UNIVERSITY OF CALIFORNIA
Los Angeles

Grounding Code-Switching Evaluation to Community Speech Patterns

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Computer Science

by

Rebecca Pattichis

2024

ABSTRACT OF THE THESIS

Grounding Code-Switching Evaluation to
Community Speech Patterns

by

Rebecca Pattichis
Master of Science in Computer Science
University of California, Los Angeles, 2024
Professor Nanyun Peng, Chair

Code-switching (CS), broadly defined as switching between multiple languages in speech and text, is a common occurrence in multilingual communities. And yet, CS has been historically disparaged in higher institutions, including in the research field of Natural Language Processing. This thesis contextualizes CS dataset collection, transcription, and analysis for better data quality. Specifically, I improve CS dataset analysis by adapting previous metrics in NLP that are based on word-level units, which are misaligned with bilingual speech. Crucially, CS is not equally likely between any two words, but follows syntactic and prosodic rules. This work therefore adapts two metrics, multilinguality and CS probability, to use the Intonation Unit (IU), an established unit for speech transcription, as basic tokens for NLP tasks. I also calculate these two metrics separately for distinct mixing types: alternating-language multi-word strings and single-word incorporations. Results indicate that there is a shared tendency among bilinguals for multi-word CS to occur across, rather than within, IU boundaries. That is, bilinguals tend to prosodically separate their two languages. This constraint is blurred when metric calculations do not distinguish multi-word and single-word items. By comparing against the same metrics and datasets using the word as a token, I also show that IUs help researchers distinguish between CS speaker patterns, whereas the word-based metrics homogenize and obscure these patterns. These results call for a reconsideration of units of analysis in future development of CS datasets for NLP tasks.

The thesis of Rebecca Pattichis is approved.

Baharan Mirzasoleiman

Kai-Wei Chang

Nanyun Peng, Committee Chair

University of California, Los Angeles

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I would like to express my gratitude to Dr. Rena Torres Cacoullos from Pennsylvania State University, along with Dr. Dora LaCasse and Dr. Sonya Trawick. It is through Dr. Torres Cacoullos' willingness to share her carefully collected community data that this thesis came to fruition at all. Additionally, I gained an immense amount of linguistic knowledge through this group's dedicated collaboration and mentorship. I will continue to carry the research and linguistic methodology that I learned here through my professional endeavors.

I would also like to thank Dr. Nanyun Peng, who advised me throughout this thesis. Dr. Peng's encouragement and excitement towards my research interests was a refreshing breath of fresh air in a field that often feels exclusive. Additionally, her well-formed technical recommendations continue to motivate the continuation of this work. I am forever thankful for the advising relationship I have built with Dr. Peng, which ultimately grounded me in this work and field.

This thesis would not have been finished without the support system of my friends. Thank you to Christina, Arjun, Amita, and Ana, for continuing to re-center me while pursuing my graduate studies at UCLA. The weekend dinner sessions allowed me to not only air out my frustrations and insecurities, but also gain immense amount of support and technical help. Without y'all, this would not have been accomplished! To Cher, Shea, Ashley, and Miza: thank you for providing material support in the form of housing, cooked meals, and exercise breaks in the last few weeks of my thesis deadline.

Thank you to my family. To my mother, for instilling in me the importance of my languages, for calling me regularly to ask about my thesis and also about my life outside of this work. To my father, for being on the phone with me on walks, for listening through all of my doubts in this work and continuing to encourage me to finish strong. And to my younger sister, Antonia, thank you for being the best friend anyone could ask for. Your support even as we are apart from each other has reminded me to laugh, breathe, and enjoy life through all of my identities. Thank you for the continued love and support!

CHAPTER 1

Introduction

Code-switching (CS) is broadly defined as switching between multiple languages in a span of speech or text. This can be interpreted at several levels, including: people switching languages in a conversation, a single person switching languages in their utterance between words, and even someone CS within a word boundary. Below is an example showcasing the second case, reproduced verbatim from the New Mexico Spanish-English Bilingual (NMSEB) corpus (Torres Cacoullos and Travis 2018: Chs 2 & 3)[1], with the translation following the original transcription. Italic and roman type represent speech originally in English (E) and Spanish (S), respectively.

con mi espejito,	S
que así me miraba y,	S
if I could see my profile,	E
... y luego me volteé pa' atrás,	S
with my little mirror,	S
that I was looking at myself and,	S
if I could see my profile,	E
... and then I turned around,	S
(03, 05:45-05:53)	

CS has gained traction in Natural Language Processing (NLP) and Linguistics for

[1] Within parantheses following each example are the NMSEB corpus transcript number and beginning-end time stamp. https://nmcode-switching.la.psu.edu

a variety of reasons. Importantly, CS is prominent in multilingual communities among speakers with similar CS patterns (cf. Deuchar 2020; Poplack 2018). Linguists concern themselves with identifying the mechanisms and consequences of CS. This has resulted in several linguistic theories around bilingual's grammatical preference for CS (e.g., the Equivalence Constraint (Poplack, 2013; Sankoff, 1998), Matrix Language Framework (Myers-Scotton, 1993), and Functional Head Constraint (Belazi et al., 1994)), as well as studying whether CS induces grammatical language change (Lorenz, 2019). CS research in NLP mainly arises with its presence in social media data (Winata et al., 2023), and was recently acknowledged as an important research direction at the Association for Computational Linguistics in 2022 (Doğruöz et al., 2023). All CS research, however, suffers from the same issue; namely, the lack of useful and representative CS language data (Lorenz, 2019; Doğruöz et al., 2021). This is due to the difficulty of automatically identifying CS in the wild, the labor it requires to effectively transcribe multilingual speech data, and analyzing the range and quality of CS apparent in data.

To this end, this work focuses on the latter point by adapting CS metrics for effectively clustering and comparing bilingual, and specifically CS, speech. I[2] center transcripts from the NMSEB data, and use this real-world bilingual speech to ground my methodology. The remaining sections in this chapter will describe the state of CS in NLP before outlining the rest of the chapters in this thesis.

1.1 Code-Switching in Natural Language Processing

CS is a product of multilingualism in communities. However, CS and parallel monolingualism should not be equated. Parallel monolingualism, which is often considered the only valid form of fluency in educational and broader institutional spaces in the United States (Martínez, 2017), critically excludes CS patterns by multilingual communities.

[2] This thesis reports on work involving multiple collaborators and ultimately submitted to several conferences. Moving forward, every chapter will explicitly mention the collaborators involved in that work.

This linguistic bias is also evident in NLP research. Notably, early multilingual language models (MLMs) such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) were both trained on parallel monolingual chunks of text (i.e., Wikipedia and a subset of the CommonCrawl Dataset) (Liu et al., 2019). And while early studies find that mBERT doesn't drop in accuracy for CS data (Pires et al., 2019), their experiments were not comprehensive, with only one dataset and one task explored (Khanuja et al., 2020). Later experiments actually show that these MLMs cannot match the performance of hierarchical meta-embeddings (HMEs) – which combine word, subword, and character level information – for a variety of downstream tasks (Winata et al., 2021). Notably, HMEs are trained on far less data, highlighting two generalizable methodological faults with current language model development:

1. *More data does not guarantee better performance:* Similar findings extend to T5 (Raffel et al., 2020), an encoder-decoder model based on the Transformer architecture (Vaswani et al., 2017), when evaluated on African languages. When researchers meticulously filter the mC4 data (Raffel et al., 2020) for 16 African languages, they find that a T5 model trained on the smaller filtered corpus can consistently improve the original T5 model's performance for downstream tasks, even doubling accuracy for the task of question answering (Oladipo et al., 2023). Indeed, an audit of web-crawled multilingual corpora (including mC4) reveal that most languages have a high percentage of non-language tokens and even incorrect languages in their subset (Kreutzer et al., 2022). This makes 'representation washing' – the claim that a diverse set of languages is represented in corpora – a critical fault in NLP research for low-resource languages (Kreutzer et al., 2022). CS falls under this categorization (Doğruöz et al., 2021), and therefore requires intentional dataset curation to ensure the data that is collected is representative of community speech.

2. *CS requires distinct language learning assumptions:* Winata et al. (2021) suggest that the masked language model task is not the ideal training objective for learning CS representations (Winata et al., 2021), which challenges previous research claiming

that mBERT outperforms cross-lingual word embeddings (Khanuja et al., 2020). In other words, while Khanuja et al. (2020) claim it to be a data problem, Winata et al. (2021) supplement this (by experimenting with more complex embedding methods and holistically evaluating on performance, speed, and number of parameters) to reveal the learning objective problem. In addition to not relying on massively collected data for CS research, we also cannot inherently rely on increasingly large model architectures or previously successful learning objectives. Instead, NLP methods for CS must be linguistically grounded (Khanuja et al., 2020).

