03-2024 Through 22-03-2024.

M Carmen Parafita Couto, Miriam Greidanus Romeli, and Kate Bellamy. 2021. Code-switching at the interface between language, culture, and cognition. *Lapurdum*.

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Margaret Deuchar, Maria del Carmen Parafita Couto, Peredur Davies, Kevin Donnelly, Fraibet Aveledo, Diana Carter, Marika Fusser, Jon Herring, Lowri Jones, Siân Lloyd-Williams, Myfyr Prys, Elen Robert, and Jonathan Stammers. 2009. Bangortalk Corpora. Accessed April 21, 2025.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. Chapman and Hall/CRC.

Nadine Eichler, Malin Hager, and Natascha Müller. 2012. Code-switching within determiner phrases in bilingual children: French, Italian, Spanish and German. *Zeitschrift für französische Sprache und Literatur*, 122(3):227–258.

Eva Eppler. 2010. *Emigranto:: The syntax of German-English code-switching*. Wilhelm Braumüller Universitäts-Verlagsbuchhandlung.

Dana-Maria Iliescu, Rasmus Grand, Rob van der Goot, and Sara Qirko. 2021. Much gracias: Semi-supervised code-switch detection for Spanish-English: How far can we get? In *Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, page 65. Association for Computational Linguistics.

Gerrit Jan Kootstra, Ton Dijkstra, and Janet G. van Hell. 2020. Interactive alignment and lexical triggering of code-switching in bilingual dialogue. *Frontiers in Psychology*, Volume 11 - 2020.

Carol Myers-Scotton. 1993. *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Ella Rabinovich, Masih Sultani, and Suzanne Stevenson. 2019. CodeSwitch-Reddit: Exploration of written multilingual discourse in online discussion forums. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, page 446, Hong Kong, China. Association for Computational Linguistics.

Victor Soto, Nishmar Cestero, and Julia Hirschberg. 2018. The role of cognate words, POS tags and entrainment in code-switching. In *Interspeech*, pages 1938–1942.

Victor Soto and Julia Hirschberg. 2017. Crowdsourcing universal Part-of-Speech tags for code-switching. *CoRR*, abs/1703.08537.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Almeida Jacqueline Toribio. 2023. 'Doing' Romance Linguistics. *Isogloss. Open Journal of Romance Linguistics*, 9(2):1–14.

Chara Tsoukala, Stefan L Frank, APJ van den Bosch, J Valdés Kroff, and Mirjam Broersma. 2019. Simulating Spanish-English code-switching: El modelo está generating code-switches. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics.

Jorge R Valdés Kroff, Rosa E Guzzardo Tamargo, and Paola E Dussias. 2018. Experimental contributions of eye-tracking to the understanding of comprehension processes while hearing and reading code-switches. *Linguistic Approaches to Bilingualism*, 8(1):98–133.

Janet G Van Hell, Kaitlyn Litcofsky, and Caitlin Y Ting. 2015. *Intra-sentential code-switching: Cognitive and neural approaches*. Cambridge University Press.

Genta Indra Winata, Alham Fikri Aji, Zheng-Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. *Preprint*, arXiv:2212.09660.