Both of these problems extend to generative models, which are notoriously trained on massive amounts of data and are large with respect to parameter size. Specifically, current popular generative models (e.g., ChatGPT (Brown et al., 2020), BLOOMZ (Muennighoff et al., 2022), etc.) do not outperform smaller, fine-tuned MLMs previously mentioned for several downstream tasks (Zhang et al., 2023). Additionally, recent research shows that generative models are unreliable or even fail to produce CS examples deemed realistic by native-speaking human annotators (Yong et al., 2023). Linguistically grounded generation methods are an alternative to massive MLMs like ChatGPT, and often require both less data and less parameters (Pratapa et al., 2018; Pratapa and Choudhury, 2021; Gregorius and Okadome, 2022). Impressively, some methods don't require training data at all, but rather rely solely on established linguistic theories for CS (Gregorius and Okadome, 2022).

Regardless, data is at the core of CS research in NLP. Whether it be necessary for training language models or to validate linguistic theories, assessing the quality and representativeness of CS data is of key importance.

1.2 Thesis Overview

I have emphasized the importance of establishing CS metrics to assess data quality and representativeness. In this work, I adapt two CS metrics – the Multilingual Index (M-Index) (Barnett et al., 2000) and Integration Index (I-Index) (Guzmán et al., 2017) – and provide key linguistic insights for CS boundaries in bilingual speech. Importantly, these metrics are better aligned to CS speech patterns, and allow for a comparison between datasets (and speakers). My contributions are as follow:

1. I quantitatively validate the Intonation Unit (IU)-Boundary constraint, which states that bilingual speakers prefer switching at prosodic boundaries.

2. I use the Intonation Unit (IU) to adapt the token level out of the word-token level for the M-Index and I-Index, which measure inequality of language representation and observed probability of switching, respectively.

3. Through IU-based CS metrics, I show that researchers should avoid amalgamating lone items (or single-word insertions) with other forms of CS. This is especially relevant to NLP, as we often focus on simpler linguistic theories that are advantaged by assuming homogeneity of CS types.

4. I compare the IU and word as the token level for CS dataset analysis, showing that the former is more informative in clustering speakers.

The rest of this thesis is broken up as follows: Chapter 2 delves further into the background of CS conceptualizations. Chapter 3 provides a description of the Spanish and English bilingual transcripts I use for this work. Additionally, I introduce key transcription conventions, including the IU and an individual tagging system for lone items. Chapter 3 details the methodological alteration made to the M-Index and I-Index to use IUs as tokens instead of words, and compares this token with the word token for five different transcripts. Chapter 5 concludes with future research directions and a discussion on dataset collection for CS.

CHAPTER 2

Background

In this chapter, I provide the background necessary to understand code-switching (CS) as a linguistic phenomenon, as well as how CS has been conceptualized in linguistics and computer science research. Specifically, Section 2.2 details previously published linguistic theories, Section 2.3 describes previous efforts to develop CS data, and Section 2.4 outlines metrics to quantitatively analyze and compare CS data.

2.1 Code-Switching

While concretely defining code-switching (CS) is contentious (Cacoullos and Travis, 2015), for this work I define CS as a speaker switching between multiple languages during a speech utterance. CS is a common phenomenon among multilingual speakers – which is a majority of the world's population – and is associated with language contact. There are notably distinct CS patterns that arise between communities. For example, CS can occur when there is contact between a standard and dialectical form of a language (e.g., Standard Modern Greek and Cypriot Greek in Cyprus (Armosti, 2009)) (Doğruöz et al., 2021), or between different languages coming into contact, such as in Puerto Rican communities in East Harlem CS between several variations of Spanish and English (Zentella, 1998).

CS patterns also change generationally. Specifically, Zentella (1998) found evidence of English beginning to replace Spanish as the dominant language among second and third generation New York Puerto Ricans. This is similar for northern New Mexico, where the population of non-English speakers reported in the U.S. Census has dropped since statehood, and where English was imposed in educational institutions (Cacoullos and

Travis, 2015). The latter context is where I gather the data for this work, and therefore more of this community's context will be discussed in Chapter 3.

Geography and age are only a few examples of the varying axes of contextualization necessary for effectively understanding, representing, and analyzing CS. Others include class status, register (formal vs. informal), and language background (Doğruöz et al., 2023). Fine-grained context is especially important for CS, as an individual's CS patterns cannot be separated from their social community and upbringing (Labov, 2006). This implies that effective CS data collection is grounded in a specific speech community (Cacoullos and Travis, 2015), or at least some defined community. Ultimately, despite code-switching (CS) being seen as a negative linguistic practice, it is actually one grounded in community and cultural practice (Zentella, 1998).

2.2 Linguistic Theories for CS

There are several linguistic theories that attempt to formalize the way that bilinguals CS. The most widely adopted of these theories within NLP research – at least in terms of language generation – is the Matrix Language Framework (MLF) (Joshi, 1982; Myers-Scotton, 1997). The MLF assumes an asymmetrical switching framework, where the grammar of the main/matrix language operates as the base, and words or phrases from the embedded language are inserted (see example (1) in Table 2.1). Several works utilize this framework when modeling CS for data generation (Lee et al., 2019; Gupta et al., 2020; Gregorius and Okadome, 2022), admittedly because it is easier to implement and less restrictive.

Although it is a popular linguistic theory for NLP researchers, it is ultimately not representative of more complex CS patterns (Doğruöz et al., 2021). The Equivalence Constraint, on the other hand, states that CS is more likely to occur at points where the languages' linear structures coincide (Poplack, 2013). Importantly, the Equivalence Constraint can handle alternational CS, which is usually more complex and requires more sophisticated fluency in both languages (Doğruöz et al., 2021). For example, in Table

(1)	... (1.4) y entró con un *ashtray* grande. and came in with a big *ashtray* (03, 18:15-18:19)	SLS
(2)	.. contratista general, ... que todavía acarreo la licencia de, .. *general building*, .. general contractor ... that I still carry the license .. *general building* (27, 05:53-05:59)	S S E

Table 2.1: Example of CS, where (1) is an English noun insertion into otherwise Spanish speech, and (2) is a multi-word switch from Spanish to English. Notably, (2) is not captured by the Matrix Language Framework, because the speaker invokes both the English and Spanish grammar structure. Both examples contain their translations directly underneath.

2.1, example (2) CS occurs from Spanish to English, and invokes both grammars (i.e., adjective after the noun in 'contratista general' vs. adjective before the noun in *general building*). Several works have also modeled this theory for dataset generation (Winata et al., 2019; Pratapa et al., 2018), although it involves more complex constraints for switching boundaries.

The MLF and Equivalence Constraint are the two most popular linguistic theories in NLP. Both are syntactic theories, and are only a subset of a much larger set of syntactic linguistic theories of CS (see Doğruöz et al. (2021) and Winata et al. (2023)). In this work I quantitatively validate a prosodic-based (as opposed to syntactic-based) linguistic theory for CS, known as the Intonation Unit (IU)-Boundary Constraint. The IU will be introduced in Chapter 3 alongside the dataset I use and its transcription conventions. The IU-Boundary Constraint, which accounts for speech patterns, can be understood alongside syntactic theories for conceptualizing CS.

2.3 Code-Switching Datasets

Aguilar et al. (2020) released the first benchmark for evaluating models on several downstream tasks for CS. Known as Linguistic Code-switching Evaluation (LinCE) benchmark (Aguilar et al., 2020), it includes datasets for Language Identification, Part of Speech, Named Entity Recognition, Sentiment Analysis, and Machine Translation for a combination of nine distinct language pairs (including Spanish and English). The datasets come from previously collected data and range from transcribed interviews (Deuchar, 2008) to social media data. In the same year, Khanuja et al. (2020) released the Generalized Language Understanding Evaluation for Code-Switching (GLUECoS) Benchmark (Khanuja et al., 2020) for two languages pairs, which draws from GLUE (Wang et al., 2018) to evaluate embedding models on a variety of language understanding tasks. Notably, in addition to including some overlap with the LinCE benchmark, they also create evaluation datasets for Natural Language Inference, which is considered to be more difficult than aforementioned tasks and missing from previous CS research.

A recent audit of 68 CS datasets brought up several critical faults in their collection and creation (Doğruöz et al., 2023) (see Table 2 and 3 of their Appendix for comprehensive results). Doğruöz et al. (2023) find that most datasets, whether speech or text-based, fail to report geographic, socio-demographic, and register variation information. This results in overgeneralizations and claims about the data quality (e.g., internet data is still majority in English, and it is unclear how much CS is present). Another critical nuance is that none of the audited datasets explicitly report who the data collectors, transcribers, or annotators are. It is crucial to know this information because 1) I know CS patterns differ among several community contexts, and 2) the relationality between the interviewer and interviewee (i.e., whether they are in the same community) impacts the degree of more complex and community CS patterns (Poplack, 1983).

Doğruöz et al. (2023) conclude by imploring researchers working with CS data to report about the geographic, register-based, and socio-demographic context of the data, along with any filtering or post-processing methods. My work answers most of these questions (including interviewer and transcriber demographics) in Chapter 3 when describing the data I draw from, as well as the post-processing steps in Section 4.1.

2.4 Metrics for Code-Switching Data

CS is unique from monolingual data in that it merits understanding the degree of CS present. This is especially important for quantitatively comparing datasets. To that effect, a variety of works have proposed metrics to quantitatively compare CS across datasets. Below I intuitively describe each widely-used CS metric, with Chapter 4 mathematically introducing the CS metrics I decide to adapt.

The first set of metrics concern themselves with measuring the underlying language distribution or representation in the data. The Multilingual Index (M-Index), which ranges from 0 to 1, measures the inequality of language distribution (Barnett et al., 2000). In other words, a value of 0 implies a completely monolingual text, while 1 represents a perfect distribution of languages. The Code-Mixing Index (CMI), also ranging from 0 to 1, measures the proportion of language tokens that are not part of the majority language (which is determined by the language with the most amount of tokens in the data) (Das and Gambäck, 2014). While the CMI claims to measure amount of CS in the data, it inherently only deals with measuring the language representation between the majority language and the minority language(s).

The Integration Index (I-Index), on the other hand, measures the observed probability of CS (Guzmán et al., 2017). With the range of 0 to 1, the I-Index simply records how many of the possible switch-points (i.e., boundaries between word tokens) actually did contain a switch between languages. Later work introduces the Normalized I-Index (Bullock et al., 2019), which notes that the original I-Index is inherently tied to the language distribution (or M-Index value), and should therefore be normalized to account for its possible minimum and maximum. For example, unless there is a perfect balance of languages (M-Index of 1), there is no way for the original I-Index to even reach its maximum value.

All of the aforementioned CS metrics were introduced for the word token level. This construction, however, brings about two separate assumptions about CS that are faulty. The first one is that CS is equally likely to occur at any boundary between two words, which is untrue given several syntactic linguistic theories such as the Equivalence Constraint. The second is that by assuming that many switch-point locations (i.e., at every word boundary), the universe of CS is unrealistically large. Consequentially, metrics such as the I-Index do not reach its full range for realistic CS patterns (Bullock et al., 2019). In this work, I introduce the IU as a valid token for CS metrics rather than the word token by adapting the I-Index, refining our understanding of CS boundaries and aligning these metrics to CS speech behaviors. Through this, I show that this adapted version allows for a more nuanced clustering of speaker's CS patterns within the same community.

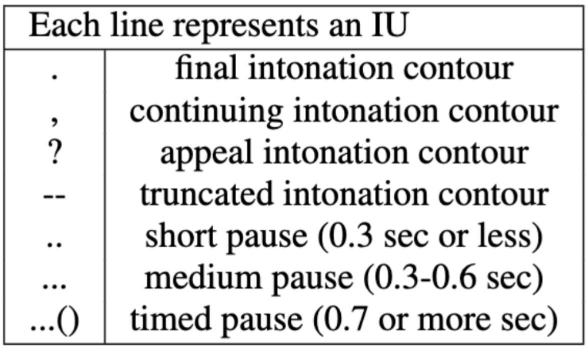
CHAPTER 3

The New Mexico Spanish-English Bilingual Corpus

I have established the data scarcity problem that exists for low-resource language practices, including code-switching (CS). I have also discussed the importance of contextualizing the data collected for CS, as it varies across several axes. In this chapter, I describe the bilingual dataset that I draw from for this work. In this, I include information about the speaker's geographic context and historical language contact, the demographics of the interviewers and transcribers, as well as the transcription conventions. The latter are especially crucial for my methodological process, which is in Chapter 4.

3.1 Dataset Description

The data for this thesis come from the New Mexico Spanish-English Bilingual (NMSEB) corpus (Torres Cacoullos and Travis 2018: Chapters 2 and 3), a collection of transcribed interview conversations from community members in New Mexico containing New Mexican Spanish and English. The speakers are bilingual members of a long-standing speech community in northern New Mexico where Spanish and English have both been spoken for 150 years. Because of this, interviewees in NMSEB are at least third-generation New Mexican, allowing us to look at CS in terms of long-term language contact (Cacoullos and Travis, 2015). Participants were chosen if they frequently use both languages, determined by a combination of self-reports, content analysis of audio, and the data that speaker's produce. Language proficiency tests have historically carried a negative connotation, and therefore would not have worked in this context for gathering participants (Cacoullos and Travis, 2015).



Each line represents an IU	
.	final intonation contour
,	continuing intonation contour
?	appeal intonation contour
--	truncated intonation contour
..	short pause (0.3 sec or less)
...	medium pause (0.3-0.6 sec)
...()	timed pause (0.7 or more sec)

Figure 3.1: Full transcription conventions.

Interviewers were recruited from the University of New Mexico. Additionally, interviewers are ethnically in-group and are relationally associated as family or acquaintances (Torres Cacoullos and Travis, 2018). This is critical because it impacts the distribution of CS, namely in that multi-word switching (i.e., more complex switching) increases with in-group members (Poplack 2013: 113). Each interviewer was conducted as a sociolinguistics interview (Labov, 2006), where interviewers focused on questions surrounding the participant's life story or experiences. This helped capture speech that is used in daily dialogue, such as *acequia* (irrigation ditch) or *troca* (truck).

3.2 Transcription Conventions

Transcribers were also relationally associated to the community, which is crucial as New Mexican Spanish is distinct to other popular varieties in New Mexico (e.g., Mexican Spanish) (Cacoullos and Travis, 2015). Each transcriber had access to the ELAN audio file, allowing for a more accurate orthographic transcription process. Next, I outline the prosodic transcription of each interview, along with the distinction and motivation between distinguishing between single- and multi-word CS.

3.2.1 The Intonation Unit

In addition to orthographic transcription, NMSEB is prosodically transcribed, following the convention for identifying an Intonation Unit (IU) (Torres Cacoullos and Travis,

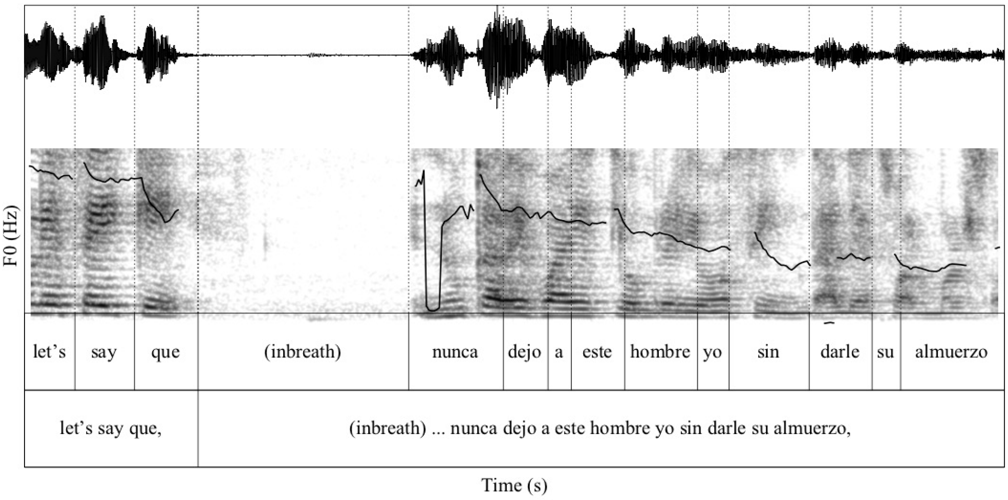


Figure 3.2: Acoustic properties of Intonation Unit include higher pitch at the beginning of the IU and slower rate of speech at the end of the IU, and sometimes pausing between IUs.

2018; Du Bois et al., 1993). In particular, IUs are "uttered under a single, coherent intonation contour" (Du Bois et al. 1993:47; see Figure 3.2 for an example). In the NMSEB corpus, each line of transcription represents an IU, with punctuation marking the transitional continuity between IUs, or the terminal pitch contour (see Figure 3.1 for complete transcription conventions). Below is an example from transcript 05, where the second column indicates 'E' for English and 'S' for Spanish.

y para nosotros *it was a snap*,	SE
.. *cause we already knew,*	E
.. *English,*	E
'and for us *it was a snap*,'	
.. *cause we already knew,*	E
.. *English,*	E
(10, 01:22-01:26)	

Table 3.1: Example of prosodic transcription in NMSEB, including the language tagging convention per IU.

Prosodic transcription has long been part of linguistic analysis (Halliday, 2015) as it allows for accurate delineation of speech and clause boundaries, rather than attempting to guess sentence boundaries (Cacoullos and Travis, 2015). Instead, the prosodic sentence is objectively defined as an IU or series of IUs containing at least one finite verb that ends in intonational completion (Chafe 1994:139). Syntactically, it displays nice properties. For example, pronouns and verbs are often in the same IU, and words in the same IU tend to have a closer syntactic relationship than those in neighboring IUs (Croft 1995:849-864).

IUs are also relevant for CS. Specifically, previous studies on the NMSEB corpus have uncovered the tendency for speakers to prefer CS across, rather than within, IUs (Steuck, 2018). Known as the IU-Boundary Constraint, it formally captures bilinguals' tendency to prosodically separate the two languages. For CS patterns in speech, then, we can think of the IU-Boundary Constraint as operating alongside the Equivalence Constraint, such that switching is far more likely between two words at the boundary of IUs. In other words, these prosodic and syntactic constraints are related in that the looser syntactic relationship between words at IU boundaries than within them may mean greater word order flexibility and therefore a higher likelihood to CS. However, the two constraints are independent in that, both at IU boundaries and within IUs, CS is subject to the Equivalence Constraint and, among Equivalence points, CS is more likely at IU boundaries than within them. Responding to these facts (i.e., likelihood to switch is not uniform across all word token boundaries), I apply CS metrics using IUs rather than individual words as the token, described further in Chapter 4.

IUs are tagged for language, such that all-Spanish and all-English IUs are tagged 'S' and 'E', respectively. IUs hosting both languages are tagged 'SE' or 'ES', or other combinations of 'S' and 'E' (as in the example in Table 3.1). Some IUs also host potentially language neutral items. Proper nouns are tagged 'P' (e.g., California) and discourse markers or backchannels are tagged 'D' (e.g., *so*, which appears both in English and in the otherwise monolingual Spanish of bilingual speakers in New Mexico (Aaron, 2004)).[3] The tagging also distinguishes single-word incorporations, or lone items, tagged 'L'.

[3] Other material not tagged as either language: fillers (e.g., *uh*, *mhm*, tagged as 'F' when they occur in their own IU), the word no when it occurs at a switch point or adjacent to other potentially language-neutral material ('N'), IUs consisting of laughter or other nonlinguistic material ('A'), and unclear speech ('X').

3.2.2 Lone items

Distinguishing single-word incorporations, or lone items, allows for more nuanced analysis of CS types. Amalgamating lone items with all other types of CS is a documented problem in CS analysis within NLP (Doğruöz et al., 2021). Indeed, by homogenizing CS as all one and the same, NLP researchers can get away with coarse implementation and analysis, such as only adopting the MLF or overstating their model's performance with CS.

Lone (single-word) items are common nouns and other, mostly content, words incorporated into otherwise monolingual discourse, such as in the NMSEB example in Table 2.1 as well as below.

'with the *flashlight* in one hand,' ELE
(16.1, 25:04-25:06)

Table 3.2: Example of single-word CS.

Lone items differ from multi-word CS in their structural properties (Poplack 2017; Torres Cacoullos and Travis 2020:256-259). While lone items are placed according to the word order/grammar of the language in which they are embedded (in line with the MLF), multi-word CS is considered to be more complex (Doğruöz et al., 2021). In contrast to single-word CS, multi-word CS strings consistently operate such that each language phrase is consistent with the grammar of that respective language (which is not included in the MLF conceptualization). Refer back to Table 2.1 for a detailed comparison.

Another interesting difference between single- and multi-word CS is the asymmetry between the two languages for lone items (Figure 3.3). In the New Mexico bilingual community, the tendency is for lone nouns to be mostly English rather than Spanish (Torres Cacoullos and Vélez Avilés, 2023). In contrast, multi-word CS is fairly balanced by CS direction (42% Spanish to English, 29% English to Spanish, 29% two or more CS within the prosodic sentence).

Ultimately, the combination of prosodic transcription alongside distinguishing between single- and multi-word CS allows us to analyze bilingual's CS tendencies in the New

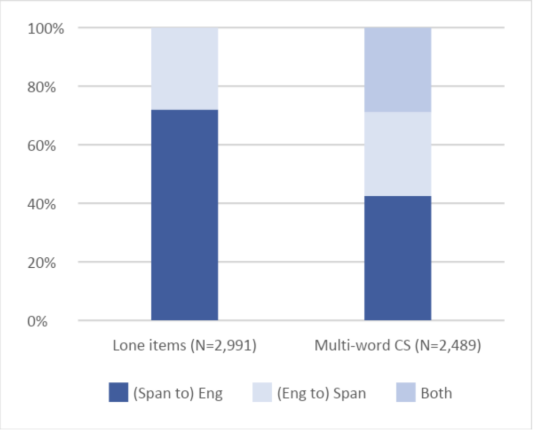


Figure 3.3: Distribution of lone items and multi-word CS by language (from Pattichis et al. (2023a):16842)

Mexican community. Specifically, I quantitatively validate the IU-Boundary Constraint by adapting aforementioned CS metrics for the IU token (rather than the word token). Through this construction, I also provide insights about speaker's CS patterns between single- and multi-word CS.

CHAPTER 4

Code-Switching Metrics Using Intonation Units

In this chapter, I quantitatively validate the Intonation Unit (IU)-Boundary Constraint, which states that speakers prefer switching across, rather than within IU boundaries, on a subset of the NMSEB corpus. Specifically, I accomplish this by adapting the M-Index and I-Index, previously developed word-based metrics, to use the IU token. Because NMSEB also distinguishes between single- and multi-word CS, I provide insight on how the two types of CS interact with prosodic boundary CS. To my knowledge, this is the first study to use the IU as the token of measurement for NLP analysis of CS datasets.[4]

4.1 NMSEB Chosen Data

I work with five transcripts from NMSEB, totaling 4.8 recorded hours, 41,000 words, and 14,135 IUs. In the transcripts chosen, 84% - 97% of the IUs are produced by the speaker rather than the interviewer or another interlocutor, satisfying my threshold for majority monological speech. This allows us to treat CS as occurring within the same speaker turn (not in response to "interactive alignment" with an interlocutor, see, e.g., Kootstra et al. (2020)). The speakers are: transcript 03, Sandra (administrator, Española); 05, Rocío (teacher aid, Santa Fe); 10, Pedro (school administrator, Taos); 16, Manuel (electrician and rancher, Rio Arriba); and 27, Eduardo (general contractor and store owner, Rio Arriba). Names given are pseudonyms and locations listed are either counties or major cities to protect speaker privacy. Anonymization also occurred within each transcript, so

[4] This work is done with collaborators Dr. Dora LaCasse, Dr. Sonya Trawick, and Dr. Rena Torres Cacoullos, and can also be accessed through the EMNLP 2023 publication (Pattichis et al., 2023b).

that any real names, nicknames, or identifiable proper nouns (i.e., small cities, places of work, high school names, etc.) were replaced and indicated with a preceding '~' symbol (Torres Cacoullos and Travis 2018:48).[5]

4.2 Methods

For this work, I use IUs as the base token (as opposed to words). I consider IUs as eligible for future analysis if they contained a language tag of either 'S' (Spanish), 'E' (English), or 'L' (Lone Item). Note that a complete IU language tag may be a combination of 'S', 'E', or 'L'; see previous examples for illustrations of language tagging.

In order to understand the impact of how combining lone items with multi-word strings affects our understanding and perspective of CS, I compute the I-Index (Guzmán et al., 2017) for different representations of the corpora (see Tables 4.1-4.3). Specifically, I-Index measures were calculated two ways: by considering only 'S' or 'E' for analysis, and by also including 'L' (re-coded as 'S' or 'E').

Since I am operating at the IU-token level, each token can contain more than one language tag (i.e., if there is a within-IU language switch). Below I describe the binary measures that allow us to maintain the integrity of the token level.[6]

4.2.1 M-Index

The Multilingual Index (M-Index) proposed by Barnett et al. (Barnett et al., 2000) is meant to measure the multilinguality of a given corpus with at least two languages from a range of 0 to 1, where the former is monolingual, and the latter means there is a perfect balance of languages. Here, k denotes the number of languages in the corpus, and p_j is

[5] The NMSEB corpus records the spontaneous vernacular of a close-knit minority language community, from interactions with in-group fieldworkers, sometimes of a highly personal nature. In accordance with the participant consent form, protocols for access by those familiar with the speech community protect against unintentional publication of stereotyping examples (Torres Cacoullos and Travis 2018:47-49; cf. Poplack 2022:212,217).

[6] Code can be found at `https://github.com/rpattichis/IU-Boundary_constraint_code`.

the number of tokens in language j divided by the total number of tokens:

$$\text{M-Index} = \frac{1 - \sum p_j^2}{(k-1) \cdot \sum p_j^2}.$$

While it is originally meant for the word-token level, I use it at the IU level. That is, instead of the numerator of p_j representing the number of words in language j, I instead make p_j the amount of IUs in language j divided by total IUs. When an IU has multiple languages contained within its bound, I tag it with the earliest language present (i.e., an IU with 'SES' will count as 'S'). For the M-Index, I only considered 'S' and 'E' as valid language tags.

4.2.2 Across-IU I-Index

Here, I use the Integration Index (I-Index) developed by Guzman et al. (Guzmán et al., 2017) to measure the probability of CS in each transcript. Specifically, the I-Index is meant to approximate the probability that any given token is a switch point. Here, n is the number of tokens, and $S(l_i, l_j)$ is 1 if two neighboring tokens are in different languages, 0 otherwise:

$$\text{I-Index} = \frac{1}{n-1} \sum_{1 \leq i = j-1 \leq n-1} S(l_i, l_j).$$

Again, while this metric was developed with an assumption of words as tokens, I count the IUs as tokens. Then, for the across-IU I-Index, I ask the question: Does a language switch happen at the boundary between the i^{th} IU and the $(i+1)^{th}$ IU? This binary measure determines the value of $S(l_i, l_j)$. Here, I consider two perspectives: first, with only the 'S' and 'E' language tags, and then with the inclusion of 'L' to understand how lone items impact switching across IUs.

Corp	**Total S/E**	**Across IU**		**W/in IU**		IUs
		no Ls	Ls incl.	no Ls	Ls incl.	w/ Ls
05	1911/532	77	101	5	23	23
27	616/2035	194	238	12	58	54
03	1040/1501	376	464	29	112	111
16	994/1266	189	252	17	71	70
10	737/894	264	276	21	49	31

Table 4.1: Number of IUs counted as 'S' and 'E'; number of IUs hosting CS by prosodic position—Across IU boundaries vs. Within IU—and by treatment of lone items—'No Ls' vs. 'Ls incl.'; IUs hosting Ls.

Corp	**% S/E**	**Across IU**		**W/in IU**		IUs
		no Ls	Ls incl.	no Ls	Ls incl.	w/ Ls
05	78/22	3.2	4.1	0.2	0.9	0.9
27	23/77	7.4	8.9	0.5	2.2	2.0
03	41/59	14.8	18.0	1.1	4.3	4.3
16	44/56	8.4	11.0	0.8	3.1	3.1
10	45/55	16.2	16.8	1.3	3.0	1.9

Table 4.2: Percentages corresponding to Table 1.

4.2.3 Within-IU I-Index

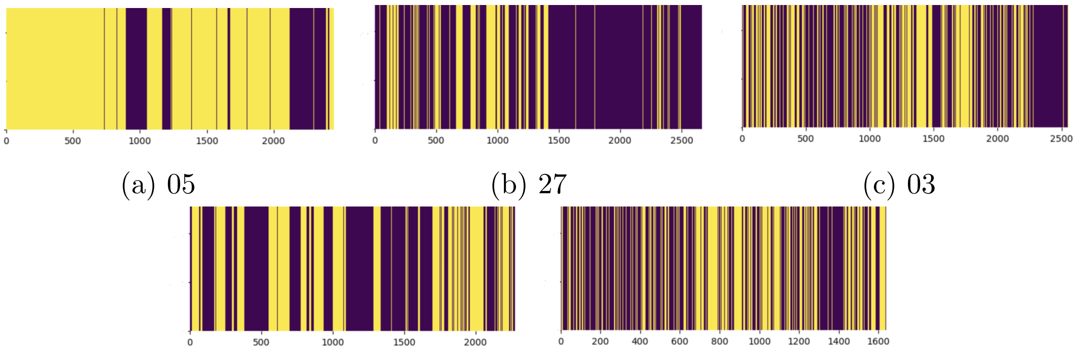
To also account for the within-IU CS that occurs in the corpora, I use the same I-Index but change the question: Does a language switch occur within the i^{th} IU? Although there might rarely be more than one switch point within an IU, I decided to keep the binary measure so as to not double count a token. Here, the only change is that $S(l_i)$ only looks at one token, rather than a comparison between two tokens. Again, for this metric, I consider the two perspectives with only 'S' and 'E' as well as the inclusion of 'L' to understand how including lone items may impact our understanding of CS.

Note that although later work by Bullock et al. (Bullock et al., 2019) propose a normalized I-Index due to its dependence on a corpus's M-Index, I have intentionally chosen transcripts that are comparable in their M-Index. Specifically, the M-Index is close to 1 for three transcripts (03, 10, 16) and .50 for two (05, 27).

Corp	M-Index	I: Across		I:W/in	
	(S/E)	no Ls	Ls incl.	no Ls	Ls incl.
05	0.52	0.03	0.04	0.0	0.01
27	0.56	0.07	0.09	0.0	0.02
03	0.94	0.15	0.18	0.01	0.04
16	0.97	0.08	0.11	0.01	0.03
10	0.98	0.16	0.17	0.01	0.03

Table 4.3: M-Index and I-Index; I-index calculated for CS according to prosodic position and according to inclusion of Lone items (see methods section).

4.3 Results



(a) 05 (b) 27 (c) 03

(d) 16 (e) 10

Figure 4.1: Language distribution graphs for each transcript, where English is in purple and Spanish is in yellow.

Table 4.1 gives the number of IUs counted as 'S' and as 'E' and the number of IUs hosting CS according to prosodic position and the treatment of lone items.[7] Table 4.2 gives the corresponding percentages. M-Index and I-Index for the five transcripts appear in Table 4.3. Of the five corpora chosen, three have an M-Index close to 1, indicating a balance of English and Spanish within the transcript. In particular, M-Indices of .94-.97 correspond to speakers who produce a more balanced 41%-45% of their IUs as 'S' and 55%-59% as 'E', while values of .52-.56 came from speakers with 22%-23% in one language (and more than 75% in the other).

[7] Counts in 'IUs w/ Ls' column do not correspond to the differences between 'Ls incl.' and 'no Ls' in the preceding columns because an 'L' at an IU boundary may count as both an Across- and Within-IU switch.

As shown in Table 4.3, the combination of the M-Index and I-Index is crucial in understanding the nuance of different speakers' CS patterns. For example, transcripts 03 and 16 have similar M-Indexes (with a difference of 0.03), but the former has almost twice the I-Index of the latter speaker. This indicates the independence of the two metrics, and the necessity to use both in tandem for a more comprehensive comparison.

I also use language distribution graphs to visualize the M-Index and I-Index within a transcript, in Figure 4.1. Here, in an ordered array of the language tag for each IU token, English IUs are colored in purple, whereas Spanish is in yellow. These visualizations help clarify what the quantitative metrics distinguish, by the number and width of the language bands. For example, transcripts 03 and 10 (language distribution graphs in Fig. 4.1c and 4.1e, respectively), have a similar M-Index and Across-IU I-Index, which is evident from their language distribution graphs. On the other hand, 05 (Fig. 4.1a) has a patently different M and I-Index from these transcripts, resulting in a smaller total number of language bands and wider bands for Spanish.

As to how the inclusion of lone items impacts our perspective of CS, it seems to have little impact on the Across-IU I-Index, as seen in Table 4.3. However, including the 'L' language tag does have a substantial impact on the Within-IU I-Index. While for the Across-IU index the increase is at most 37.5% (for 16, from 0.08 to 0.11), for the Within-IU index the increase is as large as 300% (for 03, from 0.01 to 0.04) (disregarding the cases of infinity, for 05 and 27). This makes sense – lone items inherently occur within a single IU (e.g., usually nouns positioned at the end or interior of the IU).

Merging the single-word items with multi-word CS also magnifies differences between speakers with respect to the within-IU I-Index. Within-IU I-Indices range from 0 to 0.01 with 'no Ls,' but from 0.01 to 0.04 when lone items are included. This is important because it shows that with this perspective (the only perspective for most NLP research), we fail to capture the IU-Boundary Constraint on CS whereby speakers strongly prefer switching across prosodic units rather than within them. In other words, while there is a noticeable difference between the across and within-IU I-Index regardless, the difference is consistently stark without 'L's, signifying the consistent preference to switch across prosodic boundaries that including ‘L’s weakens.

4.4 Conclusion

I put forward the Intonation Unit (IU) as the token level for CS metrics. In addition, I distinguish between single- and multi-word CS. Applying the IU-based metrics on NMSEB transcripts captures the tendency for multi-word CS to be used from one IU to another in across-IU CS. All speakers (regardless of their M-Index or Across-IU I-Index) strongly disfavor within-IU multi-word CS. In other words, I have effectively shown the IU-Boundary Constraint in action for bilingual speech. In contrast, lone items, which by definition are used within a single IU, notably raise the I-Index for within-IU CS. Merging multi-word CS and lone items thus obscures uniform adherence to the IU-Boundary constraint, despite individual differences in language distributions (M-Index) and CS rate (I-Index).

Ultimately, I highlight the IU as an important token level for future CS datasets, as well as the impact of single-word (lone) items as distinct from multi-word strings on CS metrics, supporting previous claims about their amalgamation.

4.5 IU vs. word-based CS metrics

While the previous section provided important insight into the usefulness of the IU as a token level for CS metrics, it begs the question: How do these results translate to the word-based version of the I-Index (i.e., CS probability metric) for the same transcripts? Next, I describe the post-processing to convert the NMSEB language tags into word tags, and compare a combined IU-based metric value with the word-based results.

4.5.1 Methods

Similar to the previous section, I calculate the I-Index, but this time for the word token (its original use) (Guzmán et al., 2017). But first, to make a direct comparison with word tokens, I report an I-Index value that combines our previous across and within-IU I-Index values (while continuing to avoid double-counting an IU token). Namely, I estimate the overall switching rate by adding the amount of across and within IU CS present, and subtracting the number of IUs I estimate to host both:

$$I_{IU} = \frac{1}{n}[A + \sum_{i=1}^{n} S(l_i) - [A * I_{within}]], \tag{4.1}$$

where $A = \sum_{1 \leq i=j-1 \leq n-1} S(l_i, l_j)$, or the number of switches that happen at IU boundaries, and I_{within} is the within-IU I-Index described in Section 4.2.3. Thus, the # of CS points recorded for the IU-based metric is this combined value.

To convert IU-based language tags from the original transcriptions into a language tag for each word, texts were pre-processed to remove symbols that cannot be mapped to a word, such as '...' indicating pauses. In the simplest case, when an IU has just one language tag, I duplicate that tag by the number of words present in the IU (delimited by spaces). For multiple language tags within a single IU, the goal is to identify the boundaries of each language tag. Strings are parsed according to a particular order of items (e.g., discourse markers 'D' are identified before proper nouns 'P'). For combinations of 'S', 'E', and 'L', I use a pre-trained Language Identification (LID) model trained on the LinCE dataset (Aguilar et al., 2020) for Spanish and English CS to find the language boundaries by word tokens; I also use the LID model for proper nouns not marked by a symbol indicating anonymization.[8]

Once again, to compare outcomes when lone items and multi-word CS are distinguished vs. merged, I calculate the word-based I-Index two ways, where 'L's are treated as the language in which they are embedded vs. where 'L's are counted as CS. For the former, the 'L' tag is converted to whatever the previous language tag was (e.g., if the word-level tags are 'SSSLS', then I count it as 'SSSSS'). For the latter, the 'L' tag is converted to the opposite language (so, the prior example will become 'E' for English, 'SSSES').

[8] The LID model from Code Switch can be found at https://github.com/sagorbrur/codeswitch.

Corp	**Multi-word CS**					
	# S/E	% S/E	# CS	I-Ind.	# SE/ES	% SE/ES
05	5859/1976	75/25	80	0.01	40/40	50/50
27	1970/7013	22/78	206	0.02	103/103	50/50
16	3272/4786	41/59	210	0.03	105/105	50/50
03	3509/6079	37/63	406	0.04	203/203	50/50
10	2497/3686	40/60	285	0.05	143/142	50/50

Table 4.4: Word-based CS metrics when excluding lone items as CS.

Corp	**Ls included**			
	# CS	I-Ind.	# SE/ES	% SE/ES
05	101	0.01	54/47	53/47
27	251	0.03	126/125	50/50
16	266	0.03	152/114	57/43
03	500	0.05	275/225	55/45
10	304	0.05	159/145	52/48

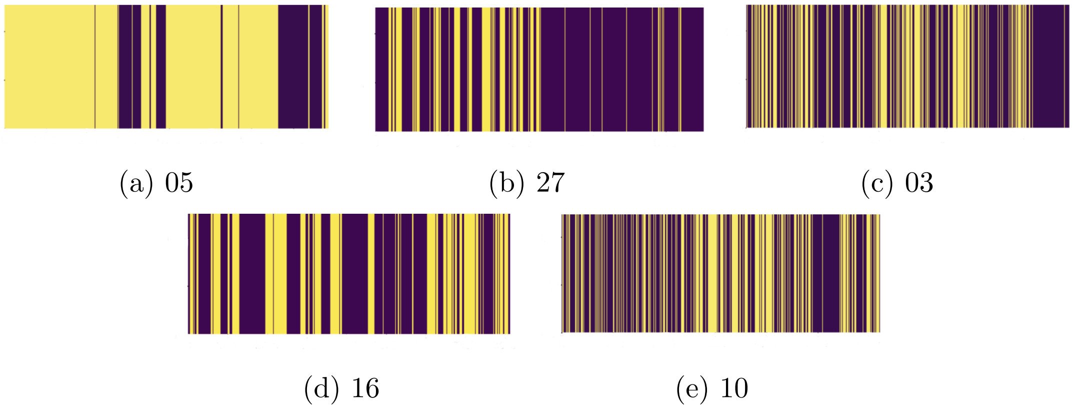
Table 4.5: Word-based CS metrics when including lone items as CS.

Corp	**Multi-word CS**					
	# S/E	% S/E	# CS	I-Ind.	# SE/ES	% SE/ES
05	1911/532	78/22	82	0.03	23/19	55/45
27	616/2036	23/77	205	0.08	63/59	52/48
16	994/1266	44/56	204	0.09	58/48	55/45
03	1040/1501	41/59	401	0.16	149/120	55/45
10	737/894	45/55	281	0.17	51/58	47/53

Table 4.6: Intonation unit (IU)-based CS metrics when excluding lone items as CS.

Corp	**Ls included**			
	# CS	I-Ind.	# SE/ES	% SE/ES
05	101	0.04	34/26	57/43
27	249	0.09	87/75	54/46
16	258	0.11	105/57	65/35
03	482	0.19	218/133	62/38
10	297	0.18	65/62	51/49

Table 4.7: Intonation unit (IU)-based CS metrics when including lone items as CS.



(a) 05

(b) 27

(c) 03

(d) 16

(e) 10

Figure 4.2: Language distribution graphs for each dataset, where English is in purple (darker color) and Spanish is in yellow (lighter color).

4.5.2 Results

The token level is the word in Tables 4.4 and 4.5, the IU in Tables 4.6 and 4.7 (with the latter tables reporting the combined I-Index value for the IU token level). Reported for each dataset are the number of ‘S’ and ‘E’ tags, the percentage of each language, the number of CS points recorded for the I-Index, and the I-Index values. Also shown is the direction of CS, from Spanish to English or the reverse direction.[9] Recall that the IU token inherently supplies the prosodic sentence (as opposed to the syntactic sentence) as the unit within which to assess switching direction. Therefore, in calculating CS direction within the prosodic sentence, I do not count a switch from ‘S’ to ‘E’ or the reverse when the interval between them contains an IU ending in final or appeal intonation (“.” and “?”, respectively). The four rightmost columns give the corresponding results when ‘L’s are included as switch points. For the ‘Ls included’ columns, I only count going into the lone item, and not stepping out of it, in line with previous methods (Wintner et al., 2023).

Interestingly, the word-based calculations result in generally low values of the I-Index, from 0.01 to 0.05. In comparison, the IU-based values display a wider range, from 0.03 to 0.17. This discrepancy in range makes sense. For the word-level calculations, we are counting far more tokens in the denominator as possible switch points. Additionally, the

[9] Recall from Section 3.2.2. that there is asymmetrical use of languages between lone items and multi-word CS.

word-based values suggest that the five datasets, arranged by ascending I-Index, differ equidistantly from each other, by .01.[10] In comparison, using the IU token level makes differences between datasets more evident and suggests clusters among them. The I-Index for 05 is lower than for both 27 and 16, which in turn are lower than both 03 and 10. This ordering and clustering according to the IU-based values seems consistent with the language distribution graphs visualizing the datasets (Figure 4.2). For example, the dataset with the lowest I-Index (05) has one color predominating over the other and also has the fewest bands, whereas the datasets with the highest values (03 and 10) have an even distribution of colors (symmetry of languages) and more bands (quantity of CS). In other words, the IU-based metrics allow for a more granular and representative comparison of CS behaviors between speakers.

A striking repercussion of including the ‘L’s is in the apparent direction of switching. With the word token, switching direction is 50/50, since for every switch into one language a switch is counted into the other. For multi-word CS, the direction of “intra-sentential” switching (‘%SE/ES’, Table 4.5) is quite balanced, ranging from 55/45 to 52/48. Including switches into ‘L’s augments the asymmetry between the languages. In all cases, the asymmetry shifts in favor of ‘SE’ (Tables 4.4-4.7, rightmost column). This follows from the preference for lone *English* noun incorporations in the bilingual community.

The distinction between lone (single-word) items and multi-word CS operationalizes distinctions that have been made under labels such as insertional versus alternational switches. There is increasing recognition in NLP that lone items impact the relation between languages and the direction of switching as appearing asymmetrical (e.g., Bullock et al. 2018:2537, Wintner et al. 2023:1480).

[10] Normalized I-Index (Bullock et al., 2019) values are similarly compressed (0.02, 0.05, 0.03, 0.06, 0.06) for the five datasets.

4.5.3 Conclusion

By comparing the word-based and IU-based I-Index metric for bilingual datasets, I showed that the word token compresses the values and ultimately obscures CS patterns. The more revealing IU-based I-Indexes follow from the patterns of real bilingual speech, where CS is neither as frequent as NLP researchers may convey (through word token boundaries) nor, crucially, is it equally likely between any two words. The present results, then, highlight the IU as a reliable token level, validate the IU-Boundary Constraint, and better align CS datasets and metrics for NLP with bilingual speech patterns, allowing for valuable comparison between speakers.

CHAPTER 5

Conclusion

5.1 Summary of Contributions

In this thesis, I ground previously developed code-switching (CS) metrics in NLP research to bilingual community speech. I draw data from the New Mexico Spanish-English Bilingual (NMSEB) corpus, which was carefully collected and transcribed orthographically and prosodically by in-group community members. Then, by adapting the Integration Index (Guzmán et al., 2017) from words to Intonation Units (IUs) as the token level, I am able to quantitatively validate the IU-Boundary Constraint, which states that speakers prefer CS at prosodic boundaries. Additionally, these results show the importance of distinguishing between single- and multi-word CS, which are operationally different. Lastly, by conducting an empirical comparison between the word-based and IU-based I-Index for my transcripts, I show that the former compresses the I-Index values, while the latter allows for a more nuanced clustering and comparison of speaker's CS patterns.

Ultimately, I advocate for incorporating IUs into future CS datasets through transcription and analysis, especially for comparison among data. This work tackles the two issues raised in CS data, namely quality and contextualized analysis, by utilizing carefully curated data in this work, and aligning NLP research to bilingual speech rather than the other way around. In sum, this thesis is the first to use IUs for CS metrics. The results provide critical insight for future transcription and synthetic data generation methods that can improve CS datasets, which will ultimately impact all downstream NLP tasks.

5.2 Future Directions

I see relevance for this work in Controlled Natural Language Generation (NLG). Much of the CS work is slow/iterative because it is hard to find quality CS data (Tarunesh et al., 2021; Doğruöz et al., 2021). Thus, my work impacts the future construction of CS datasets, specifically by calling for the injection of linguistically well-founded CS patterns to improve NLG methods. Currently, CS generation has focused on word substitutions (Solorio and Liu, 2008; Tarunesh et al., 2021). However, I have shown here that bilingual behavior is not limited to word substitution. In distinguishing and accounting for multi-word language alternations, future NLP metrics can draw on the Equivalence constraint (syntax) and the related but independent IU-Boundary Constraint (prosody) for more refined CS data production.

Important work on auditing CS data already reveals the lack of contextualization in current data collection methods within NLP (Doğruöz et al., 2023). While they looked at different axes of speakers/users represented, a further step should investigate the frequency and distribution of single- vs. multi-word CS. The results here show that there is indeed a quantitative difference between single- and multi-word CS for speakers, which supports previous calls for distinguishing CS types (Doğruöz et al., 2021). A further step, then, could look at auditing CS datasets to elucidate the distribution of CS present. Due to the field's skew towards the MLF (which corresponds to simpler, insertional CS), I believe there will be a general lack of multi-word CS among very popular CS datasets and benchmarks.

I argue that CS data requires intentional collection, transcription, and analysis, making labor and time spent unavoidable. Specifically, speech datasets might heavily consider the orthographic and prosodic transcriptions detailed in Chapter 3 for downstream granular analysis.

5.3 Final Thoughts

NLP is an exponentially growing field within computer science, now reaching broader non-research audiences with the publicized nature of chatbots like ChatGPT (Brown et al., 2020). And while research advancement is exciting, it is always the margins or non-majority of a field that implore us to take our time to ensure quality, representation, and autonomy. In the case of NLP, efficiency and speed are not problems for Standard American English (SAE), as a majority of the internet is in this language. Indeed, the disparity in resource allocation and attention among languages deems SAE to be high-resource, while other varieties such as African American English are considered low-resource.

This problem, which is inherently structural, will not be fixed with more research to automatically and efficiently collect more data for these languages and varieties. Even if it works, without community-based approaches, the efforts at this data collection will not matter because it will not be representative of language patterns, deeming language technologies that build off of them irrelevant. This thesis, then, attempts to place itself in the realm of slow, iterative, and patient work within NLP, which employs data that took a community and decades to collect. By gate-keeping access to NMSEB, the curators are able to ensure that community data is safeguarded to only those that are willing to consult experts and listen to community knowledge. This thesis exemplifies that this collaborative work is possible, and should be centered more often.

REFERENCES

Aaron, J. E. (2004). “So respetamos un tradición del uno al otro”: *So* and *entonces* in New Mexican bilingual discourse. *Spanish in context*, 1(2):161–179. 15

Aguilar, G., Kar, S., and Solorio, T. (2020). LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association. 9, 25

Armosti, S. (2009). The phonetics of plosive and affricate gemination in cypriot greek. *Unpublished PhD dissertation, University of Cambridge*. 6

Barnett, R., Codó, E., Eppler, E., Forcadell, M., Gardner-Chloros, P., Van Hout, R., Moyer, M., Torras, M. C., Turell, M. T., Sebba, M., et al. (2000). The lides coding manual: A document for preparing and analyzing language interaction data version 1.1–july 1999. *International Journal of Bilingualism*, 4(2):131–271. 4, 10, 19

Belazi, H. M., Rubin, E. J., and Toribio, A. J. (1994). Code switching and x-bar theory: The functional head constraint. *Linguistic inquiry*, pages 221–237. 2

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901. 4, 32

Bullock, B., Guzmán, W., and Toribio, A. J. (2019). The limits of Spanglish? In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 115–121. 10, 11, 21, 28

Bullock, B. E., Guzmán, G. A., Serigos, J., and Toribio, A. J. (2018). Should code-switching models be asymmetric? In *Interspeech*, pages 2534–2538. 28

Cacoullos, R. T. and Travis, C. E. (2015). Gauging convergence on the ground: Code-switching in the community. 6, 7, 12, 13, 14

Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press. 15

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics. 3

Croft, W. (1995). Intonation units and grammatical structure. *Linguistics*, 33(5):839–882. 15

Das, A. and Gambäck, B. (2014). Identifying languages at the word level in code-mixed Indian social media text. In Sharma, D. M., Sangal, R., and Pawar, J. D., editors, *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387, Goa, India. NLP Association of India. 10

Deuchar, M. (2008). The Miami corpus: Documentation file. *Bangortalk, bangortalk. org. uk/docs/Miami_doc. pdf*. 9

Deuchar, M. (2020). Code-switching in linguistics: A position paper. *Languages*, 5(2):22. 2

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. 3

Doğruöz, A. S., Sitaram, S., Bullock, B. E., and Toribio, A. J. (2021). A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics. 2, 3, 6, 7, 8, 16, 31

Doğruöz, A. S., Sitaram, S., and Yong, Z.-X. (2023). Representativeness as a forgotten lesson for multilingual and code-switched data collection and preparation. *arXiv preprint arXiv:2310.20470*. 2, 7, 9, 31

Du Bois, J., Cumming, S., Schuetze-Coburn, S., and Paolino, D. (1993). Outline of discourse transcription. In Edwards, J. A. and Lampert, M. D., editors, *Talking data: Transcription and coding in discourse*, pages 45–89. Lawrence Erlbaum Associates. 14

Gregorius, B. and Okadome, T. (2022). Generating code-switched text from monolingual text with dependency tree. In Parameswaran, P., Biggs, J., and Powers, D., editors, *Proceedings of the The 20th Annual Workshop of the Australasian Language Technology Association*, pages 90–97, Adelaide, Australia. Australasian Language Technology Association. 4, 7

Gupta, D., Ekbal, A., and Bhattacharyya, P. (2020). A semi-supervised approach to generate the code-mixed text using pre-trained encoder and transfer learning. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2267–2280, Online. Association for Computational Linguistics. 7

Guzmán, G., Ricard, J., Serigos, J., Bullock, B., and Toribio, A. (2017). Metrics for modeling code-switching across corpora. In *Interspeech*, pages 67–71. 4, 10, 19, 20, 24, 30

Halliday, M. A. K. (2015). *Intonation and grammar in British English*, volume 48. Walter de Gruyter GmbH & Co KG. 14

Joshi, A. K. (1982). Processing of sentences with intra-sentential code-switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*. 7

Khanuja, S., Dandapat, S., Srinivasan, A., Sitaram, S., and Choudhury, M. (2020). GLUECoS: An evaluation benchmark for code-switched NLP. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics. 3, 4, 9

Kootstra, G. J., Dijkstra, T., and van Hell, J. G. (2020). Interactive alignment and lexical triggering of code-switching in bilingual dialogue. *Frontiers in Psychology*, 11. 18

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suarez, P. O., Orife, I., Ogueji, K., Rubungo, A. N., Nguyen, T. Q., Müller, M., Müller, A., Muhammad, S. H., Muhammad, N., Mnyakeni, A., Mirzakhalov, J., Matangira, T., Leong, C., Lawson, N., Kudugunta, S., Jernite, Y., Jenny, M., Firat, O., Dossou, B. F. P., Dlamini, S., de Silva, N., Çabuk Ballı, S., Biderman, S., Battisti, A., Baruwa, A., Bapna, A., Baljekar, P., Azime, I. A., Awokoya, A., Ataman, D., Ahia, O., Ahia, O., Agrawal, S., and Adeyemi, M. (2022). Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72. 3

Labov, W. (2006). *The social stratification of English in New York city*. Cambridge University Press. 7, 13

Lee, G., Yue, X., and Li, H. (2019). Linguistically motivated parallel data augmentation for code-switch language modeling. In *Interspeech*, pages 3730–3734. 7

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. 3

Lorenz, E. (2019). 2018. bilingualism in the community. code-switching and grammars in contact, written by rena torres cacoullos, and catherine e. travis. *Journal of Language Contact*, 12(2):523–531. 2

Martínez, R. A. (2017). Texas education review, volume 5, issue 1: Dual language education and the erasure of chicanx, latinx, and indigenous mexican children: A call to re-imagine (and imagine beyond) bilingualism. 2

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., et al. (2022). Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*. 4

Myers-Scotton, C. (1993). Common and uncommon ground: Social and structural factors in codeswitching. *Language in society*, 22(4):475–503. 2

Myers-Scotton, C. (1997). *Duelling languages: Grammatical structure in codeswitching.* Oxford University Press. 7

Oladipo, A., Adeyemi, M., Ahia, O., Owodunni, A., Ogundepo, O., Adelani, D., and Lin, J. (2023). Better quality pre-training data and t5 models for african languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168. 3

Pattichis, R., LaCasse, D., Trawick, S., and Cacoullos, R. (2023a). Code-switching metrics using intonation units. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16840–16849. vi, 17

Pattichis, R., LaCasse, D., Trawick, S., and Cacoullos, R. (2023b). Code-switching metrics using intonation units. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16840–16849, Singapore. Association for Computational Linguistics. 18

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics. 3

Poplack, S. (1980 (2013)). Sometimes I'll start a sentence in Spanish y termino en espanol: toward a typology of code-switching. *Linguistics*, 51(s1):11–14. 2, 7, 13

Poplack, S. (1983). Bilingual competence: Linguistic interference or grammatical intergrity? 9

Poplack, S. (2017). *Borrowing: Loanwords in the Speech Community and in the Grammar.* Oxford University Press. 16

Poplack, S. (2018). *Borrowing: Loanwords in the speech community and in the grammar.* Oxford University Press. 2

Poplack, S. (2022). 16 Data management at the uOttawa Sociolinguistics Laboratory. *The Open Handbook of Linguistic Data Management*, page 209. 19

Pratapa, A., Bhat, G., Choudhury, M., Sitaram, S., Dandapat, S., and Bali, K. (2018). Language modeling for code-mixing: The role of linguistic theory based synthetic data. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553, Melbourne, Australia. Association for Computational Linguistics. 4, 8

Pratapa, A. and Choudhury, M. (2021). Comparing grammatical theories of code-mixing. In Xu, W., Ritter, A., Baldwin, T., and Rahimi, A., editors, *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 158–167, Online. Association for Computational Linguistics. 4

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67. 3

Sankoff, D. (1998). A formal production-based explanation of the facts of code-switching. *Bilingualism: language and cognition*, 1(1):39–50. 2

Solorio, T. and Liu, Y. (2008). Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981. 31

Steuck, J. (2018). *The prosodic-syntactic structure of intra-sentential multi-word code- switching in the New Mexico Spanish-English bilingual community*. PhD thesis, Pennsylvania State University. 15

Tarunesh, I., Kumar, S., and Jyothi, P. (2021). From machine translation to code-switching: Generating high-quality code-switched text. *arXiv preprint arXiv:2107.06483*. 31

Torres Cacoullos, R. and Travis, C. E. (2018). *Bilingualism in the community: Code-switching and grammars in contact*. Cambridge University Press. 1, 12, 13, 19

Torres Cacoullos, R. and Travis, C. E. (2020). Code-switching and bilinguals' grammars. In Adamou, E. and Matras, Y., editors, *The Routledge Handbook of Language Contact*, pages 252–275. Routledge. 16

Torres Cacoullos, R. and Vélez Avilés, J. (2023). Mixing adjectives: A variable equivalence hypothesis for bilingual word order conflicts. *Linguistic Approaches to Bilingualism*. 16

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. 3

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Linzen, T., Chrupała, G., and Alishahi, A., editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics. 9

Winata, G., Aji, A. F., Yong, Z. X., and Solorio, T. (2023). The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics. 2, 8

Winata, G. I., Cahyawijaya, S., Liu, Z., Lin, Z., Madotto, A., and Fung, P. (2021). Are multilingual models effective in code-switching? In Solorio, T., Chen, S., Black, A. W., Diab, M., Sitaram, S., Soto, V., Yilmaz, E., and Srinivasan, A., editors, *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153, Online. Association for Computational Linguistics. 3

Winata, G. I., Madotto, A., Wu, C.-S., and Fung, P. (2019). Code-switched language models using neural based synthetic data from parallel sentences. In Bansal, M. and Villavicencio, A., editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 271–280, Hong Kong, China. Association for Computational Linguistics. 8

Wintner, S., Shehadi, S., Zeira, Y., Osmelak, D., and Nov, Y. (2023). Shared lexical items as triggers of code switching. *Transactions of the Association for Computational Linguistics*, 11:1471–1484. 27, 28

Yong, Z. X., Zhang, R., Forde, J., Wang, S., Subramonian, A., Lovenia, H., Cahyawijaya, S., Winata, G., Sutawika, L., Cruz, J. C. B., Tan, Y. L., Phan, L., Phan, L., Garcia, R., Solorio, T., and Aji, A. (2023). Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages. In Winata, G., Kar, S., Zhukova, M., Solorio, T., Diab, M., Sitaram, S., Choudhury, M., and Bali, K., editors, *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore. Association for Computational Linguistics. 4

Zentella, A. C. (1998). *Growing up bilingual: Puerto rican children in New York*. Blackwell. 6, 7

Zhang, R., Cahyawijaya, S., Cruz, J. C. B., Winata, G., and Aji, A. (2023). Multilingual large language models are not (yet) code-switchers. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics. 4